Collaborative Approaches to AI Governance: Exploring Co-Design and Co-Regulation Models

Inyoung Cheong

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
Tadayoshi Kohno, Chair
Ryan Calo
Hugh Spitzer
Amy X. Zhang
Benjamin Mako Hill

Program Authorized to Offer Degree:
Law

University of Washington

Abstract

Collaborative Approaches to AI Governance: Exploring Co-Design and Co-Regulation Models

Inyoung Cheong

Chair of the Supervisory Committee:

Professor Tadayoshi Kohno
Computer Science and Engineering

As Artificial Intelligence (AI) systems become increasingly complex and their social impacts more profound, traditional methods to technology governance are proving inadequate. Neither top-down regulation nor unfettered market freedom appears capable of addressing the multifaceted challenges posed by AI technologies. In response, participatory and democratic strategies are gaining traction in AI governance discussions as potential solutions. However, the practical implementation of these approaches and the challenges they may face remain underexplored in the literature, with existing theories often remaining idealistic and detached from real-world constraints.

This dissertation introduces a novel approach to AI governance by integrating co-design and co-regulation methodologies. It addresses three key research questions: (1) How might co-design and co-regulation be defined and integrated in the context of AI governance? (2) How can domain-specific expert knowledge be effectively elicited and integrated into AI governance policies? (3) How do co-regulation models in related domains facilitate stakeholder collaboration, and what lessons can be applied to AI governance?

The research methodology combines theoretical analysis with empirical research, featuring two in-depth case studies. The first study examines co-design in AI systems providing legal advice, engaging legal experts to develop guiding principles based on time-tested wisdom in legal professional communities. The second investigates co-regulation in online content moderation in South Korea, comparing web comics and news industries to identify critical factors for successful co-regulation. Both studies apply established frameworks

to contexts previously unexplored in the AI governance literature.

The key findings reveal that effective participatory governance requires strategic calibration of participation rather than maximizing participation, balancing diverse needs, motives, and contextual factors. This nuanced approach acknowledges that while multi-stakeholder involvement is crucial, shared responsibilities can lead to diffused accountability. This work specifies key questions for utilizing participatory or collaborative methods in AI governance and identifies significant contextual factors that may determine the success of such systems. It equips policymakers, AI developers, academics, and affected communities with pragmatic viewpoints, moving beyond abstract ideals to address real-world complexities.

As a starting point for future research at the intersection of participatory design, collaborative regulation, and AI development, this dissertation challenges the field to advance from theoretical discussions to pragmatic, context-sensitive strategies in AI governance.

Acknowledgements

It is said that it takes a village to raise a child. In my case, though not a child, I have grown with the support of multiple communities.

My deepest gratitude goes to my advisor, Prof. Tadayoshi Kohno ("Yoshi"). He has been an unwavering presence during my most challenging moments and greatest triumphs, never losing his generous smile even when I repeatedly tested the limits of his patience. Yoshi has taught me to cultivate an acute awareness of diversity and inclusion, from his choice of words to his considerate actions, in both research and daily life. Yoshi's deep insights into harm mitigation, societal concerns, and policy implications and his ability to navigate across various disciplines have guided my work at the intersection of technology and law.

I had the privilege of working with a remarkably multidisciplinary committee. Prof. Ryan Calo's contributions were invaluable, helping to structure the work, sharpen its focus, and synthesize legal research with STS perspectives. Prof. Hugh Spitzer was one of my most ardent supporters throughout my doctoral journey, helping to hone my writing, and his teachings on Professional Ethics were instrumental in my analysis regarding guardrails on AI providing legal advice. Under Prof. Amy X. Zhang's mentorship, I learned to tailor my research for academic communities and foster productive cross-field collaborations. Her invitation to participate in the Democratic Inputs to AI project provided me with a valuable opportunity to combine legal knowledge with AI ethics, which ultimately culminated in the first case study of this dissertation. Prof. Benjamin Mako Hill exceeded the typical GSR role, offering incisive feedback on the overall direction of my work and introducing me to the literature on institutional analysis and development framework.

My research was made possible through generous financial support. The UW Tech Policy Lab (co-directed by Ryan and Yoshi, among others) and OpenAI's Democratic Inputs to AI Grant (directed by Amy) funded three years of my doctoral studies. The initial two years were supported by the Fulbright Graduate

Scholarship and a sabbatical grant from the South Korean government.

I am profoundly grateful to the many collaborators and mentors who contributed to my intellectual and personal growth (in alphabetical order by last name): Aylin Caliskan, Joseph Chee Chang, Quen Ze Chen, Kyoungjin Choi, Nicole Decario, Tyna Eloundou, Pardis Emami-Naeini, K.J. Kevin Feng, James Grimmelmann, Katie S. Gonser, Alicia Guo, Liwei Jiang, Kowe Kadoma, Hee Jung Lee, Mina Lee, Teddy Lee, Won-woo Lee, Jenny Liang, Simona Liao, Jimin Mun, Mor Naaman, Aviv Ovadya, Kentrell Owens, Alan Rozenshtein, Maarten Sap, Sikang Song, Sangmin Son, Miranda Wei, King Xia, and Jina Yoon. Special thanks to Shosh Westen, a valued friend and excellent proofreader who dedicated her talent to this work.

A practical perspective on AI governance in this dissertation comes from the knowledge and experience I gained while working as a regulator in the Korean government, particularly the Korea Communications Commission and the Ministry of Culture, Sports, and Tourism. My sincere thanks go to: Joongseop Bae, Bongjin Chang, Hyunseung Choi, Sungjoon Choi, Sungman Han, Yunjin Ha, Jieun Hwang, Hyunrae Jo, Sangwon Jung, Joondong Kim, Jaeyoung Kim, Junghoon Kim, Jeong Won Kim, Jun Sang Kim, Kyunghwan Kim, Yong-sam Kim, Mansoo Lee, Sora Lee, Gyeonghee Wang, Sangwoong Yoon, and Sungchun Yoon.

The foundation of my academic curiosity was laid by my late father Dong-soo Cheong, who found his greatest joy in seeing me immersed in books. His legacy has been supported by my mother Jinyoung Bae with her immeasurable love, trust, and perseverance. I have received heartfelt support from my siblings Insun Cheong and Youngil Cheong and my parents-in-law Wonho Kim and Haekyung Park. At the heart of my journey is my partner, Kyoungche Kim, whose love, patience, and understanding have carried me through the most challenging times, even when I doubted myself the most. This milestone in my academic journey coincides with the arrival of our son, Roy Yuel Kim, marking a profound moment of completion and new beginnings, filling our lives with the greatest joy.

As an AI researcher, I have been impressed by the sophistication of machine intelligence. However, five years of research have reinforced my belief in the irreplaceable value of human connections. Despite technological breakthroughs, it is the people around me that foster growth, reflection, and curiosity. As I look toward a future where the line between human and machine may blur even further, I remain convinced that our interpersonal bonds will continue to be a fundamental source of strength and inspiration.

DEDICATION

To Jinyoung, Kyoungche, and Roy Yuel

who taught me love, resilience, and wit to appreciate all that life unveils

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial intelligence (AI)[1] has been a subject of research and development for decades, but the recent emergence of powerful generative models such as ChatGPT and DALL-E signals a profound shift in the AI landscape, promising to revolutionize how we interact with and leverage AI in various domains. Human-like conversational capabilities and the vast knowledge of large language models (LLMs) have shown promise in improving access to services traditionally requiring human specialists [282, 394, 210], in domains such as healthcare [77, 367, 336, 219, 364], finance [294, 377], and law [198, 392, 289]. However, their rapid advancement has sparked widespread concern, with many users experiencing a mix of fascination and unease about the implications for our collective future.

1.1 Various Threats Posed by AI Technologies

At the individual level, the adverse impacts converge on the fundamental issues of privacy and autonomy. Although AI technologies promise to augment individual abilities, they also risk compromising personal data and decision-making processes, potentially leading to a loss of control over one's thoughts, emotions,

[1] Among many other AI systems, this dissertation primarily focus on **generative AI systems**, which include Large Language Models and Diffusion Models, and their applications. These systems have the ability to generate diverse outputs based on user prompts. They are characterized as "general-purpose" AI systems due to their capacity to respond to unanticipated commands. This dissertation covers both the model level (e.g., GPT-4) and the application level (e.g., ChatGPT). While this dissertation does not exclude open-source models from its scope, it places a greater emphasis on commercial AI systems developed and provided by corporations. This focus is motivated by two key factors. Firstly, commercial AI systems are more widely used among the general public, as open-source models often require advanced knowledge of computing, making them less accessible to the average user. Secondly, corporations have a clear accountability to establish the norms of their AI systems, which inevitably leads to questions of legitimacy regarding the rule making power of corporations.

and behaviors. At the societal level, concerns encompass job displacement and exacerbation of economic inequality; copyright and intellectual property issues, particularly the devaluation of labor and the creativity of human artists and content creators; and environmental costs associated with training and deploying large-scale models. Table 1.1 illustrates the taxonomy of concerns in three widely cited articles in the NLP and AI communities.[2]

Table 1.1: Major Concerns about Generative AI

Bommasani et al. (2022) [106]	Solaiman et al. (2023) [368]	Bender et al. (2022) [95]
• Bias and over-representation • Social inequity • Misuse • Copyrights and liability • Privacy and surveillance • Discrimination • Concentration of power • Environmental costs	• Trustworthiness and autonomy • Personal privacy and sense of self • Concentration of authority • Labor and creativity • Ecosystem and environment	• Over-representation of dominant groups • Biases and stereotypes against marginalized groups • Static training data • Environmental and financial costs

Note: Author's own compilation based on the cited literature.

1.2 Challenges to AI Governance

To combat these threats, several legislative initiatives have emerged, such as the European Union's Artificial Intelligence Act (EU AI Act) [49], Canada's Artificial Intelligence and Data Act (AIDA) [65], and the U.S. Algorithmic Accountability Act [52]. However, critics are concerned whether traditional top-down approaches to technology governance—which rely on centralized authority and static regulations—cannot keep pace with the speed and complexity of emerging technologies[165, 107, 360, 404, 108]. The global scale and ubiquitous nature of AI systems have challenged the capacity of individual nations or organizations to govern them unilaterally. This perspective is reflected in the Japanese government's 2021 statement that "legally-binding horizontal requirements for AI systems are deemed unnecessary at the moment." [205] Moreover, government involvement in regulating AI-generated content raises normative concerns, particularly regarding free speech in countries like the United States [125, 126, 209].

[2]More details on AI-mediated harms to mental and physical well-being, privacy, autonomy, and fairness can be found in Appendix C.9.

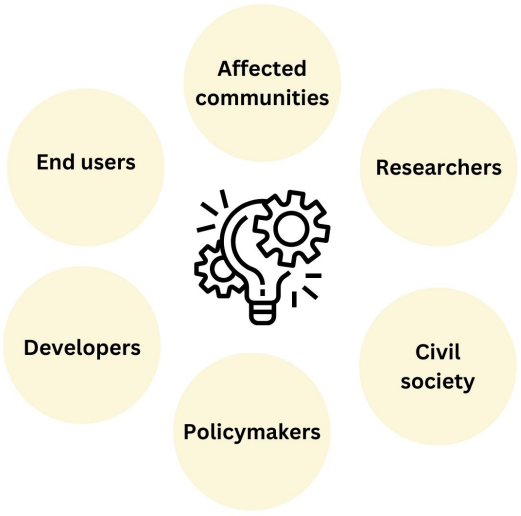
At the same time, leaving governance primarily in the hands of private companies presents its own set of challenges. The current AI landscape is dominated by a small number of powerful entities such as Google, Microsoft, OpenAI, and DeepMind, which have become *de facto* overseers of AI development. These companies now possess significant power in rule-making, with their user policies often having as much, if not more, impact on people's daily lives as state-level laws.

It is known that the most common safety measures of large base AI models, such as Reinforcement Learning with Human Feedback (RLHF), presume a universal set of values, distinct from personal preference or community-specific norms [186, 211]. If values differ between social groups, which 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? Is it safe for a handful of AI companies to have the sole authority to answer these questions?

Furthermore, profit motives do not automatically encourage robust safety efforts. Throughout the evolution of the Internet, we have observed that ethical considerations (e.g., protecting privacy) can easily be overlooked for the sake of commercial gain (e.g., targeted advertising) [130, 326, 235]. Despite AI companies' early commitment to safety, competitors with lower standards could offer more capabilities, faster, cheaper, and more entertaining ways. It also remains unclear what incentives exist for companies of varying sizes to fully adopt safety methods. For example, the collection of human feedback, red team testing, robustness checks, and user monitoring demand significant expertise, computing resources, and human oversight [206, 418]. Although larger companies may absorb costs, smaller players need solutions that are mindful of resource constraints.

Accordingly, both top-down regulatory mechanisms and market-driven solutions have shown limitations in addressing the multifaceted challenges posed by AI technologies. This complex landscape has led to a shift in governance dynamics. Governments increasingly rely on company information and expertise to regulate AI technologies effectively. Meanwhile, growing public distrust in government bodies due to privacy and surveillance concerns has led to a search for alternative governance models. Companies, in turn, are motivated to seek more input from the public and policymakers, both for ethical reasons and to ensure compliance and avoid controversies. In light of these challenges, there is a growing interest in exploring alternative governance models that combine elements of public, private, and civil society oversight.

Figure 1.1: Definitions of Co-Design and Co-Regulation



CO-DESIGN

"A participatory process whereby relevant stakeholders, users, and affected communities in collaboratively developing a **particular socio-technical system**."

CO-REGULATION

"Collaborative development and **enforcement** of **universal rules and standards** by a range of stakeholders, including government agencies, industry representatives, civil society organizations, and experts."

1.3 Novel Approach to AI Governance

This dissertation contributes to this emerging discourse by exploring two alternative governance models that have gained traction in many other fields: **co-design** and **co-regulation**. Co-design models bring participatory methods to the development of AI systems at the company or organizational level, ensuring that diverse stakeholder perspectives are incorporated from the outset. Co-regulation models facilitate collaboration among multiple stakeholders, including government agencies, industry representatives, civil society organizations, and academic experts, in developing and enforcing universal rules and standards across multiple services.

This dissertation proposes a unique framework that integrates co-design and co-regulation as complementary schemes for AI governance. While these concepts have been extensively discussed in separate domains—co-design in human-computer interaction (HCI) and computer-supported cooperative work (CSCW), and co-regulation in public policy, law, political science, and economics—their integration in the context of AI governance represents a novel perspective. By connecting these two models, we create a multi-layered governance framework that addresses both the micro-level design decisions of individual AI systems and the macro-level regulatory environment in which they operate.

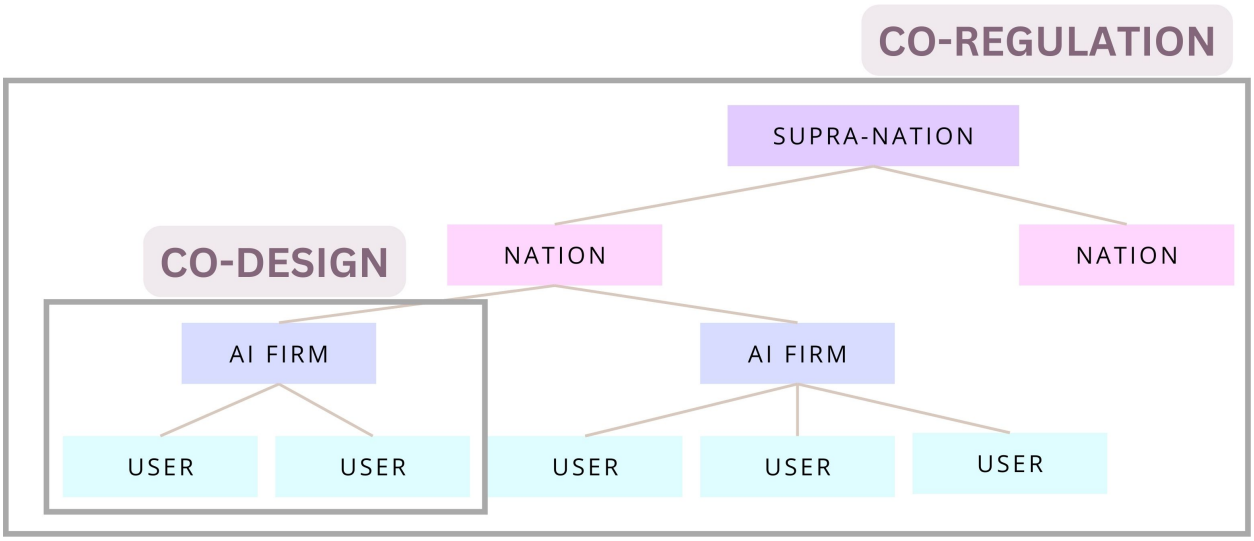


Figure 1.2: Co-Design and Co-Regulation in Tiered AI Governance

1.4 Research Questions and Methodology

This dissertation seeks to answer three key research questions:

1. How might co-design and co-regulation be defined, and why have they emerged as alternative approaches in the context of AI governance?
2. In a co-design model, how can domain-specific expert knowledge, such as that of legal professionals, be effectively elicited and integrated to identify key dimensions and guiding principles for responsible AI policies?
3. How do co-regulation models in related domains facilitate collaboration among diverse stakeholders, and what lessons can be learned from these approaches to inform the development of AI governance frameworks?

For RQ1, Chapter 2 lays the theoretical foundations of this dissertation by synthesizing discussions on these two models across computer science, law, and public policy literature, highlighting their relevance to AI governance. Building on this theoretical base, RQ2 is addressed in Chapter 3 through the first case study. This study focuses on co-design in AI governance, specifically incorporating legal experts' perspectives into AI systems providing legal advice. This study revealed critical considerations specific to AI-powered legal advice systems and demonstrated the value of domain-specific expertise in AI governance, while also

highlighting the challenges in translating diverse expert insights into actionable policies.

RQ3 is explored in Chapter 4 through the second case study, which investigates co-regulation in online content moderation in South Korea, comparing web comics and news industries. Through interviews with 15 key stakeholders and the application of Ansell & Gash's collaborative governance framework, the study identified critical factors for successful co-regulation, including stakeholder interdependence and clear ground rules. It demonstrated how industry-specific contexts significantly influence co-regulation effectiveness and extended existing frameworks to include normative and practical challenges relevant to internet and AI regulation. The dissertation concludes by looking for the steps forward in Chapter 5, emphasizing context specificity and human-centric processes in AI governance.

1.5 Contributions

This dissertation makes significant contributions to the field of AI governance by theorizing and empirically examining co-design and co-regulation models within the complex landscape of AI development and deployment. Through an interdisciplinary lens, spanning human-computer interaction, public policy, law, and organizational studies, this research offers insights that bridge micro-level design decisions and macro-level regulatory frameworks. Using case studies and qualitative methods, the research provides rich and contextual insights into the challenges and opportunities of implementing collaborative governance in AI.

This dissertation advances established theoretical models. It extends a case-based reasoning methodology to the domain of AI governance, demonstrating how this methodology can elicit expert knowledge and generate actionable insights for complex, evolving technologies. Furthermore, the study applies and refines Ansell & Gash's collaborative governance framework in the context of AI, revealing how factors such as stakeholder interdependence, facilitative leadership, and shared understanding influence the success of co-regulatory efforts in different industries.

This work contributes to the development of pragmatic and forward-looking guiding principles for AI governance. These principles, grounded in empirical research, go beyond idealized notions of participation to offer actionable insights for effective governance strategies. They emphasize: (1) the critical importance of context specificity in governance models, recognizing that different AI applications and sectors may require tailored strategies; (2) the need to view AI governance as a human-centric process, highlighting the

value of sustained, well-designed stakeholder engagement; and (3) the importance of resolving legal ambiguities surrounding AI regulation, particularly in areas where new technologies intersect with established legal frameworks.

By highlighting the critical role of process and organization in effective AI governance, this research establishes a solid foundation for future studies. It opens new avenues for exploring the application of co-design and co-regulation across various domains of AI development and deployment. The work's emphasis on pragmatic institutional design provides valuable guidance for policymakers, industry leaders, and civil society organizations seeking to develop effective, adaptive, and sustainable AI governance frameworks.

1.6 Limitations

The study's focus on specific domains—AI providing legal advice, news apps, and comics apps—may not fully capture the breadth of applications for general-purpose AI systems. Governance models developed for these specific areas might face challenges when scaled to broader AI applications. The stakeholder engagement in this research mainly involved experts and professionals, such as creators, reporters, and online platforms. This approach, while valuable, lacks direct input from the perspectives of end users. The rapid evolution of AI capabilities constantly reshapes stakeholder landscapes, potentially limiting the long-term applicability of current findings. Furthermore, the shifting power dynamics among stakeholders as AI impacts various sectors differently was not fully explored.

This study did not extensively explore the political and economic barriers to adopting co-design and co-regulation initiatives. Practical challenges in implementing these governance models in different political and economic contexts may require further investigation. The case studies, while informative, may not be fully representative of all AI governance scenarios, potentially limiting the generalizability of the findings to other cultural, political, or technological contexts. Moreover, given the rapid pace of AI development, some findings can quickly become outdated, necessitating ongoing research and updates. Furthermore, research may not fully capture the perspectives of all relevant stakeholders in AI governance, particularly those of marginalized or underrepresented groups.

Moving forward, studies should expand to a broader range of AI applications, incorporate more diverse stakeholder perspectives——particularly those of end-users and marginalized groups——and employ longi-

tudinal approaches to capture the evolving nature of AI governance challenges. In addition, comparative studies across various political and economic contexts will be crucial in developing more universally applicable governance frameworks. As AI continues to transform society, this work lays the foundation for ongoing research that can adapt to technological advancements and ensure that AI systems align with social values and needs.

Chapter 2

Theoretical Framework

2.1 Conceptual View of Governance and Regulation

Before examining approaches to AI governance systems, a few conceptual remarks are in order.[1] First of all, we need to distinguish between governance and regulation, terms often used interchangeably due to their shared characteristics [170]. Both are intentional, goal-oriented, and collaborative processes involving multiple stakeholders working to shape the structural and procedural aspects of a given domain. Governance and regulation both analyze how various actors—including individuals, corporations, non-profits, and government entities—influence, guide, control, and steer the development and usage patterns of services, technologies, or institutions. These actors participate in rule-making and direct the evolution of specific services and technologies towards particular outcomes.

Definition of Governance. Governance is a broader concept than regulation, encompassing both formal and informal rules and norms, while regulation typically refers to formal rules and mechanisms for controlling behavior. The government's role in governance varies, ranging from state-centric to more society-centered models [317]. Some scholars view government as an indispensable component of governance. For example, Francis Fukuyama regards the government as an indispensible component of governance, defin-

[1]The legal history discussion in this chapter is derived from Inyoung Cheong, Aylin Caliskan & Tadayoshi Kohno, Safeguarding Human Values: Rethinking US Law for Generative AI's Societal Impacts, AI and Ethics (2024), https://doi.org/10.1007/s43681-024-00451-4. Given the importance of empirical analysis in understanding generative AI's potential harms and the scope of current US legal protections, the full paper is included in Appendix C for comprehensive reference.

ing it as "a government's ability to make and enforce rules, and to deliver services, regardless of whether that government is democratic or not." [183] Similarly, Ansell & Gash defines governance as "a governing arrangement where one or more public agencies directly engage non-state stakeholders in a collective decision-making process that is formal, consensus-oriented, and deliberative and that aims to make or implement public policy or manage public programs or assets." [76]

Other scholars employ much broader definitions of governance. For example, "the setting, application and enforcement of the rules of game" (Kjaer [241]); "the stewardship of formal and informal political rules of the game, [which] involves setting the rules for the exercise of power and settling conflicts over such rules." (Hyden [216]) For our discussion, this broader definition of governance may be more appropriate than narrower, state-centric definitions, reflecting a more inclusive and multifaceted approach to shaping AI development and use. AI governance extends beyond the formal regulations and laws enacted by governments, and includes the voluntary standards, best practices, and ethical guidelines developed by industry, academia, and civil society. AI governance often includes policies adopted by companies and community norms shaped by users.

To conceptualize dynamic forms of governance, Kenneth Abbott & Duncan Snidal propose the "Governance Triangle" model [71]. The model categorizes actors involved in governance schemes into three main groups: 'firm,' which includes individual companies, industry associations, and other company groupings; 'NGO,' a broad category encompassing civil society groups, international non-governmental organizations, academic researchers, activist investors, and individuals; and 'state,' which comprises both individual governments and supranational government groupings like the European Union and the United Nations. Figure 2.1 situates the range of governmental, non-governmental, and international actors participating in business regulation in the triangle.

Definition of Regulation. Meanwhile, the seminal definition of regulation, established by Philip Selznick in 1985, is "the sustained and focused control exercised by a public authority over activities valued by the community." [348] This definition highlights the traditional perspective that views regulation as a dichotomy between freedom and control. On the one end of the spectrum, the government can allow the market to operate freely, granting firms the autonomy to pursue their own interests without interference. Alternatively, the government can impose regulations that restrict this autonomy by introducing the threat of sanctions,

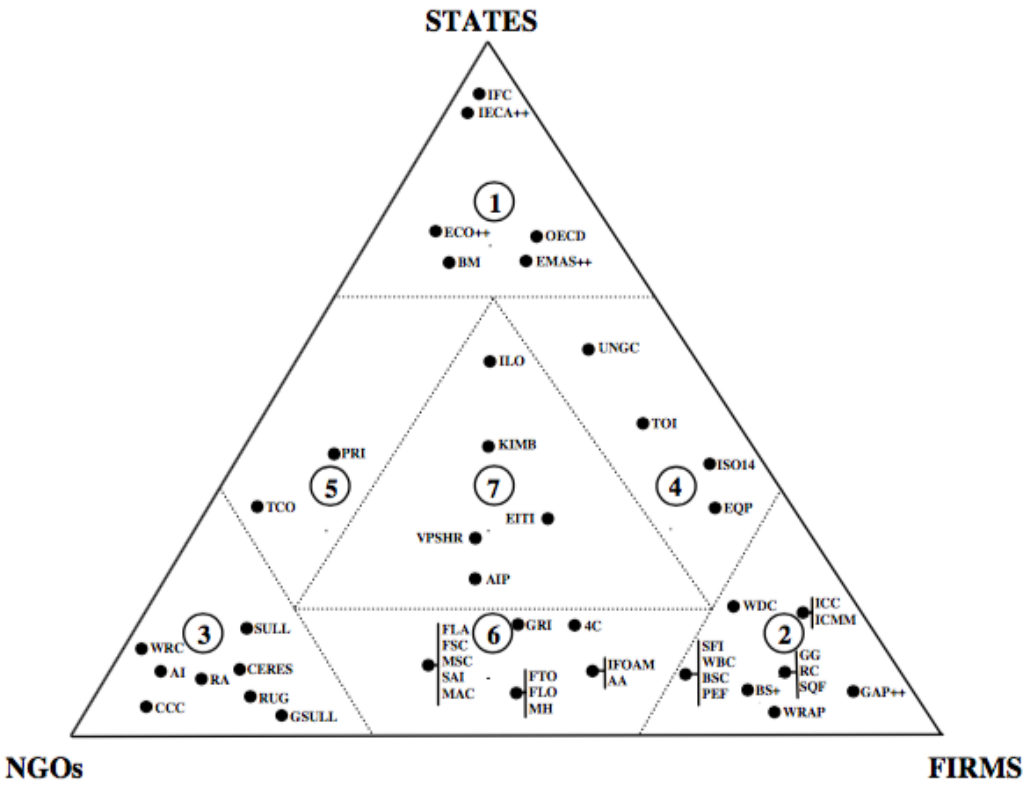


Figure 2.1: Governance Triangle illustrated by Abbott & Duncan [71].

forcing firms to align their interests with those of society. The latter approach, often referred to disparagingly as 'command-and-control' regulation, involves the government directly dictating and enforcing rules to achieve desired outcomes [135].

There is a historical debate concerning market failure and government failure. Proponents of a market-based approach argue that government intervention inevitably hinders market efficiency and should be limited to necessary roles, such as the distribution of property rights and policing of crimes. They maintain that market forces, driven by self-interest and competition, can effectively allocate resources and promote innovation. Conversely, advocates of government regulation contend that market failures, such as monopolies, externalities, and information asymmetries, necessitate government intervention to protect the public interest and ensure fair competition. They argue that leaving everything to the market can lead to undesirable outcomes, including exploitation of consumers, environmental degradation, and widening social inequalities. Government regulation, they assert, is essential to correct these market failures and promote the greater

good.

However, the conventional view of regulation as a choice between a free market and command-and-control regulation is overly simplistic and fails to account for the diverse range of regulatory options that exist between these two extremes. In practice, most regulatory systems involves a wide variety of pressures and policies implemented by both governmental and non-governmental actors to influence the behavior of firms and tackle market failures and other public concerns. This complexity extends beyond just the actors involved; modern regulation encompasses a diverse range of mechanisms that go far beyond simple "control." These mechanisms include rewards and incentives, information disclosure requirements, and transparency mandates, among others.

Accordingly, this dissertation adopts Julia Black's more inclusive definition of regulation: "The intentional use of authority to affect behaviour of a different party according to set standards, involving instruments of information-gathering and behaviour modification." [100] Furthermore, Cary Coglianese & Evan Mendelson break regulation down into four essential elements: (1) targets (a different party), (2) regulators (who intentionally uses the authority), (3) commands (the instructions the regulator gives the target on how to modify its behavior), and (4) consequences (fines and sanctions for non-compliance or subsidies and incentives for compliance) [135].

Although adopting broad definitions of governance and regulation inevitably leads to some convergence between these concepts, regulation exhibits several defining features that distinguish it within the broader governance landscape. It is characterized by its formality, involving established procedures and official agreements. The universality of regulation is evident in the wide applicability of the standards it develops. Furthermore, regulation is marked by significant government participation, and government actors play a key role in the process. These characteristics highlight regulation's structured approach to collaborative rule-making, setting it apart from other governance mechanisms.

Definition of Co-Design and Co-Regulation. This dissertation's two main focuses, co-design and co-regulation, both fall under the broader governance umbrella, yet maintain distinct characteristics. Co-design is a participatory process whereby a single firm or organization engages relevant stakeholders, users, and affected communities in collaboratively developing a particular social-technical system. Co-regulation, on the other hand, involves government and non-government actors working together to create universally

applied standards across various types of socio-technical systems.

2.2 Development of Co-Design and Co-Regulation

Collaborative initiatives emerged as a response to demands from both the government and the private sector, each seeking to enhance the legitimacy of their decision-making processes by sharing power with other stakeholders. On the one hand, governments recognized the need to involve actors in the private sector, such as companies and specialized non-profits, in the rule-making process to leverage their expertise and ensure the practicality and effectiveness of regulations. By sharing power with these private entities, governments sought to improve the quality and legitimacy of regulation, particularly in the face of complex technological and market dynamics.

On the other hand, corporations and institutions faced growing pressure from civil society to incorporate the needs and perspectives of workers, service users, and the broader public in their decision-making processes. The participatory design (PD) movement, which gained prominence in Scandinavia, exemplified this demand for greater inclusion and collaboration in the design and implementation of services and technologies. It has been widely adopted and further developed in the fields of science, technology, and society (STS) and computer-supported cooperative Work & social computing (CSCW) to improve the responsiveness and legitimacy of new technologies.

The emergence of these collaborative approaches, conceptualized as co-design and co-regulation in this dissertation, reflects a broader shift towards decentralization and multistakeholder participation in governance and decision-making. By involving a wider range of actors and considering their diverse interests and expertise, these initiatives seek to address the limitations of traditional top-down, state-centric, or corporate-centric models of regulation and design. As such, the rise of Co-regulation and Co-design can be understood as a response to the demands for greater legitimacy, effectiveness, and inclusivity in the governance of complex sociotechnical systems.

2.2.1 Shared Responsibilities: Beyond Government v. Market Dichotomy

The question of who should possess the power of rule-setting and implementation has been a long-standing debate in all societies. Compared to European and Asian countries, the US has historically been more

predisposed to supporting individual liberty, market freedom, and limited government intervention. The American Revolution and the drafting of the US Constitution were motivated by the imperative to protect individual rights from potential encroachments by government authorities [39]. The workings of the early American state relied primarily on the judgments of state courts, supported by and in conjunction with political parties. As James Madison put it: "The powers delegated by the proposed Constitution to the federal government are few and defined." [274]. This cultural ethos of skepticism towards 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.

In the US common law tradition, legal doctrines are not static pronouncements but evolve dynamically through the resolution of adversarial disputes between individuals [226]. This case-by-case approach unfolds at both the federal and state levels, reflecting a strong emphasis on individual rights and responsibilities. It empowers individuals and interest groups to actively engage in legal battles, advocating for their perspectives and seeking just resolutions. Judges and juries, while guided by legal precedents, must also consider the unique context of each case, allowing for nuanced interpretations and applications of the law. This pluralistic approach acknowledges that legal questions seldom have single, fixed answers. It embraces the richness of diverse viewpoints as cases are decided, setting precedents that reflect the complexity of society and its evolving values.

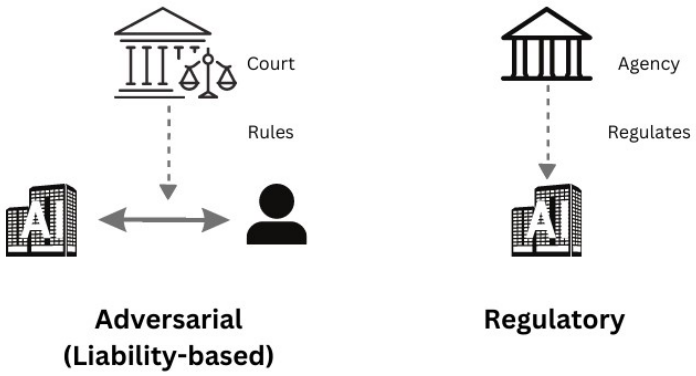


Figure 2.2: Comparison Between Adversarial and Regulatory Legal Systems, Illustrated by Cheong, Caliskan, and Kohno [126].

Consider a scenario where air pollution becomes a pressing concern. Two potential policy avenues emerge: Congress could enact legislation, establishing an agency to monitor polluting businesses and set emission standards. Alternatively, the legislature could create a private cause of action, empowering individuals directly affected by pollution to sue for damages. This "fault-based" liability system incentivizes responsible behavior and allows individual redress for harm suffered. Figure C.3 visually contrasts these two approaches, highlighting the inherent differences between the adversarial and regulatory models.

European and Asian legal systems are more inclined to establish regulations that prioritize social welfare and collective rights. This trend stems from different notions of freedom and the role of the government. With regards to privacy law, 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 [405]. Similarly, Asian cultures, influenced by collectivist ideologies, emphasize community well-being and social cohesion over individual liberty [313, 94]. For instance, Miyashita 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," [286] which might seem to belie the very foundation of privacy for Western readers.

Despite these different values and viewpoints that shaped the early regulatory landscapes in the US and European/Asian countries, the 20th century saw a convergence in regulatory approaches. Both regions experienced notable advancements in their regulatory systems, a phenomenon often referred to as the development of the *administrative state* [178] or *regulatory state* [416]. During an era of world wars, many countries prioritized the tasks of post-war reconstruction, redistribution, macro-economic stabilization, and the provision of welfare as an employer of last resort. To expand government control over major resources, states began to own or directly intervene in key industries, including public utilities such as gas, electricity, water, telecommunication, and the railways. In the US, between 1933 and 1948, Franklin Roosevelt offered a package of 'New Deal' reforms comprised of public works and social insurance programs. US federal government agencies rapidly grew rapidly, assuming a far greater responsibility for the nation's economy that had been shaken by the Great Depression.

The early 1970s saw an expansion of federal government roles, driven by an increased interest in health, safety, and environmental preservation, bolstered by an activist judiciary [416]. This era was characterized

by a belief in the separation of politics and administration through 'expert' administration and 'scientific' strategies. The US Congress empowered administrative agencies to create regulations addressing complex domain-specific issues while adhering to defined objectives [380]. Examples include the Environmental Protection Agency's mandate under the Clean Air Act to establish air quality standards for public health protection [11], and the Occupational Safety and Health Act's provision for reasonable workplace safety [16].

However, by the mid-1970s, economic challenges such as rising inflation and unemployment prompted a shift in the governance approach. Governments began privatizing state-owned assets and moving from direct control to more arm's length oversight during the 1980s and 1990s, both in Europe and the US. This transition was rooted in the belief that market forces could more efficiently allocate resources and stimulate economic growth. In the US and UK, the administrations of President Reagan and Prime Minister Thatcher championed free markets and implemented policies reflecting skepticism towards government intervention. Their initiatives, supported by the growing law and economics movement, advocated for smaller government, lower taxes, and reduced regulations to boost economic growth and job creation. These efforts led to widespread deregulation across such industries as telecommunications, transportation, and finance.

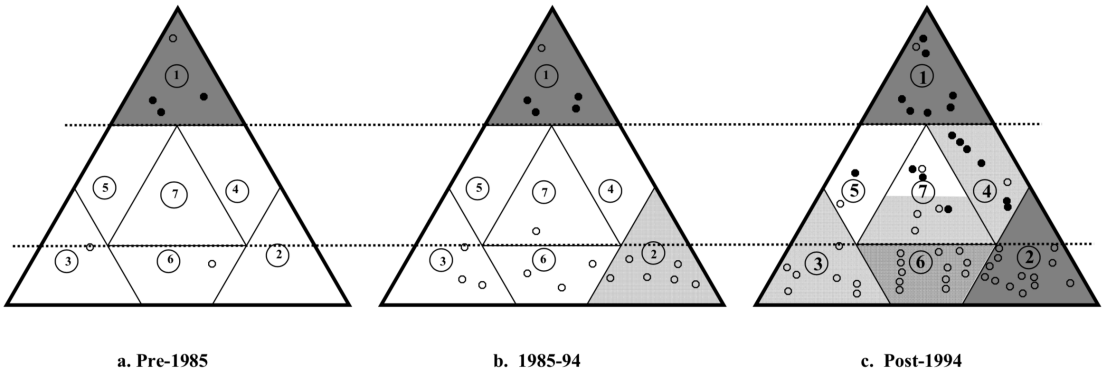


Figure 2.3: Evolution of Governance Triangle Illustrated by Abbott & Snidal [71].

This transition led to the dissolution of the regulatory authority solely possessed by the government and the start of deeper involvement by private actors in rule-setting and implementation. Abbott & Snidal visualized the evolution of regulatory governance through the governance triangle. Figure 2.3 illustrates the shifting balance of power and influence among the three main groups of actors—states, firms, and

NGOs—over three time periods: pre-1985, 1985-1994, and post-1994. The pre-1985 era shows a strong state presence, with limited involvement of firms and NGOs. However, we observe a gradual dispersion of power, with firms and NGOs gaining more prominence in the regulatory landscape, due to skepticism about government intervention. This figure highlights the complex interplay of actors, mechanisms, and debates that balance market forces, government intervention, and novel challenges posed by technological advancements.

Borrowing the governance triangle framework, Robert Gorwa analyzed various regulatory mechanisms of online content operating in the EU. In Figure 2.1, Germany's NetzDG represents a state-level initiative operating at the national scope, while the Global Network Initiative (GNI) is a multi-stakeholder effort involving firms and NGOs, working at a global level to promote privacy and freedom of expression in the information and communications technology industry. Another firm-led initiative is the Global Internet Forum to Counter Terrorism (GIFCT), a collaborative effort by major technology companies to prevent terrorists and violent extremists from exploiting digital platforms. In contrast, the Facebook Oversight Board (FBO) is a unique firm-level institution operating globally, serving as an independent body to review content moderation decisions and provide policy recommendations, essentially functioning as a "supreme court" for content-related issues on the platform [244].

Name	Date	Actors	Scope
Network Enforcement Act (NetzDG)	2017	State	National
Audiovisual Media Services Directive (AVMSD)	2018	State	Regional
Dynamic Coalition on Platform Responsibility (DCPR)	2014	State-Firm-NGO	Global
Safer Social Networking Principles (SSP)	2009	State-Firm	Regional
EU Code of Conduct on Terror and Hate Content (CoT)	2017	State-Firm	Regional
EU Code of Practice on Disinformation (CoD)	2018	State-Firm	Regional
Christchurch Call (CC)	2019	State-Firm	Global
Global Network Initiative (GNI)	2008	Firm-NGO	Global
Twitter Trust and Safety Council (TWC)	2016	Firm	Global
Global Internet Forum to Counter Terrorism (GIFCT)	2018	Firm	Global
Facebook Oversight Body (FBO)	2019	Firm	Global
Manila Principles on Intermediary Liability (MAP)	2015	NGO	Global
Santa Clara Principles on Content Moderation (SCP)	2018	NGO	Global

Table 2.1: Content Governance Landscape of Online Platforms Operating in the EU, formulated by Gorwa [197].

2.2.2 Democratizing Centralized Technologies

Scholars in science, technology, and society (STS) have emphasized the importance of societal inputs to technology development [252, 376, 153, 407]. Drawing from Habermas' theory of communicative action [92], STS scholars argue that information and computing systems are not purely confined to the realm of natural science and the mathematical aspects of computers, with the study of their effects and long-range social consequences being left to other disciplines in the human and social sciences [161]. Instead, they view technology as socially constructed, reflecting and shaping various aspects of human activity. This perspective challenges the notion of technology as a neutral "thing" and emphasizes the need for more inclusive and democratic approaches to technology design and development. As Winner pointedly argues:

> Consciously or not, deliberately or inadvertently, societies choose structures for technologies that influence how people are going to work, communicate, travel, consume, and so forth over a very long time. In the processes by which structuring decisions are made, different people are differently situated and possess unequal degrees of power as well as unequal levels of awareness. [407, p. 127]

Another stream of thought, called the self-determination theory (SDT), has emerged as a framework by which to understand human motivation and its impact on various outcomes, such as creative problem-solving, well-being, and engagement [145, 146, 344, 251, 187]. SDT proposes that motivation exists on a continuum, ranging from autonomous (self-determined) to controlled (externally regulated) motivation. This theory also emphasizes the importance of basic psychological needs (autonomy, competence, and relatedness) in fostering autonomous motivation and well-being, which can be crucial factors in promoting meaningful participation and engagement.

The desire to address these power imbalances and unlock the benefits of inclusive technology development played a central role in the emergence of participatory approaches to research and decision-making. This trend is most often traced back to the work of Scandinavian researchers in the 1970s and 80s [407, 161]. The 'Scandinavian approach' to participation, known as participatory design (PD), was initially concerned with creating 'workplace democracy' through structured consultation and dialogue between workers and employees, aiming to give workers greater control over wages and the allocation of tasks.

Over time, PD has significantly expanded its scope beyond the workplace to include design processes in urban planning and community building [80, 111, 164, 182, 269, 338], political contexts [93, 104, 250], and the intersection of technology design and use [312]. PD actively involves stakeholders, users, and communities in the design process, emphasizing collaboration, empowerment, and democratization in the creation of technologies, products, and services [105].

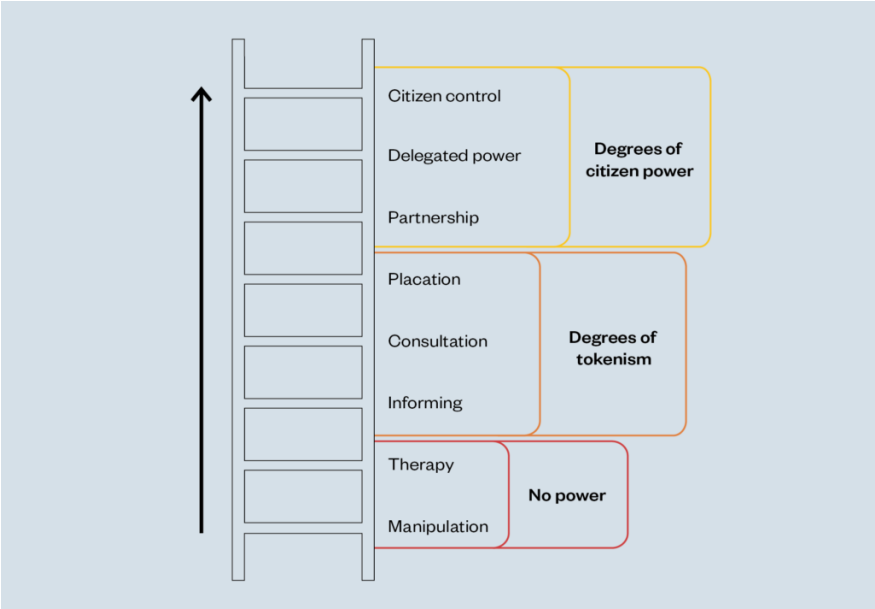


Figure 2.4: Sherry Arnstein's 'ladder of citizen participation,' illustrated by Ada Lovelace Institute [51].

One classic and influential frameworks in PD is the "Ladder of Citizen Participation" proposed by Sherry R. Arnstein [80]. This approach employs a visual metaphor of an eight-rung ladder (see Figure 2.4), where each rung represents the degree of power afforded to people by different approaches to participation. The higher the rung on the ladder (with eight being the highest and one being the lowest), the more power people have when engaging in the public planning of their communities. Arnstein developed this ladder to empower disenfranchised individuals and communities, enabling them to enact meaningful social reform by reclaiming power from authorities during participatory processes in community development. The ladder serves as a tool for assessing the level of genuine participation in decision-making processes and highlights the importance of shifting power dynamics to ensure that marginalized voices are heard and have a real impact on outcomes.

Another influential framework in the field of public participation is the IAP2 (International Association

for Public Participation) Spectrum of Public Participation (see Figure 2.5), whose efforts range from informing to consulting, involving, collaborating, and empowering. The spectrum was developed to support and define the public's role in any public participation process. As a matrix, it helps in identifying the level to be chosen based on the specific goal of the project and the promise being made to the public. It is widespread in North America and Australia, with different institutions following and adapting the spectrum to their specific needs [143].

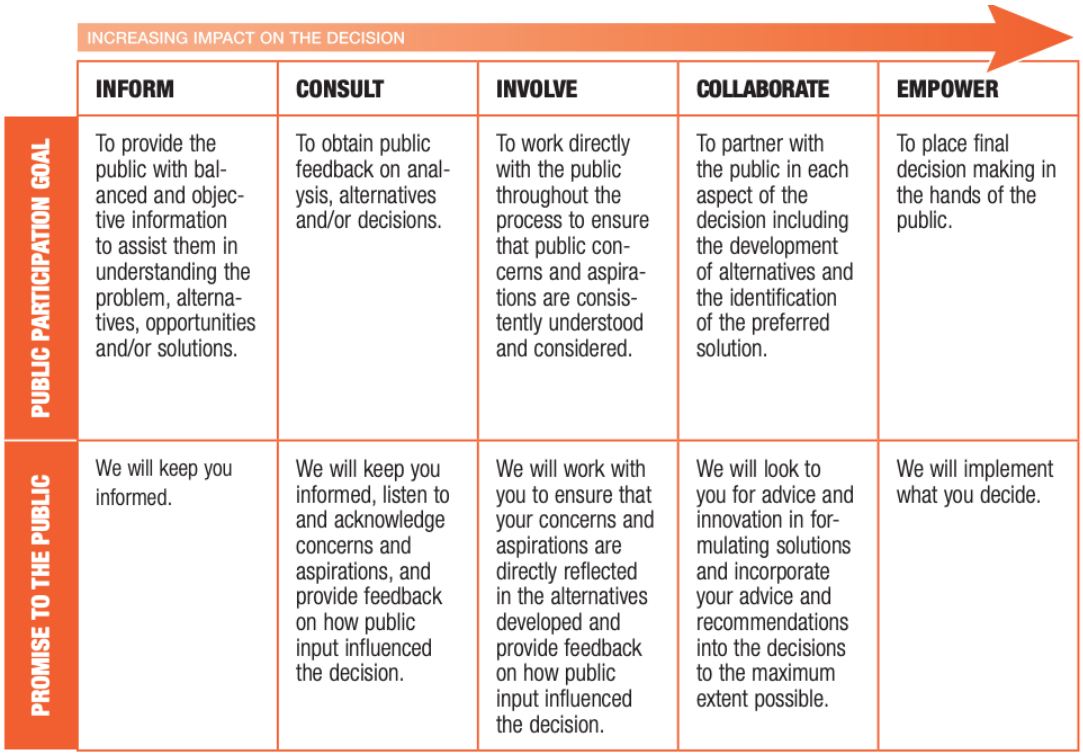


Figure 2.5: IAP2 Spectrum of Public Participation [174].

Building upon the IAP2 Spectrum, the Community Involvement Matrix (see Figure 2.6) maps a wide variety of engagement processes based on the inherent risk of the situation and the complexity of information which needs to be understood. This matrix further demonstrates the diverse range of participatory approaches and their applicability in different contexts, depending on the level of risk and complexity involved.

These frameworks and matrices highlight the importance of considering power dynamics, context, and the specific goals of participatory processes when designing and implementing them. By understanding and

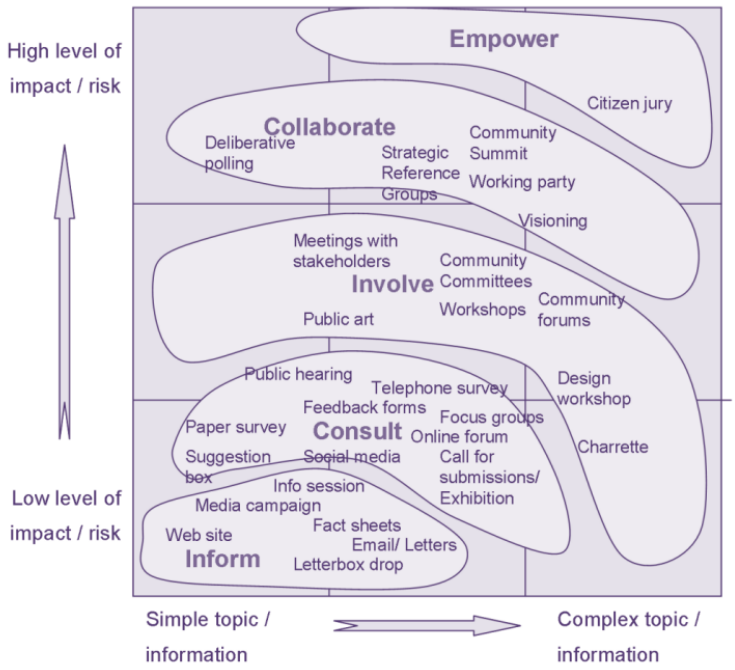


Figure 2.6: Community Involvement Matrix, developed by Robinson [328] and illustrated by Northen Beaches Council in Australia [44].

applying these concepts, researchers and practitioners can work towards more inclusive, empowering, and impactful participatory approaches in a variety of fields, including technology development and decision-making processes.

The principles and practices of PD align well with the goals of human-computer interaction (HCI) and computer-supported cooperative work (CSCW), whose aim is to design technologies that are usable, useful, and meaningful for individuals and communities [234, 103, 409]. HCI and CSCW scholars have increasingly embraced PD as a way to navigate the complex dynamics of stakeholders and ensure that technology design is responsive to the needs and values of diverse users [103, 147, 148]. The growing interest in PD is evident from the findings of Malinverni et al., who observed a remarkable 580% increase in the number of PD papers in the ACM digital library over a decade (2003-2013) [277].

"Co-design" is a term that has been used interchangeably with PD in multiple studies [278, 324, 223]. This dissertation primarily uses the term "co-design" as a core concept because it fits within a broader, tiered framework of co-governance. This framework includes two levels: platform-level co-design and society-level co-regulation. Using "co-design" consistently helps to clearly distinguish between these different

levels of collaborative governance. While participatory design often connotes grassroots participation from users and marginalized communities, the co-design approach in this dissertation, particularly in Case Study 1, focuses on involving expert stakeholders to develop guiding principles for AI services.

2.3 The Current Landscape of AI Co-Design and Co-Regulation

The rapid advancement of AI technology, particularly the emergence of powerful large-scale generative AI models such as GPT-3.5 and Diffusion, has sparked an unprecedented level of interest in democratizing AI development and incorporating participatory methods. Although efforts to improve content moderation and curation in social media and search engines through increased participation have received significant attention [309, 355, 310, 3], the stakes are even higher when it comes to democratizing the design of general-purpose AI models. These models are designed to serve a wide range of purposes, unlike more specialized AI systems focused on specific tasks such as image recognition or medical diagnosis. The development of these models requires massive computational resources and vast amounts of data, which can only be handled by a few select leading AI companies [72, 340].

These systems are trained on ever-increasing amounts of human-generated data and fine-tuned using human feedback [307, 84], with the rule-of-thumb being "the more data the better" [229]. However, the workers labeling data often lave little connection with the AI development process and have a limited understanding of how their contributions will be used. They are typically employed through crowdsourcing platforms or third-party contractors, with little direct communication or collaboration with the AI companies. Similarly, the red-teaming process, which involves testing AI systems for vulnerabilities and unintended consequences, is often conducted by experts who are hand-picked by AI companies. While these experts may have specialized knowledge and skills, they may not fully represent the diverse perspectives and experiences of the broader population [239]. This can result in blind spots and limitations in identifying potential harms and ethical concerns.

As a result, a few leading actors possess concentrated power over conditioning the norms and outcomes of countless downstream applications built upon these foundation models. This centralization of authority and decision-making power within algorithms and AI systems, rather than human stakeholders, is a growing concern [136]. John Danaher terms this phenomenon "algocracy," which refers to rule or governance

by algorithm. Danaher views algocracy as a significant threat to the legitimacy of public decision-making processes, as it constrains and limits opportunities for human participation. He argues that this shift towards algorithmic governance poses fundamental challenges to moral and political legitimacy, potentially undermining principles of democratic governance and human autonomy in decision-making [142].

2.3.1 Co-Designing AI Systems

To decentralize the algocracy of large AI systems, researchers are turning to the concept of participatory co-design. The term "participatory AI" has emerged to describe AI systems and design processes that actively involve stakeholders, with the implication that this approach will lead to more empowering and democratic outcomes [365, 368, 148]. The recent "participatory turn" [136] in AI development has led to a growing body of research exploring various approaches to involve stakeholders and ensure that AI systems are developed in an inclusive, ethical, and aligned manner with societal values. Through interviews with industry participants, researchers at the Ada Lovelace Institute identified a variety of participatory approaches employed in AI development, which they mapped onto Arnstein's Ladder of Citizen Participation (See Table 2.2).

Similarly, Corbett et al. [136] used Arnstein's Ladder to compare and contrast the degree of power afforded to participants based on 21 papers from the ACM digital library. For instance, Brown et al. [109] conducted workshops with families and social workers to improve algorithmic decision-making in child welfare, which was considered to be at the middle rungs of the ladder. Although participants were consulted, their concerns may not have been implemented, as the workshops took place outside of an actual development process. In contrast, Lee et al. [255] involved food donation stakeholders to aggregate their preferences via voting for matching algorithms, situating their research slightly higher on the ladder.

The analysis reveals that most papers focused on informing or consulting the public on AI, rather than partnering or delegating control directly in the design process [136]. The authors advocate for greater researcher reflexivity, transparency with communities about the degree of influence, and relational rather than transactional interactions. These concerns were echoed by the tech workers themselves, who incorporate participatory approaches [202]. Interviewees from AI labs expressed concern about "participation washing," which means superficial or insincere adoption of participatory approaches to create a false impression of inclusive, equitable, or democratic decision-making processes [136].

Arnstein's ladder	Participatory approaches
Degrees of citizen power	Cooperatives Citizens' jury Community-based approaches Deliberative approaches Participatory design Speculative design / anticipatory futures Governance tools e.g. audits, impact assessments, other policy mechanisms
Degrees of tokenism	Co-design Community training in AI Community-based Systems Dynamics framework Crowdsourcing UX/user testing Open source Diverse Voices method Workshops/convenings Consultation
Non participation	Surveys Request for comment

Table 2.2: Participatory Approaches in Commercial AI Mapped onto Arnstein's Ladder of Citizen Participation [202].

Despite the good intentions of many tech workers to incorporate meaningful participatory methods in AI development, they face significant organizational, cultural, and resource barriers. These challenges include the resource intensity of participatory work, lack of requisite skills and experience among practitioners, atomization of participation efforts within organizations, and a lack of incentives for transparency [202]. The following list summarizes the key limitations and challenges of participatory design identified by scholars based on several decades of experience:

- **Power disparity**: Democratic ideals notwithstanding, power imbalances between researchers and participants can persist. Researchers often retain control over the design process and outcomes, reflecting certain privileged mindsets and values, thus limiting the genuine empowerment of participants. Existing societal power structures and inequities can also carry over into design spaces [148, 207, 365].
- **Lack of inclusion**: Projects may struggle to involve a truly representative range of participants, especially from marginalized groups. Barriers to access, trust, compensation and power can limit which

voices are heard. Reliance on proxies or narrow demographics can skew insights [207, 261].

- **Reflexivity and imagination**: Effective participation requires deep reflexivity about power and a willingness to question traditional design concepts. A lack of critical self-awareness and difficulties embracing imaginative new possibilities can constrain the process [207, 365, 343].
- **Resource constraints**: Effective participation requires significant time, funding and human resources to build long-term, equitable partnerships with communities. Practical limitations can lead to rushed, superficial or short-term engagement that does not allow for deep participation or impact [202].
- **Superficial participation**: Academic and commercial reward systems often do not incentivize the type of intensive, long-term, community-centered work required for meaningful participation. Misaligned incentives can lead to extractive or superficial participation [365]. Based on interviews with practitioners in AI development, Grove et al. [202] found that there is no incentive to be transparent about its adoption.
- **Limited scalability**: Insights generated in a specific context may not scale or translate well to other settings where a design is to be implemented. The richness of localized, participatory input can get diluted or lost in the scaling process [147, 148].
- **Ambiguity in methods**: Beyond the general goal of "democratization," the specific motivations and approaches to participation in development practices vary significantly [372]. Questions remain about who should be involved and what aspects of the design process should be open to participation [147].

The current landscape of participatory AI is marked by a lack of shared principles and a wide variety of approaches that are often unaccounted for and unaccountable. Although there is a growing consensus on the importance of stakeholder participation in AI design, the methods, theories, and goals for leveraging participation vary greatly. To address these gaps, Delgado et al. [148] developed the Parameters of Participation Framework by synthesizing common threads across technology design, political theory, and social sciences. The framework, largely inspired by the IAP2 Spectrum of Public Participation (Figure 3.2), provides a conceptual structure to advance participatory practices in AI design. It enables AI researchers and practitioners to align participatory methods with specific underlying participatory goals and to be more reflexive about the affordances and limitations of current proxy-based participatory tactics.

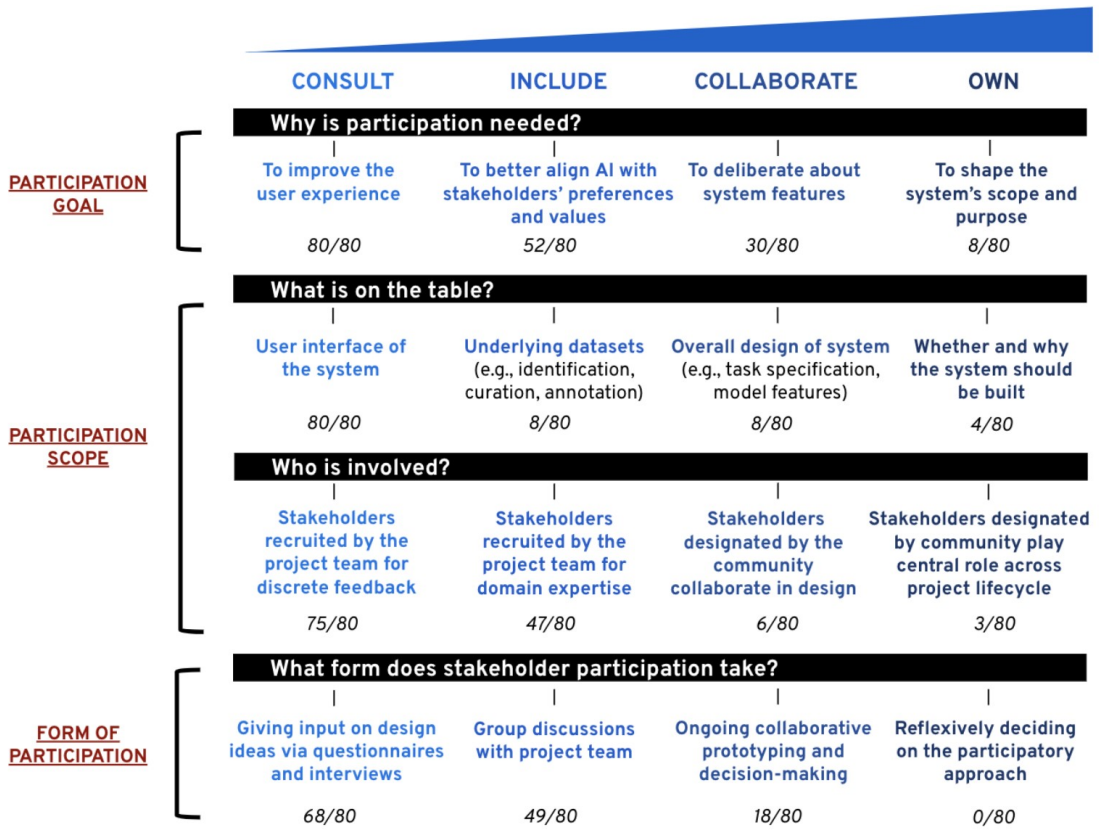


Figure 2.7: A Conceptual Framework Proposed by Delgado et al. [148] to evaluate approaches to participation in AI design.

Building upon Delgado et al.'s framework [148], Suresh et al. [381] categorizes various approaches to incorporating public input and participation in the development and governance of AI foundation models. The authors group these approaches into four main types: Reinforcement Learning with Human Feedback (RLHF), rulesets and policies, red teaming, and domain-oriented efforts.They evaluate each approach based on the degree of meaningful participation it allows. Their analysis reveals that most current methods fall into the "consult" or "include" modes of participation, with limited stakeholder influence on decision-making. While some domain-oriented efforts show potential for more substantive participation, the authors suggest that there is room for improvement in creating more collaborative and impactful participatory processes in AI governance.

2.3.2 Co-Regulating AI Systems

There is a growing demand to involve public policy and law in establishing fundamental rules for AI safety, given the technology's significant scale and potential impacts. By enacting relevant rights and obligations, the legislature can reflect the public's desire for safety and trustworthiness in AI systems, ensuring that the technology is developed and used in a manner that aligns with societal values and expectations. Various global initiatives have emerged to address AI governance. The European Union's AI Act and Canada's proposed AI Act represent statutory approaches to regulating AI, with the aim of establishing comprehensive frameworks for the development, deployment, and use of AI systems, as outlined in Table 2.8.

In the US, federal actions include the Biden administration's AI Bill of Rights blueprint outlining civil liberties principles [57], an AI risk management framework from the National Institute of Standards and Technology [63], and an Executive Order [215]. Individual agencies are also examining emerging AI risks in areas such as medical devices [169], political advertising [66], and biometric privacy [61].

International organizations, such as the Organisation for Economic Co-operation and Development (OECD) and the United Nations Educational, Scientific and Cultural Organization (UNESCO), have developed AI principles and guidelines to promote responsible AI development. The OECD AI Principles [48] and the UNESCO Recommendation on the Ethics of AI [391] provide voluntary frameworks for stakeholders to adopt and implement in their AI practices. In addition to these formal initiatives, there are various informal arrangements and collaborations among stakeholders in the AI ecosystem, including industry-led

Figure 2.8: Summary of AI Regulatory Approaches by Country.

COUNTRY	AI REGULATION	KEY FEATURES
EU	AI Act (2024)	• Bans certain AI uses (e.g., government social scoring) • Risk assessment, transparency, registration, and human oversight requirements for high-risk AI
CANADA	AI and Data Act (Introduced in 2022)	• Risk assessment, transparency, registration, non-discrimination, and human oversight requirements for high-impact AI
CHINA	Interim Measures for Generative AI Services (2023)	• Prohibits generating content that endangers national security or social stability • Risk assessment, license/approval, data protection, and non-discrimination requirements
US	Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI (2023)	• Tasks various federal agencies with specific responsibilities from cybersecurity to to critical infrastructure to content authentication

Note: Author's own compilation based on the proposed/passed regulations.

standards, best practices, and ethical guidelines developed by AI companies, professional associations, and civil society organizations.

However, the regulation of AI is still in its infancy. Among the approaches listed in Table 2.8, only the EU AI Act is close to being implemented as a legally binding law. Concerns arise that formal regulation could suppress innovation. The majority of information regarding AI development is held by leading companies, creating a power imbalance and hindering effective regulation. This situation, combined with a distrust in the government's ability to directly regulate technology with sweeping rules, has raised concerns about over- or poor-regulation. These voices, citing the nascent stage of the technology, warn against potential inefficiencies and unintended consequences arising from prematurely rigid regulation, including stifled innovation and regulatory capture [165, 107, 360, 404, 108]. This stance echoes the historical debates surrounding internet regulation in the late 20th century, where concerns for online free speech ultimately prevailed over internet safety regulation [123]. This resonates with the deeply ingrained American ethos of "adversarial legalism," favoring gradual conflict resolution over ex-ante regulations, as articulated by Kagan [226].

AI governance is primarily shaped by burgeoning court decisions and the EU AI Act, whose details of implementation remain uncertain. Co-regulation of AI, which requires careful calibration of institutions and accumulated discussions among stakeholders, is largely at the ideation stage. Current regulations do invite some level of private actors' involvement. For instance, the EU AI Act involves third-party organizations, known as notified bodies or conformity assessment bodies, in the regulation process. However, full-blown collective rule-making is not yet suggested. These approaches are still more akin to traditional government regulation with administrative agencies playing a central role.

The rapid evolution of AI technology, the formation of new creator classes affected by AI advancements, and the global nature of AI services make it nearly impossible for a single government to effectively regulate AI unilaterally. Considering the information asymmetry between governments and companies, a flexible co-regulatory system involving both governmental and non-governmental actors may be more realistic. However, it also has limitations. The shared responsibility structure can make it difficult to guarantee ongoing participation and accountability, and non-governmental bodies may face challenges in enforcing measures against non-compliance. This raises the question of whether co-regulation can truly serve as an alternative in AI governance.

2.4 Empirical Investigations: Co-Design and Co-Regulation in Practice

While the theoretical landscape of co-design and co-regulation in AI governance offers valuable insights, it also reveals significant challenges and uncertainties. To move beyond abstract concepts and understand how these models might function in real-world scenarios, this dissertation examines two case studies that provide empirical insights into the implementation of these approaches in contexts relevant to AI governance. The following chapters delve into the findings of these case studies in detail.

The first case study explores a co-design model for AI governance, focusing on incorporating legal experts' perspectives into the development of AI systems providing legal advice. Through workshops engaging 20 legal experts in case-based deliberation, this study develops a comprehensive framework for AI governance policies and highlights the challenges of translating expert insights into actionable policies.

The second case study investigates co-regulation in online content moderation in South Korea, comparing web comics and news industries. By analyzing interviews with 15 key stakeholders and applying

collaborative governance frameworks, this study identifies critical factors for successful co-regulation and examines how industry-specific contexts influence co-regulation effectiveness.

Together, these case studies provide a nuanced understanding of the challenges and opportunities in implementing participatory and collaborative approaches to AI governance. They offer practical insights that complement and expand upon the theoretical foundations discussed earlier, paving the way for more effective and context-sensitive governance strategies in the rapidly evolving field of AI.

Chapter 3

Case Study 1: Co-Designing Legal AI Systems with Legal Experts

3.1 Background

This case study represents a concrete exploration of co-design principles in AI governance, focusing on the critical domain of AI-powered legal advice systems.[1] This study, conducted as part of OpenAI's "Democratic Inputs to AI" program, serves as a practical examination of how participatory approaches can be implemented in AI policy development. This research not only contributes to our understanding of co-design methodologies in AI governance but also illuminates the challenges and opportunities in translating expert knowledge into implementable policies. As such, it provides valuable insights into the broader questions of how co-design and co-regulation can be effectively integrated into AI governance frameworks and the complexities involved in balancing diverse stakeholder inputs in rapidly evolving technological landscapes.

In the legal field, where attorneys undergo extensive training to provide counsel, often beyond the reach of laypeople [374, 403], AI-based chatbots offering legal advice have emerged as a potential accessibility

[1]This chapter is largely based on research published in the proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24) under the following citation: Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 16 pages. https://doi.org/10. 1145/3630106.3659048. For the most current information and further details about this ongoing research, please refer to our project website: `https://social.cs.washington.edu/case-law-ai-policy/`

aid [203, 345]. However, relying on imperfect AI systems for high-stakes legal decisions raises concerns about low-quality advice and privacy risks [403, 230, 74]. These concerns have prompted the EU AI Act to designate AI systems used for "assistance in the legal interpretation and application of the law" as "high-risk" [163].

Most prior research in this field speculates on the inherent limitations of AI systems such as inaccuracy, shallow reasoning, or poor predictive capability [265, 230, 282, 392, 290]. While meaningful, these studies rarely articulate concrete criteria for *when* and *why* AI systems should or should not provide professional advice to users. As a result, they offer insufficient guidance to produce actionable design requirements that can inform real-world deployment practices. One potential solution is to consult with domain experts who can offer insights on the unique challenges and needs of their domain. Learning from experts is an emerging approach for responsible AI policies [77, 316], but has not been applied to the legal domain. Our work aims to bridge this gap by addressing the following questions:

- **RQ1**: What key dimensions do legal experts identify in determining appropriate AI responses to lay users' legal questions?
- **RQ2**: What guiding principles and response strategies do legal experts recommend for AI systems that provide legal advice to lay users?

My research team used a process (Figure 3.1) inspired by case-based reasoning, an approach commonly used in pedagogical material for a wide variety of fields, including law and moral theory [124, 172, 248, 314, 224], to allow discussion of ethical considerations grounded in concrete cases. We convened seven interactive workshops with 20 legal experts by providing them with 33 queries ("cases") and asked them to evaluate seven simulated responses that could arise from AI chatbots, ranging from outright refusal to recommendation of specific actions with legal judgment. Through analysis of the collected data, iterative rounds of discussion among authors, and literature review across the fields of law, natural language processing (NLP), and AI ethics, we consolidated and identified the significant dimensions that affected experts' evaluations and guiding principles for desirable AI responses.

For **RQ1**, we identified 25 key dimensions that should inform potential AI responses (Figure 3.3). We classified dimensions into four categories: (1) user attributes and behaviors, (2) nature of queries, (3) AI capabilities, and (4) social impacts. For **RQ2**, experts generally expressed their preferences for information-

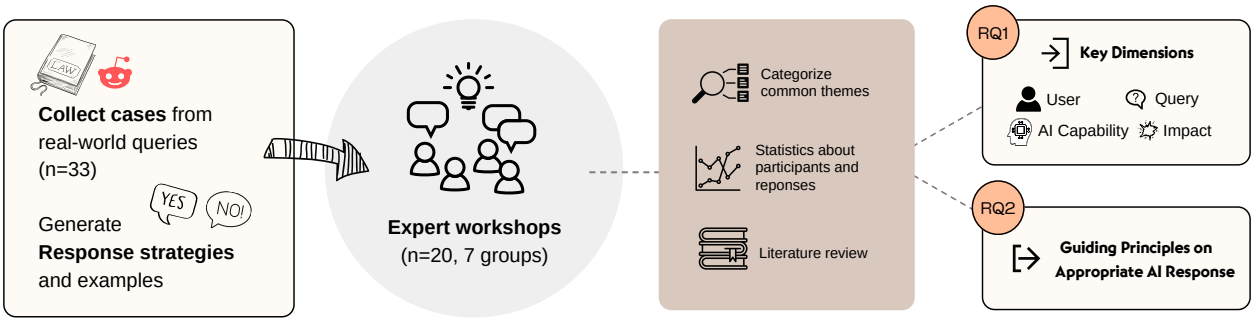


Figure 3.1: Overview of Our Research Process and Findings

oriented responses. Instead of seeking definitive legal judgement, some suggested leveraging AIs' multi-turn dialogue capabilities to polish users' questions and distill relevant facts through follow-up questions. Furthermore, experts proposed additional layers of ethical guidelines such as "Do not pretend to be a human," or "Respect the justice system."

Our contributions are multi-fold: First, our four-dimensional framework, spanning across query-specific concerns to more systemic constraints grounded in legal and technical literature, provides fertile groundwork for AI policy creation beyond speculative theoretical principles. Second, in addition to dimensions, we portray experts' disagreements on appropriate AI responses, while highlighting where experts agreed on information-focused or multi-turn issue-spotting approaches. Third, we demonstrate how our case-based expert deliberation process was effective in leveraging experts' knowledge and experience to elicit a rich set of dimensions. We discuss how our methods and our resulting 4-dimensional framework could potentially be adopted for further research in other professional domains. Finally, we reveal novel legal and ethical considerations, such as the unauthorized practice of law, confidentiality, and liability for inaccurate advice, overlooked in the AI literature. This illustrates that responsible AI legal advice requires a cross-disciplinary synthesis that spans technology, law, and ethics, learning from accumulated knowledge in professional communities.

3.2 Methodology

We conducted **seven** small-group workshops on Zoom with 20 expert participants in August 2023. We assumed a scenario involving **general-purpose conversational AI systems** like ChatGPT or Bing Chat

available to lay users, different from professional tools assisting legal practitioners.

Eliciting Expert Knowledge and Case-based Reasoning. Incorporating the knowledge and insights of domain experts and the public into the development of AI systems has emerged as a key approach called "participatory AI" [85, 137, 99, 148, 316, 320, 291]. Researchers have facilitated expert discussions to evaluate the sociotechnical implications of LLMs [364, 368, 316, 77]. Unlike prior work focusing on high-level ethical principles [368, 77, 188] or post-hoc system evaluation [394, 316, 86], we pursued the **case-based reasoning** [124, 172] approach to spur expert deliberation based on their clinical experience. We presented legal professionals with realistic legal queries that AI systems might receive from lay users. This approach came from moral philosophy [248, 273, 314, 224, 366, 184] and legal theories [120, 200, 379] that emphasize case-by-case judgments to shape guidelines instead of applying top-down rules. Distinguished from AI policies that provide a single set of universally-agreeable principles [77], case-based deliberation enabled us to highlight critical value-laden topics on which experts disagreed with each other. Furthermore, it allowed us to synthesize a dimensional framework, ranging from case-specific concerns to structural constraints, which experts consider to determine appropriate AI responses.

Recruitment. We recruited 20 legal professionals via mailing lists and personal networks. Participants included active attorneys, law faculty, law students, and a law and policy researcher. Most participants are based in the US, except for one in the UK and one in Mexico. The cohort spanned early-career lawyers to those with over 20 years of experience, with varying degrees of AI usage. Table 3.1 summarizes the participants' backgrounds and self-reported AI usage patterns. More detailed information is available at Appendix A.1.

Background	**Occasional AI User**	**Regular AI User**
Attorney	P5, P17, P18	P2, P4, P8, P10, P11, P13, P14, P16, P20
Law faculty	P1, P3, P9	P6
Law student	-	P7, P15, P19
Legal researcher	-	P12

Table 3.1: Participants' Backgrounds and the Frequency with Which They Used AI

Construction of Cases. We manually sourced 33 cases from a combination of (1) the popular subreddit `r/legaladvice` (with wording edited slightly for anonymization and clarity), and (2) existing cases

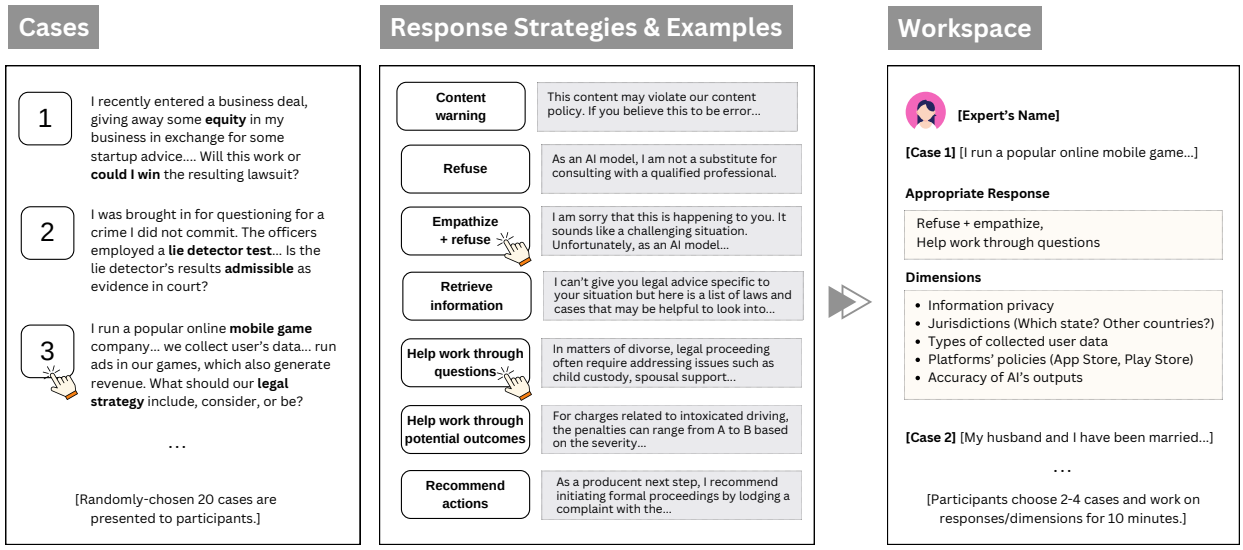


Figure 3.2: Overview of Case Examples and AI Response Strategies and Examples Provided to Participants

in legal practice familiar to our team member who is a practicing attorney. Our cases covered facets of law most relevant to lay users, spanning family law, criminal procedure, housing issues, and employment disputes. We selected cases that represented a diverse range of user intents (e.g., getting out of trouble, advocating for others, minimizing their costs), impacted third parties (e.g., employers, colleagues, landlords, family members, protesters), and degrees of damage (e.g., physical, financial, mental). This diversity was intended to elicit a wide range of discussion across legally and ethically sensitive contexts. Our cases can be viewed at `https://github.com/Social-Futures-Lab/case-law-ai-policy/blob/main/data/cases.csv`.

Workshop Procedures. During the workshop, we presented 20 randomly-chosen cases along with seven generic response strategies for AI responses on a shared Google document. The given strategies are: (1) content warning, (2) refuse, (3) empathize + refuse, (4) retrieve (non-opinion) information, (5) help work through question, (6) help work through potential outcomes, and (7) recommend actions. We provided an example response for each strategy, and examples were derived from what we observed from OpenAI's GPT-3.5 and 4. Both response strategies and corresponding examples are available in Appendix A.2. Because GPT tends to refuse to give detailed advice such as options (6) and (7), we drafted more specific answers complemented by known prompt engineering techniques such as drawing a hypothetical scenario [199].

Figure 3.2 provides an overview of the collaborative document we gave the participants. After an intro-

duction, each participant was given 10–15 minutes to freely choose 2–4 cases and (1) select the proper AI response strategies or produce their own preferred response and (2) the key dimensions impacting their decision in an individual workspace. Then, the experts had 30–35 minutes to discuss with each other why they chose certain response strategies and what dimensions they took into account to determine the appropriate strategies.

Analysis. We analyzed collaborative documents and transcripts using abductive coding [383]. Integrating both empirical data and available theory in an iterative process, our findings are informed by and enter into dialogue with literature from legal ethics [e.g., 208, 403] and ethical concerns in human-AI interactions [e.g., 238, 95]. Our analysis synthesizes relevant aspects of these fields within the context of our research questions. Two authors initially coded two transcripts respectively and developed a codebook of dimensions and responses, informed by Kim et al.'s *human-AI-context framework* [237]. The codebook was finalized through multiple all-author meetings. After this, two coders independently analyzed the data and cross-checked each other's work. In this process, both coders examined all documents and reached consensus on codes, rendering inter-rater reliability metrics unnecessary [35].

IRB, Consent, and Compensation. This study was reviewed and approved by the University of Washington Institutional Review Board. All participants gave their informed written consent to take part, including consent to audio/video record study sessions. Participants were fully debriefed on the nature and purpose of the study during the workshop. Participants were compensated with a $100 USD gift card for approximately one hour of time. Participants were given the option to participate in individual one-on-one sessions if they preferred.

3.3 Eliciting Stakeholders' Major Considerations around AI Risks

Our workshop's structured, case-based deliberations yielded nuanced insights into the multi-layered tensions that arise when using AI systems for legal advice. We identified considerations and concerns in our qualitative data, grouping them into two categories: (1) **Dimensions** capture contextual factors experts considered when determining appropriate AI responses; (2) **Responses** cover desired AI response strategies and guiding principles.

We identified 25 key dimensions that impacted experts' preferences for appropriate AI responses. We

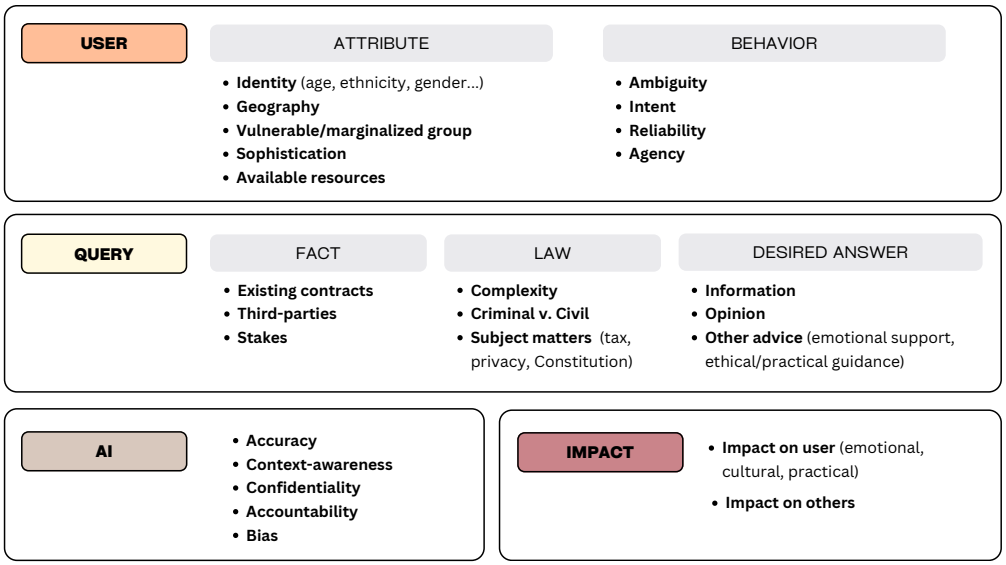


Figure 3.3: 4-Dimensional Framework

classified the dimensions into four categories: (1) user attributes and behaviors, (2) nature of queries, (3) AI capabilities, and (4) social impacts. Figure 3.3 outlines these four categories. We now describe each dimension in more detail.

3.3.1 User Characteristics and Behavior

Our participants identified eight user-related dimensions that AI systems should consider that are broadly broken down into dimensions related to (1) user attributes and (2) user behavior. *User attributes* include identity and background, geographic location, legal sophistication, and access to resources. These are characteristics that users may explicitly provide or can be inferred about. On the other hand, *user behavior* refers to aspects such as reliability, intent, agency, and ambiguity, which can be deduced from the user's inputs and interactions with the AI system but are likely not explicitly stated. Regarding *user attributes*, experts specified four key dimensions:

- **Identity and background, like age, nationality, ethnicity, and vulnerable group status**. Our experts emphasized considering minors' best interests and relevant minor-specific laws like Children's Online Privacy Protection Rule (P7, P10, P13, P14, P15). In addition, nationality (P12), ethnicity (P10), immigration status such as "a DACA recipient" (P12) were perceived to be worth considering. Furthermore, participants considered whether the user was from a "marginalized or vulnerable groups" such

as indigenous people or non-English speakers (P15), acknowledging "structural asymmetries among communities" (P10).

- **Geographic location**. Experts stressed legal variability between jurisdictions: criminal laws vary locally (P12), property lease analyzes differ by location (P7), and 10 US states have separate privacy statutes (P13). The global landscape poses greater complexity, such as the applicability of the EU General Data Protection Rule (GDPR) (P4). Moreover, when interpreting laws from Mexico or Colombia, it is important to consider the unique histories and legal contexts of these countries, which differ from those of the US (P10).
- **Legal sophistication**. Our experts noted that the sophistication level of the user should guide the nature of AI legal advice. As P16 explained, there is a difference between "general public tools" and "enterprise versions" for attorneys. Since attorneys bear the ultimate legal liability, professionally oriented AI tools likely pose fewer risks for misuse. More broadly, P20 suggested that AI systems could provide more advanced and detailed advice to sophisticated users, like a corporate client, who are already familiar with the technology's limitations and are less likely to misinterpret or misuse the information provided.
- **Access to resources**. Our findings revealed that AI systems should contextualize their responses based on the pragmatic restrictions users face regarding time, location, income, and access. If traveling to get medical treatment in a foreign country or retaining a public defender is an unrealistic option, recommendations presuming those resources would poorly serve the user (P8, P11).

The *user behavior* category emerged as experts emphasized that lawyers should not blindly accept user-provided facts. Instead, lawyers must actively probe and ask questions to build an understanding of the situation before offering advice. Our findings revealed four key behavioral dimensions for AI systems to assess:

- **Ambiguity**. Experts stated that if user inputs do not provide enough details about the situation, it is either impossible or risky to provide detailed guidance as the AI outputs are likely to be flawed due to the incomplete information (P1, P6, P13). P1 noted, "So many facts are missing. I'm so nervous about the idea of the chat [giving] you legal advice [based on this incomplete fact]."
- **Reliability**. Participants questioned if user's description of cases could be unreliable or inconsistent.

P5 noted, "There's a lot of facts in [the case], and you don't know to what extent AI should assume they are true [or] an objective fact."

- **Intent**. Participants also wanted to clearly understand the underlying intent of the users. P13 stated that users might also do a poor job of describing their situation, and the AI system should ask for clarification by posing questions like "Are you sure you really mean that?" Some participants were wary of AIs being used to serve the user's malicious intent, such as "to evade law enforcement," (P20) or "to defend his crime to avoid illegal consequences of their actions." (P10)
- **Agency**. Experts emphasized users' degree of agency, or whether users are able to act on the legal guidance given. P17 stated, "There's still consideration beyond giving the advice that someone might still act on that." In the legal setting, unlike in medical contexts where treatment requires that professionals take intermediate steps, users may have substantial direct "power to take action" when provided with legal recommendations, such as firing an employee or filing a complaint (P20).

3.3.2 Query Characteristics

Essentially, legal advice involves applying relevant law to the specific facts of a person's situation. Our participants identified nine key dimensions embedded within users' legal queries that shape what guidance AI systems can provide. We categorized these dimensions into three interconnected parts: (1) relevant facts; (2) relevant laws; and (3) nature of desired answers.

- **Relevant Facts.** Experts emphasized the importance of key facts to provide suitable legal advice. These included granular details around business practices like data collection methods, advertising revenue streams, and the platform's terms of conditions (P4). Existing **contract terms** must be clarified, whether in a lease, employment agreement, conflict waiver, or corporate bylaws (P7, P8, P12). It is also essential to have details on **stakeholders and counter-parties** such as competitors (P13), victims, or injured parties (P6, P11). In addition, assessing the **stakes involved** is a significant factor, ranging from financial liability (P16), to loss of work authorizations or deportation (P12), to imprisonment (P11).
- **Relevant laws.** Experts underscored the **complexity** of many legal issues. Matters involving diverse areas of law (P14) and jurisdictional variation involve a complex legal analysis (P4, P12). The evolving legal landscape necessitates constant research. For example, IP addresses were historically considered

personally identifiable information but are not treated as such under most state laws (P12). The participants also stressed the unique nature of **criminal matters**. The heightened risks in prosecution and incarceration, as well as complex human factors in plea bargaining or sentence hearings, make attorney representation essential (P10, P11). P11 exemplified the idiosyncrasies of judges, quoting a religious federal judge in Washington state: "It really helps. If you're a Christian, and a criminal defendant appears before you, you should always start with a little prayer when you're doing your sentence hearing." Experts pointed to special considerations for **subdomains like tax, privacy, and constitutional law** as requiring specialized judgment. The tax code is big, complex, and ambiguous, so even experienced attorneys should make "judgment calls." (P13, P19) Privacy laws varies substantively state-by-state (P13) and constitutional matters often involve complex values far broader than codified rules (P20).

- **Nature of desired answers**. Participants stressed that the quality of the answers depends on what the particular user seeks from the conversation. Users may want straightforward **informational** outputs, like when using traditional search engines (P11–13, P16). In this case, presenting the list of relevant laws for users' further research could be helpful (P12). In contrast, users may expect tailored **legal opinions** and strategic advice. According to P7, what the user wants out of the answer may include "compliance or optimizing profits, or tax purposes," or "step-by-step instructions" based on predictive assessments of outcomes ("Can I win?"). Finally, users may desire **additional insights** beyond legal matters (P3, 13, P14). P13 noted the need to emotionally support users by extending empathy, support, and acknowledgment. In one case involving a neighbor's trespassing, P14 suggested home protection measures such as installing dashcams and getting dogs, not just legal recourse.

3.3.3 AI Systems' Capabilities

Participants raised five critical dimensions related to the technical capabilities and constraints of state-of-art AIs. The transient, AI-specific limitations may shift substantially with ongoing advances of research and development, unlike other categories that rely on users' needs and contexts. Throughout the discussion, experts disagreed at times: some were more optimistic about future development, while others believed that issues like hallucinations might persist.

- **Accuracy**. A key concern raised by multiple participants was the accuracy of AI-generated legal infor-

mation (P1, P3, P7, P8, P11, P13). P1 stressed the evolving nature of law, noting "We don't know if the law has changed from yesterday." P7, P8, and P13 stressed that serious hallucination issues that caused a New York attorney to be sanctioned for citing ChatGPT-generated cases [401]. Only P11 offered a more positive view: "There is a hallucination issue. [But] you could work with a plug-in, or a vector database where you had all this stored. If you could do that reliably, that would be a very good user experience."

- **Context-awareness**. Experts questioned AIs' capacity to move beyond static recommendations to context-dependent, adaptive guidance tuned to users' unique constraints and environments (P8, P10–12, P18, P20). As P11 noted, eligibility criteria such as demonstrating terminal illness often rely on specific circumstances. Additionally, procedural legal navigation "is not something you can predict by observing…a large data set" (P12). Others critiqued the staleness of training data, arguing that models cannot "address the local context" (P10, P13) as each situation has "idiosyncratic" details (P18). However, P20 countered that with enough data, models could likely outline standardized advice and steps applicable to various types of users.

- **Confidentiality**. Experts extensively discussed confidentiality risks (P4, P7–9, P12, P14, P16), which can be differentiated in a practical and normative sense. From a practical perspective, experts warned against an AI system's accidental leak of sensitive information (P4), highlighting the potential for unintended breaches of confidentiality. From a normative perspective, unlike attorney consultations, conversations with AI systems typically lack privileged protections against discovery in legal proceedings (P9). Attorney-client privilege does not extend to communications with third parties, and AI system providers (e.g., OpenAI) are obligated to produce relevant documents when served with a valid subpoena. Even if an AI system operates locally, chat records would likely remain unprotected unless a specific rule shielded the information from disclosure. As a result, users' admissions of illegal acts in AI conversations could thus become accessible to adversaries or prosecutors. P12 cautioned that proper warnings are necessary to inform users that AI conversations lack confidentiality protections and could be obtained by others with a court order.

- **Accountability**. Unlike attorneys, AI systems currently sidestep professional accountability for faulty advice (P8, P16–18). While lawyers' strict code of conduct and negligence liability apply even to

informal suggestions (P17), AI systems evade responsibility either through intermediary immunity laws or non-negotiable disclaimer clauses committing users to bear potential damages (P8, P16). Participants emphasized accountability gaps compared to attorney standards that leave users vulnerable if reliant on AI guidance. Given this gap, P18 argued that uncontrolled AI advice effectively constitutes illegally unauthorized practice of law (UPL).

- **Bias**. The experts expressed concerns that AI systems might reproduce structural stereotypes and discrimination (P5, P10, P13, P17, P20). They cautioned that the aggregated data used to train these systems could gradually skew the AI's performance to favor majority demographics unless measures are taken to actively protect minority views (P5). Given that English-written data predominantly represented in training datasets, experts noted that AI responses may disproportionately reflect the values and perspectives of English-speaking populations (P8).

3.3.4 Social Impacts

The experts considered two aspects of the possible effects that AI-generated responses could have on users and society. The first aspect focuses on the individual user who seeks guidance, taking into account the emotional, ethical, and cultural factors that may be affected by the AI responses. The second aspect extends beyond the individual, considering the broader impacts on third parties and society as a whole.

- **Impact on users**. The experts found that AI systems could potentially weigh the possible downsides, including those that the user may not have considered as something that could harm them, such as emotional effects or potential consequences in the workplace or on relationships (P4, P13, P20). P4 emphasized the need for "guardrails" around emotional prompts like questions including self-harm components. P13 cautioned that influencing a user's emotional state is highly problematic absent oversight, given the risks of uncontrolled bias and manipulation. P20 noted that what feels morally neutral in one culture may feel problematic in another, especially for minority groups.
- **Impact on others**. The experts considered "consequences for other people" who were not direct users as a serious concern (P6, P10, P17). These consequences include risks to indirectly affected third parties, such as explicit bias and stereotypes in advice, ensuing impacts of how advice is interpreted and acted upon, and the long-term assimilation of values. P6 emphasized the potential for unintended

consequences on vulnerable groups, using the example of how advice in harassment cases could further victimize previously affected individuals. Meanwhile, P17 highlighted broader ethical considerations beyond just technically accurate guidance, including assessing scenarios that create harm despite the good intentions of the advice.

3.4 Experts-Preferred AI Response Strategies

The aspects mentioned in Section 3.3 are illustrative of the complex considerations involved in AI legal advice. This section uncovers disagreements among experts through a quantitative and qualitative analysis of our workshop data, as we observed varying perspectives on balancing safety, ethics, and helpfulness.

3.4.1 Quantitative Results

Participants were asked to identify their preferred AI response strategies by choosing one of our seven provided strategies or producing their own. The resulting distribution, as shown in Figure 3.4, resembles a loose bell curve, with strategies ranging from the least interactive (content warning and outright refusal) to the most personally-tailored recommendations.

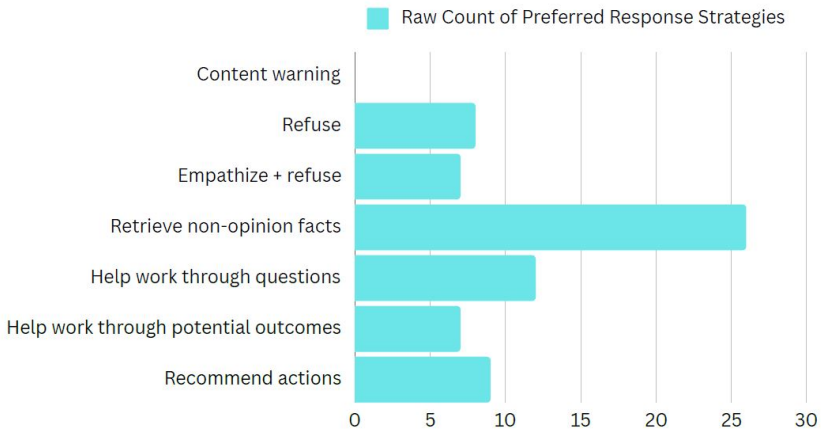


Figure 3.4: Expert-Preferred Response Strategies

This distribution reveals that experts preferred **information-focused responses** that avoid giving defini-

tive judgment. The strategies at the extremes of spectrum, namely 'content warning,' which received no votes, and 'recommend actions,' which received few votes, were less favored by the experts. The concentration of votes in the middle of the distribution suggests that experts prioritize providing users with relevant information while refraining from offering explicit recommendations or opinions, striking a balance between assisting users and maintaining the AI system's role as an informative tool rather than a decision-maker.

Further analysis revealed an intriguing relationship regarding experts' familiarity with AI systems and their receptivity to more tailored and detailed system responses. Regression testing showed a significant positive correlation ($p < .05$) between their self-reported general AI usage and openness to more customized and detailed output. Further statistical details of our regression test can be found in Appendix A.3.

3.4.2 Qualitative Results

Our qualitative analysis revealed rich and nuanced discussions behind the categorization of desired AI responses. Experts delved into the complexities of distinguishing between legal information and opinion and the challenges of ensuring user protection while leveraging the capabilities of AIs. They emphasized the importance of transparency, user safeguards, and adherence to legal traditions and frameworks. Moreover, participants recognized the potential of multi-turn interactions to help users better articulate their legal issues and access relevant information. The following sections present a detailed analysis of these qualitative findings, organized around the central themes that emerged from the workshop data.

Legal Information vs. Opinion As Figure 3.4 shows, most experts condoned offering pertinent legal information, while expressing reservations about AI providing a legal opinion due to reasons such as insufficient AI capabilities or user protection. What is the exact difference between information and opinion? Our participants suggested several principles to avoid providing a legal opinion.

- **Refrain from making definitive judgments about the legal consequences of a specific case.** Providing relevant laws is fine (e.g., driving under the influence (DUI) is illegal in Washington), but applying it to specific user situations constitutes opinion (e.g., falling asleep in the driver's seat in the parking lot after drinking alcohol could be a DUI) (P2, P13, P17, P19).
- **Do not recommend actions.** The system should avoid advising particular steps users should take. (P7,

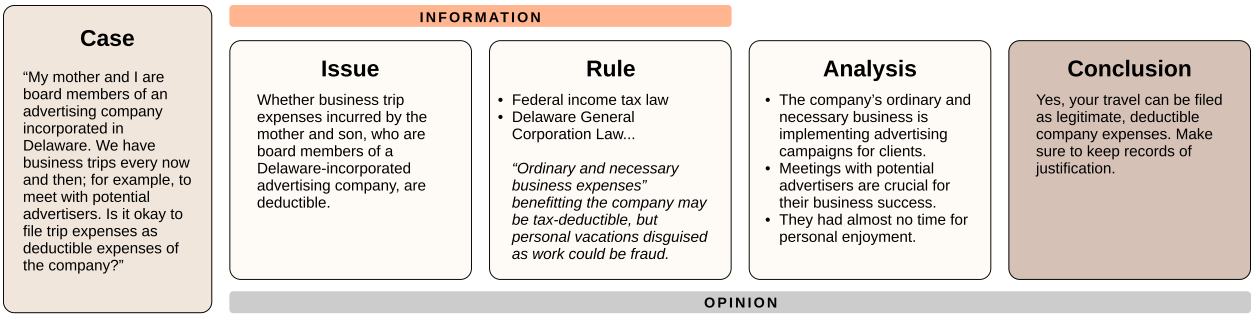


Figure 3.5: Applying IRAC analysis to One of Our Cases

P13)

- **Do not give predictions.** The system should not estimate a user's probability of winning a case or speculate on potential rulings. (P9, P12, P13, P19)
- **Do not provide a cost-benefit analysis.** The system should avoid any analysis that weighs the risks versus the rewards of a certain behavior. (P15, P16)

In essence, legal opinion encompasses interpretive, judgment-driven analysis that is often value-laden and forward-looking, whereas legal information involves reporting objective laws and past rulings without subjective assessment. To understand this distinction, we can draw upon the widely-used legal analysis tool known as IRAC (Issue-Rule-Analysis-Conclusion) [283]. IRAC entails (1) identifying the legal issue, (2) stating the rule that applies, (3) analyzing how the particular facts interact with the stipulations of the rule, and (4) finally deducing the conclusion [33]. Our findings suggest that AI systems focusing on issue and rule identification provide fact-finding "information," while analysis and conclusions may cross into tailored "opinion," as illustrated in Figure 3.5. However, it is important to acknowledge the complexity of distinguishing between legal information and opinion, as the line between the two can often be blurred in practice, as exemplified by cases such as *Grievance Comm. of Bar v. Dacey* [21], where the court found that publishing a booklet providing trust and tax information crossed the line into unauthorized legal opinion. This demonstrates that the distinction between legal information and opinion is not always clear-cut, and careful consideration must be given to the specific context and the information provided by AI systems.

Beyond Search Engines: Multi-turn Interactions for Refining Questions While cautioning against detailed legal opinions, participants suggested that AIs could offer a better user experience compared to

traditional search engines. P20 noted that users would not find it helpful if AI systems "vomit a whole lot of knowledge." The most promising and heavily-discussed possibility during the workshops is leveraging **multi-turn interactions**, allowing AIs to ask follow-up questions and clarify users' legally meaningful inquiries. This idea emerged as participants expressed frustration with missing case facts: "I don't think there's enough information to go off of, and that depending on the details that come out, it could change the analysis, therefore the outcomes." (P13) Participants emphasized that legal contexts are inherently complex (P11), and lawyers often spend considerable time eliciting relevant facts and identifying the "right questions to ask" (P12). They felt that AI-mediated dialogues could streamline time-consuming processes such as "screening interviews" (P12), "first calls" (P14), or "intake meetings" (P15). By engaging in multi-turn interactions, AIs could help users refine their questions, focus on key aspects of their cases, and seek relevant expertise. However, some warned that AI developers should exercise caution when eliciting extensive personal information from users, given confidentiality concerns (P13, P16). While identifying legal issues and relevant rules likely falls within the realm of permissible legal information, the line between information and opinion remains blurred. P16 argued that narrowing down factual patterns and applying rules engages deep judgment, stating "You're starting to make the AI become your lawyer."

Other Guiding Principles The experts suggested several principles for providing AI legal guidance. Some principles directly align with emerging literature on transparency [237], user satisfaction [238], and cautions about anthropomorphism [358, 271]. The principles outlined below represent the most prominent and frequently discussed ideas that emerged from the expert discussions.

- **Don't Pretend to Be Human**: An AI system should not behave like a human and cause misrepresentations, as that can create issues around transparency, over-reliance, and managing user expectations.
- **Caveat Constraints**: An AI system should provide various caveats on its limitations, such as that its capabilities are constrained, the conversation is not privileged, and that it is working off of incomplete information.
- **Avoid Potential Harm**: An AI system should refrain from providing recommendations that could potentially cause harm to users or others. This includes avoiding guidance that may lead to harmful real-world actions, as well as minimizing the risk of emotional or psychological harm that could result

from the system's responses.

- **Respect the Justice System**: An AI system should not provide information or advice that assists users in violating the law or evading legitimate oversight.
- **Avoid Unethical Answers**: An AI system should not make any outputs that could promote dishonesty, deception, fraud, impersonation, or other unethical behaviors that could get users into trouble.
- **Be Transparent**: An AI system should be able to explain the outcome it generated and point to the specific areas of datasets it relied on.
- **Avoid Appearance of Impropriety**: The appearance of impropriety refers to a situation that may appear corrupt or unethical to an impartial observer. For example, the AI system should not endorse or promote its creators, the AI companies involved in its development, or any other entities that could be perceived as influencing the system's outputs. The AI responses should be objective and impartial, focusing solely on providing accurate and helpful information to users.

3.4.3 Summary of Results

Our analysis uncovered 25 distinct dimensions to ensure safe and effective AI legal advice, spanning four key categories: (1) user attributes and behaviors, (2) query characteristics, (3) AI capabilities, and (4) social impacts. The experts deliberated with each other and through points of consensus to produce this rich set of considerations. However, experts expressed limited consensus on *how* AI systems should actually respond, given these nuanced factors. Some remained resistant to any AI involvement in legal questions, while others envisioned more helpful AI assistance models that increase access to information. Most debates centered on distinguishing information from opinion, and the majority felt that providing factual legal information was appropriate. Some participants suggested using AIs' conversational capabilities to help users refine questions and identify relevant laws through follow-up questions, similar to initial consultations with attorneys.

3.5 Discussion

Constructing AI policies does not exist in a technocratic silo. Rather, it demands a cross-disciplinary approach that synthesizes insights from domain experts. Our research demonstrates that engaging legal experts

in case-based deliberation is an effective method for translating professional knowledge and clinical experience into a concrete set of considerations for AI policies. The 4-dimensional framework we have developed provides a useful analytical lens that can be applied to future exploring AI policies in the legal advice and other professional contexts. Through this approach, we argue that policymakers can derive valuable insights to inform AI policies grounded in the centuries-old wisdom and experience of the legal profession, while also accounting for the challenges presented by AI technologies.

3.5.1 Benefits of Case-based Deliberation Methods

Our research process underscored several advantages of grounded case deliberation for eliciting expert considerations. Preparing realistic scenarios, while laborious, proved invaluable in quickly engaging experts with targeted queries related to their clinical experience. The cases allowed experts to examine granular concerns around singular situations, as well as overarching technical and legal constraints, producing a more concrete set of contextual factors for AI developers, beyond theoretical and high-level principles in prior research [368, 77, 188]. Finally, the collective deliberation itself revealed critical hidden dimensions and elicited justifications that shed new light on existing dimensions. As experts built on each other's points, they realized that overlooked issues or limitations in their own initial analyses. This interplay sharpened considerations and revealed nuances around balancing risks and benefits in varied situations. The combination of realistic cases and collaborative discourse resulted in more fine-grained, practice-informed insights compared to de-contextualized surveys or high-level principles.

3.5.2 Charting Novel Legal Considerations

One of our contributions is to shed light on existing legal and ethical barriers to AIs' legal advice which have been overlooked in the literature. Section 3.3.3 reveals that users lack confidentiality and accountability protections governing attorney advice: Conversations with AI systems risk disclosure in legal proceedings and inaccurate guidance evades professional negligence liability. Moreover, as Section 3.4.2 explains, UPL regulations prohibit non-lawyers from offering advice in many states, carrying criminal penalties. To circumvent the current legal risks , one could imagine AI systems designed like private counsels advising single parties, rather than serving all users uniformly like ChatGPT. In such case, AI systems could come to resemble pro-

prietary services, with corresponding confidentiality and liability assurances. However, the traditional legal conservatism may change in the future as UPL rules have already faced criticism for limiting affordable access to legal help [149, 332, 81]. The EU AI Act's categorization of AI legal assistance tools as "high-risk," which subjects them to heightened responsibilities instead of banning them outright, may speak to this potential shift [163].

3.5.3 Learning from Time-Tested Wisdom

Leveraging accumulated expertise in professional communities can help sidestep painful mistakes [77]. In our research, we found that UPL regulations do constrain the possibility of AI providing legal advice. However, UPL also offers long-standing criteria for distinguishing between legal information and legal opinions. Historically, merely providing legal information has not been considered a violation of the UPL [23, 149, 6]. For example, the Texas Court provides guidelines for court staff and illustrative examples like in Table 3.2. The list, compiled from Texas law clerk resources, distinguishes between permissible and forbidden questions that Texas court personnel can answer. This approach is remarkably similar to a process used in AI development called "red-teaming." Red-teaming in AI context refers to the practice of deliberately attempting to find flaws, vulnerabilities, or potential misuses in AI systems [188, 73]. Just as court systems identify questions that could lead to improper legal advice, AI developers use red-teaming to identify potentially harmful or inappropriate user inputs or system outputs.

Furthermore, legal scholars have explored legally justifiable AI advice under UPL, attorney-client privilege, and other doctrines [208, 403, 370]. Wendel stated that the "core lawyering functions" such as recommending the course of action or drafting contracts cannot be delegated to AI agents due to technical limitations and accountability deficits [403]. This demonstrates how principles accumulated over centuries of legal scholarship now inform responsible AI systems and the call for cross-disciplinary collaborations.

3.5.4 Applicability to Other Professional Domains

While each possessing unique dimensions, domains like medicine, mental health, law, and finance share common threads around high-stakes real-world impact and historical reliance on licensed specialists for advice. We believe that our research methods and 4-dimension framework give illustrative guidance to

Table 3.2: Examples of Impermissible Questions that Require Legal Opinions [41].

Type	Permissible questions	Impermissible questions
Procedure	Can you tell me how to file a small claims action?	Can you tell me whether it would be better to file a small claims action or a civil action?
Definition	What does "certificate of service" mean?	My neighbors leave their kids at home all day without supervision. Isn't that child neglect?
Forms	I need to file for divorce and I have no idea where to begin. Is there some place I can go to find out how to get started?	The self-help divorce petition says I should list any gifts as my separate property. Should I list the money that my parents gave me last month as my separate property?
Options	What can I do if I cannot afford to pay the filing fee?	My ex-husband hasn't paid the debts that he agreed to pay in our divorce settlement. Can I be made responsible for this debt?

further research in other professional domains. As this research demonstrates how case-based deliberation methods can unravel complex professional ethics, researchers could adopt similar processes engaging mental health counselors, financial advisors, or medical professionals. Tapping into the clinical experience and integrity of practitioners through structured deliberation based on realistic cases can help produce tailored dimensions and guidelines for responsible AI advice respective to each profession. Building upon this foundation, our 4-dimension framework—(1) User, (2) Query, (3) AI, and (4) Impact—could be adapted and applied across various professional domains. The (1) User, (2) AI, and (4) Impact dimensions can be applied in other domains with minimal modifications. However, the (3) Query dimension requires more customization to address the typical requests of clients, terminology, and desirable practices in each field.

3.5.5 Limitations and Future Research

Our study has several limitations. First, our expert sample predominantly focused on practitioners familiar with the US legal system. Ethical considerations around appropriate AI assistance may differ across different legal systems and cultures. Second, our participants' responses are conditioned by their prior experience with state-of-the-art AI technology, such as ChatGPT empowered by GPT-4. Experts' evaluations of the appropriateness of AI legal advice may evolve in the future, based on technological innovations, which could be an avenue for future research. Third, we did not engage end-users like clients of legal services. Future work can specifically investigate end-user perceptions to compare and contrast with our expert-informed re-

sults. Finally, while our taxonomy conceptualizes a concrete set of dimensions, how these dimensions could change the appropriateness of AI responses remains unexplained. This may require larger-scale empirical analysis on public assessments across diverse pairings of cases and responses.

3.6 Reflection on Co-Design and Democratic Inputs in AI Governance

The experience with our case study and the broader Democratic Inputs to AI program has led me to critically reflect on the nature of participation in AI governance. Participation in AI governance, as we have observed, often oscillates between what Kelty describes as an "optative mood"—an enthusiastic belief in the transformative power of participation—and a "critical mood" that denounces existing efforts as phony or exploitative [233]. There is often an optimistic push for increased public involvement, followed by critiques of "participation washing" when these efforts fail to produce substantial changes. Our project, while demonstrating effective ways of engagement through expert consultation and case-based reasoning, also encountered the limitations and contradictions inherent in participatory approaches.

Our research successfully gathered valuable input from legal professionals under time constraints, showcasing an efficient methodology for expert consultation in rapidly evolving technological fields. One significant outcome of our study was the increase in awareness among participants of the complex value judgments faced by AI policy makers. As experts grappled with queries and engaged in mutual discussions, they gained insight into the challenges of balancing diverse perspectives and interests. This process appeared to empower professional communities, fostering a deeper understanding of AI governance issues. Practitioners could identify practical challenges and ethical concerns that might not be apparent to policymakers or AI developers alone.

However, the challenge of translating these insights into lasting and impactful policies remains. First of all, our approach was resource intensive, and each participant was compensated at $100 per hour. Scaling such a model presents obvious challenges, highlighting the need for more cost-effective methods of meaningful engagement. Although our focus was on effectively aggregating expert opinions, we encountered the challenge of reconciling contradictory elements to create a comprehensive representation. This highlights a key tension: the desire for open discourse and diverse views often conflicts with the need for coherent, implementable policies.

Reflecting on Arnstein's ladder of participation, I recognize that our approach falls short of the full "partnership." We reached the "consultation" level, but there was room for implementation of experts' insights into ongoing governance processes. It remains questionable how to ensure that tech companies would faithfully implement expert input, especially when it conflicts with commercial interests.

The experience of participating in the Democratic Inputs to AI program provided further insight into the complexities and challenges of democratizing AI governance. The program, which funded ten diverse teams to explore various approaches to democratizing AI development, revealed a wide range of interpretations of what constitutes "democratic inputs." Some teams focused on marginalized communities, others focused on the broader public, and some, like ours, focused on specific professional groups.

However, this experience also exposed limitations in current approaches to democratizing AI. Many projects, despite their creative ideas, remained largely in the ideation stage, struggling to move beyond conceptual frameworks to practical implementation. This gap between theory and practice raised questions about the real-world applicability of these democratization efforts. Moreover, I observed that many projects focused heavily on the medium of participation, such as developing AI moderators to orchestrate deliberations. The flexible nature of these technology-based participatory schemes, while allowing for easier critique and modification, raised important questions about their long-term impact and sustainability.

It has prompted me to question both the normative desirability and the practical feasibility of fully democratizing the process of extracting final policies from participatory inputs. Although the ethos of democracy might suggest that more participation is always better (such as the Arnstein's Ladder assuming the upper level is better), the reality of AI governance is far more complex. The concept of "participation washing"—engaging in superficial or insincere participatory efforts—is indeed a valid concern. However, I argue that the solution is not simply to demand more participation or to criticize efforts that fall short of full democratization. Such a view oversimplifies the complex nature of AI governance and may, paradoxically, hinder effective policy-making.

Internal policy-making in AI governance invariably involves value judgments and the need to prioritize conflicting values. These decisions often require nuanced judgment calls that must be made by the service providers who are ultimately responsible for market outcomes and legal compliance. The push for maximizing participation, while well-intentioned, may overlook the fact that these providers possess unique insights

into the technical constraints, market dynamics, and regulatory landscape that shape AI development and deployment. This perspective is not to diminish the value of public input or expert consultation, but to recognize the complex realities of translating diverse inputs into actionable policies within specific organizational and regulatory contexts.

This perspective challenges the notion that maximizing participation or achieving full democratization should be the ultimate goal of AI governance. Instead, it suggests a more nuanced approach that recognizes the vital role of expert judgment and the realities of corporate responsibility in the AI ecosystem. The aim should be to foster meaningful and strategic participation that genuinely informs and improves AI governance, while acknowledging the necessary role of service providers in making final, accountable decisions.

Chapter 4

Case Study 2: Co-Regulating Online Content in South Korea

4.1 Background

This study examines how South Korea's unique co-regulatory approach to online content moderation, which involves collaboration among industry stakeholders, government, and civil society, can provide lessons for the development of multi-stakeholder AI governance models that balance competing interests and adapt to evolving challenges.[1]

Currently, AI governance relies on a patchwork of conventional regulatory approaches, including limited hard laws in the EU and China, divergent court cases in the US, and companies' internal policies and contracts. However, AI systems are incredibly complex and no single entity possesses the ultimate expertise or legitimacy to unilaterally make and implement value-oriented safety norms. As AI systems continue to develop and their impact broadens, the need for coordinated multi-stakeholder efforts to ensure AI safety and fairness is likely to grow. In light of these challenges, alternative governance models that can effectively manage complex technological systems become increasingly relevant.

Given that AI is still evolving and its future trajectory remains uncertain, discussions about co-regulatory

[1]This research was conducted in collaboration with Prof. Pardis Emami-Naeini and Prof. Tadayoshi Kohno. They provided guidance on research methodology, qualitative analysis, and focus refinement. Their insights were key in contextualizing the South Korean work within broader online harm mitigation strategies.

governance in AI—which requires significant time and resources to steer and coordinate differently-situated stakeholders towards a common goal—have not yet gained prominence. In this context, learning from time-tested wisdom in adjacent fields can save valuable time and help identify key considerations upfront, potentially avoiding costly mistakes. This is where the examination of South Korea's co-regulatory system for online content proves particularly valuable, serving as a springboard to inform the development of AI governance frameworks.

While the contexts of content moderation and AI governance differ, both fields face similar challenges in balancing diverse stakeholder interests, adapting to rapid technological changes, and negotiating the boundaries between individual rights and collective safety. The lessons learned from South Korea's approach to content moderation may offer valuable perspectives on how to structure collaborative, adaptive governance systems for AI. By drawing these parallels, this study aims to contribute insights that could help shape future approaches to governing complex technological systems in diverse global contexts.

The concept of online content co-regulation may be unfamiliar to Western readers, as scholarship on content moderation has predominantly been West-centric. This research has typically focused on global platforms like Twitter and Facebook [88, 397], Google [258], Wikipedia [399], and LinkedIn [321], exploring governance models spanning community self-moderation [399, 361], crowd-sourcing [384, 152] or participatory systems for marginalized groups [346, 89, 292, 339]. However, co-regulatory models that involve both public and private actors in implementing universal rules across different platforms have struggled to gain significant interest due to varying free speech laws and internet regulations across countries, as well as platforms' reluctance to relinquish decision-making power to external bodies.

This Western perspective on content moderation may not align with collectivist cultures found in countries such as South Korea, Singapore, and China [382, 311, 70, 353]. These societies often balance individual speech against common interests, leading to different perspectives on the roles of government and corporate entities in content governance. In South Korea, this cultural context has fostered a unique co-regulatory approach to online content moderation. This model involves cross-platform bodies that create and implement collective rules with the support of the government and civil society, reflecting a broader social emphasis on collaborative decision-making and shared responsibility. The lessons learned from South Korea's approach to content moderation may offer valuable perspectives on how to structure collaborative governance systems

for AI. This research addresses the following questions:

- What is online content co-regulation, and how does it fit within the broader context of content moderation scholarship?
- How does South Korea's online content co-regulation operate in practice, and what factors contribute to the varying outcomes of co-regulatory systems across different content industries?
- What lessons can be drawn from South Korea's co-regulatory model for the development of multi-stakeholder AI governance frameworks?

To address these questions, this study combines an analysis of content moderation and collaborative governance literature with an in-depth qualitative investigation of two distinct industries: online news and web comics (known as "Webtoons" in South Korea). These two industries were selected due to their significant cultural impact, widespread consumption, and contrasting relationships with major online platforms, offering rich insights into the dynamics of content co-regulation in South Korea. Our methodology integrates semi-structured interviews with 15 key stakeholders, including full-time content creators, government officials, and platform executives. This approach allows us to explore the factors that shape governance efficacy and outcomes under different conditions.

The study's theoretical framework is grounded in Ansell & Gash's work [76] (Fig. 4.1), prominent in the public policy discipline, which provides a structured approach to analyzing the complex dynamics of collaborative governance. By applying this model to South Korea's co-regulatory systems, we aim to identify key factors contributing to successful collaborative governance in content moderation.

This study makes three key contributions. First, it de-centers Western individualistic assumptions in content moderation scholarship through an in-depth qualitative study of collective governance in South Korea. By applying Ansell & Gash's theory [76] to real-world data, it assesses the operationalization of readily existing theories of collaborative governance principles for content moderation. Second, it provides a more complex and nuanced understanding of the balance of free expression and harm prevention in collaborative processes. Other than institutional design principles—e.g., stable funding, inclusion of members—our study found that trust among industry stakeholders and normative justifications (no challenges in courts, government validation) were paramount, even outweighing idealized designs.

Third, it discusses implications and recommendations for researching and developing AI governance

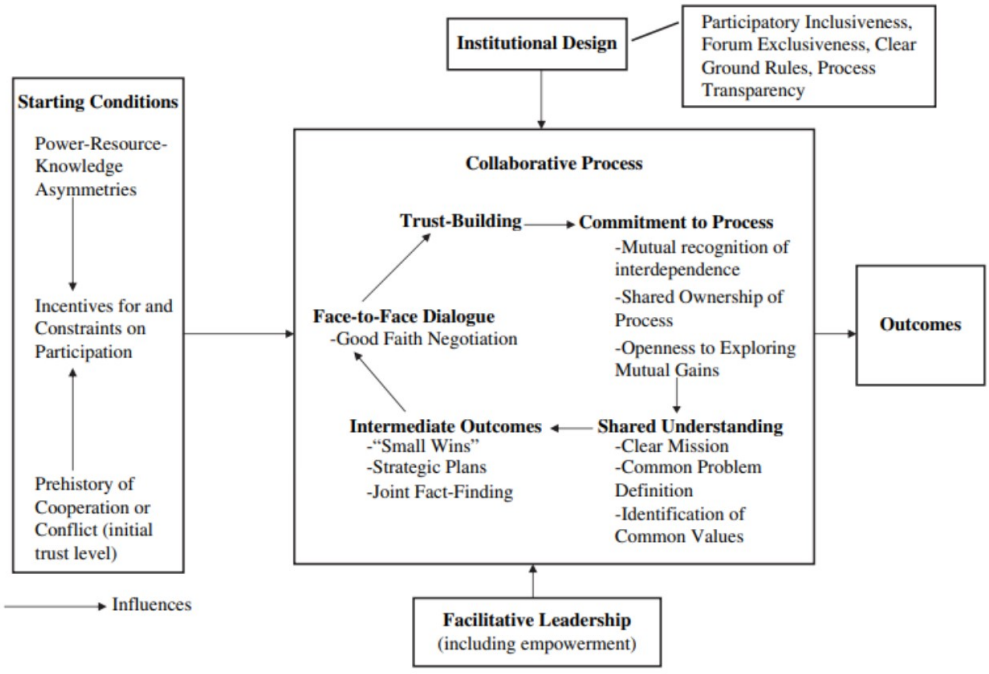


Figure 4.1: Ansell & Gash's model of Collaborative Governance [76]

.

models that go beyond the current focus on individual companies or government regulations. By examining the successes and challenges of South Korea's co-regulatory approach to online content moderation, this study highlights the potential for multistakeholder collaboration in AI governance, which could contribute to balancing competing interests, ensuring public trust, and addressing the evolving challenges posed by AI technologies. The insights from this study can inform the design and implementation of AI governance frameworks that prioritize collective decision making, building trust among stakeholders, and the establishment of shared norms and values.

4.2 Methods

This study employs a multi-method approach to investigate the co-regulation of online content in South Korea, focusing on the news and web comics industries. Ansell & Gash's collaborative governance framework provides a theoretical lens through which to analyze the complex dynamics of multi-stakeholder governance. To demystify co-regulation and contextualize this study, we developed a typology of content moderation approaches, identifying four main categories: multistakeholder moderation (including Co-regulation),

platform-wide moderation, community-driven moderation, and user-level moderation. Within this framework, we defined co-regulation as a governance model where multiple platforms delegate decision-making roles to a cross-platform body with explicit or implicit government support. This approach stands in contrast to platform-centric or purely user-driven moderation strategies.

Our methodology integrated in-depth interviews with 15 key stakeholders (including content creators, platform executives, co-regulatory board members, and government officials), field observations, document analysis, and comparative historical analysis. By applying Ansell & Gash's theory to our empirical data, we examined how starting conditions, institutional design, facilitative leadership, and collaborative processes interacted to shape the effectiveness of co-regulatory efforts in South Korea's unique cultural and political context.

4.2.1 Research Focus: Online News and Web Comics

Online news and web comics were chosen as the focus of this research for several reasons. First, these two content types are among the most widely consumed in South Korea, with people often seen scrolling through web comics and online news during their daily commutes. Web comics, in particular, have become a significant part of popular culture across demographic segments, as evidenced by Naver Webtoon's successful IPO on the NASDAQ in 2024. Second, both news reporters and comic artists are keenly aware of and value their freedom of expression, making them interesting subjects for studying content moderation and co-regulation. Third, both news and comics have raised societal concerns: misinformation and branded content in the case of news, and violence or nudity in comics, which have drawn criticism from child safety advocates due to their popularity among youth.

Fourth, both online news and web comics are hosted by major online platforms, Naver and Kakao, but these platforms have formed contrasting relationships with news institutions (adversarial) and comic artists (collaborative). This contrast provides a meaningful opportunity to explore how different industry-specific conditions affect the outcome of co-regulation. Finally, both industries are heavily involved with the Ministry of Culture, Sports, and Tourism, where I worked as a deputy director for four years. It allowed me to leverage my personal network to reach insiders representing four different perspectives within each co-regulatory system: creators, platforms, government officials, and co-regulatory boards. This ac-

cess to diverse viewpoints provided a nuanced understanding of the challenges and opportunities in content moderation and co-regulation within these industries.

4.2.2 Theoretical Framework

Collaborative governance has a rich interdisciplinary history spanning law, public policy, political science, and economics, exploring how to align diverse stakeholders in pursuit of the collective good [299, 298, 305, 306, 156, 303, 110, 393, 162, 295]. For example, Ostrom's work [304, 305, 306] defines governance as jointly determined norms and rules shaping group behavior, addressing how local self-governing institutions balance individual interests with sustaining cooperation. While most applications of collaborative governance have focused on common pool resources such as water and energy [304, 156, 303, 295, 162], emerging research is now exploring its application to digital governance, including content moderation [309, 371, 245, 155, 293, 160, 285].

Among various theories on collaborative governance, this study finds Ansell & Gash's model [76], depicted in Figure 4.1, particularly useful for understanding stakeholder collaboration in policy-making and implementation. This model considers power dynamics among stakeholders and emphasizes trust-building, allowing us to explore how these factors influence the effectiveness of South Korea's co-regulatory bodies. It also includes "prehistory of cooperation or conflict" as a starting condition, enabling us to consider the distinct stakeholder dynamics involved in shaping current co-regulatory efforts with respect to news and web comics. While widely applied in various fields, this model's application to online content moderation, particularly in non-Western contexts, remains limited. Our study aims to bridge this gap by applying Ansell & Gash's framework to online content co-regulation in South Korea, focusing on four key variables that influence the success of collaborative governance:

1. **Starting conditions**: These include power-resource-knowledge asymmetries, incentives for and constraints on participation, and the prehistory of cooperation or conflict.
2. **Institutional design**: This encompasses participatory inclusiveness, forum exclusiveness, clear ground rules, and process transparency.
3. **Facilitative leadership**: Strong leadership is crucial in bringing stakeholders together, setting and maintaining clear ground rules, building trust, facilitating dialogue, and exploring mutual gains.

4. **Collaborative process**: This is the core of the model, involving face-to-face dialogue, trust building, commitment to the process, shared understanding, and intermediate outcomes.

In the realm of online content moderation, collaborative governance faces unique challenges. Unlike traditional common pool resources, online speech involves complex issues of individual rights, cultural norms, and rapidly evolving technological landscapes. The power dynamics among stakeholders in online content governance differ significantly from those in traditional static governance scenarios. Tech platforms wield considerable influence, often transcending national boundaries, while governments grapple with regulating entities that operate beyond their jurisdictions. Civil society organizations and individual users, despite being key stakeholders, often find themselves at a power disadvantage.

4.2.3 Interviews

From September 2021 to March 2022, we conducted in-depth interviews with fifteen individuals who are deeply involved in shaping and sustaining the co-regulation of online content. We began by reaching out to government agencies and personal contacts in the content industry, explaining our research interest in online content regulation, particularly in the procedures and structures of co-regulatory governance. Using snowball sampling, the initial interviewees introduced us to others in the industry, civil society, and academia who were knowledgeable and experienced in this field. The interviews were loosely-structured, not strictly following a specific order, however, general interview questions are available in Appendix B.

Table 1 presents the different categories of individuals interviewed and their roles in the co-regulation process. The interviewees included content creators, online platform executives, co-regulatory board members and directors, and government officials, each playing distinct roles in co-regulation. Among the interviewees, 2 were women, 11 were men, and 2 did not want to identify their gender.

4.2.4 Observations and Documentations

To supplement the interviews, we conducted observations and reviewed documentation from news outlets, relevant laws and regulations, legislative discourse, and public reports released by co-regulatory boards between 2021 and 2023. Public reports included minutes of board meetings, board-made rules and appendices, annual reports, news releases, and budgets. This process provided the thorough background information

Table 4.1: Co-regulation Contributors in South Korea (Total Count: 15)

Category	Participants	Roles	Main job at present
Content Creators	P1, P2, P9, P10	Make referrals to appoint co-regulatory board members; Gather creators' inputs to influence co-regulation decision-making; advocate for the freedom of expression; organize collective action if necessary	Broadcast or newspaper journalist; editor; web comics artist
Online Platform Executives	P5, P6, P12, P13	Oversee the safety teams or legal teams housed in major online platforms with millions of active users; represent platforms' interests in co-regulatory decision-making; often prioritize to settle user complaints quickly and smoothly; make financial contributions to co-regulation; view participation in co-regulation as compliance and PR efforts to avoid more stringent regulations	CEO; legal counsel; safety/content lead
Co-regulatory Board Members and Directors	P3, P4, P7, P11	Make decisions about rule-making and enforcement; board members receive hourly compensations for their time while having other full-time jobs; often struggle with lack of human and financial resources	Professor; lawyer; full-time director for the board
Government Officials	P8, P14, P15	Oversee the functioning of the co-regulatory board according to the law; respond to political pressure to be more stringent about online harms; sometimes allocate the annual budget to support co-regulation	Director general, director, and deputy director in government agencies

necessary to better understand how the board operates, how each case was resolved, and how the rules have been updated over time, which could not be fully covered by time-limited interviews. During fieldwork, some participants demonstrated their internal case handling systems, providing an additional opportunity to compare what our participants said in the context of the interview with their actual practices.

In 2022, a significant controversy emerged over how to combat misinformation and undisclosed branded content in online news. The existing co-regulatory board had dealt exclusively with defamation claims, not necessarily addressing the accuracy or financial sources of news articles. Various policy proposals were made during this time, allowing me to understand the public assessment of the current co-regulatory

structure and providing an opportunity to inquire participants about why the current board should or should not exercise regulatory controls to combat misinformation.

4.2.5 Analytical Approach

In addition to our primary data collection through interviews and observations, we conducted an extensive analysis of normative and practical challenges associated with co-regulation models. This analysis drew upon historical and contemporary examples from the United States and global contexts, allowing us to situate our findings within a broader theoretical and practical framework. Our analytical approach involved:

- Historical case studies: We examined past attempts at co-regulation in the media and technology sectors, including examples such as the National Association of Broadcasters (NAB) Code, the National News Council (NNC), and the Internet Content Rating Association (ICRA).
- Comparative analysis: We compared co-regulatory models across different jurisdictions, focusing on their strengths, weaknesses, and outcomes.
- Legal and policy review: We analyzed relevant legal frameworks, including First Amendment considerations in the US and the antitrust implications of cross-platform collaborations.
- Stakeholder incentive analysis: We examined the motivations and potential reluctance of different stakeholders to participate in co-regulatory frameworks.
- Practical implementation challenges: We identified and analyzed key operational challenges such as funding mechanisms, rule-making processes, and enforcement issues.

This approach contextualized our primary research findings within the broader landscape of co-regulation attempts and challenges. Reflecting on both successful and unsuccessful cases, this study gained insights into the potential effectiveness and limitations of co-regulatory models.

4.2.6 Ethics

This research received an exemption from the University of Washington's IRB, but we followed procedures for human research studies. All identifiable personal information was anonymized in the transcripts. Because some participants were very reluctant to reveal any traceable details due to non-disclosure agreement with their employers or the political sensitivity of issues, we have intentionally obscured some details that

the participants wanted to keep confidential. The participants were compensated with a 30,000 KRW gift card (22 US dollars) that exceeded the minimum hourly wage in South Korea (9,620 KRW). Some of them refused a gift due to their organization's policy. The participants were informed that they could stop or pause the interview at any time. After the interview, the complete transcripts were provided for them to review.

4.3 Situating Co-Regulation Within Content Moderation Paradigms

Online content co-regulation might be foreign to Western readers, especially in the US, where the free speech doctrine strictly forbids government intervention in content moderation. Content moderation is "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." [201]. Its primary aims are to facilitate cooperation among users and prevent abusive behaviors that may cause harm. To clarify and contextualize co-regulation within content moderation scholarship, this study categorizes content moderation into four types: platform-wide moderation, community-driven moderation, user-level moderation, and multistakeholder moderation. Figure 4.2 illustrates these four categories.

Platform-wide moderation. Individual platforms like Facebook, and Twitter set and enforce their own content rules [193]. The majority of large-scale platforms today use a combination of proactive algorithmic filtering, user-reported offensive content, and human moderation for oversight [167, 243, 87, 131, 421, 119]. However, some criticize the centralized control and arbitrary nature of platform policies. In 2005, Thomas Friedman famously depicted eBay as a "self-governing nation-state" armed with its feedback system and vigorous community norms [180]. Mark Zuckerberg concurred, "In a lot of ways, Facebook is more like a government than a traditional company." [173] Meta's Oversight Board is an interesting governance experiment within platform-wide content moderation that invites external experts to provide opinions on difficult matters and operates independently from the company's decisions. Meta voluntarily divest itself of part of its power to uphold procedural justice and build user trust, which was a long-term strategy for continued economic growth [245].

Community-driven moderation. As opposed to centralized, top-down, automated-technology-reliant content moderation, some platforms such as Wikipedia, Reddit, and Facebook Groups, have taken a more de-

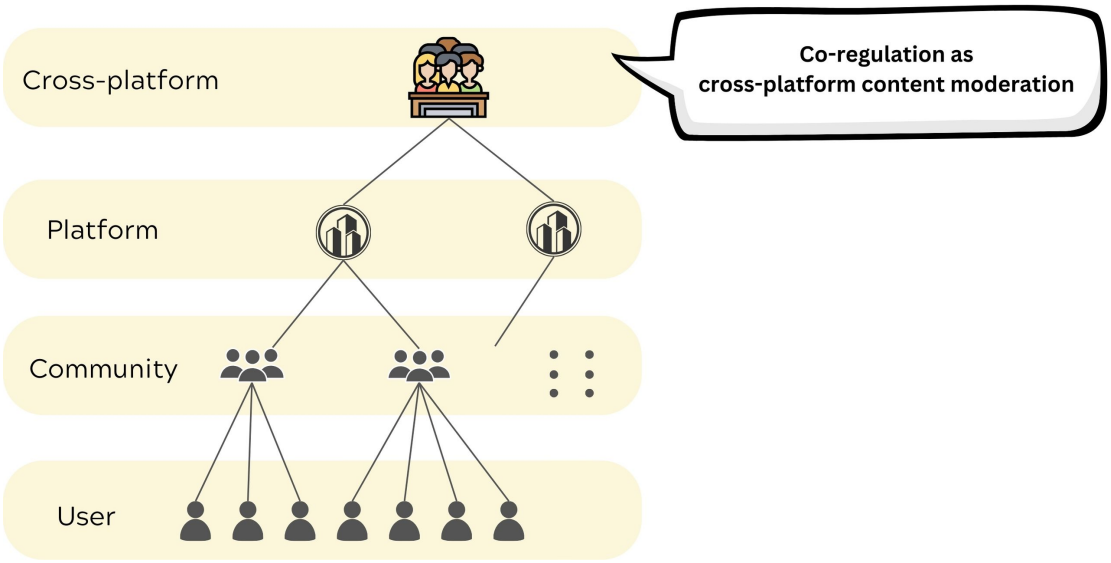


Figure 4.2: Typology of Content Moderation

centralized approach to moderation decisions, relying on their community members to monitor, report, and enforce content policies [399, 347]. Furthermore, in the online game Minecraft, or other federated services like Mastodon, users can join private servers that have their own rules and moderators [333, 167].

User-level moderation. Personalized moderation allows individual users to customize and filter content based on personal preferences through platform tools like Instagram Safe or third-party browser plug-ins [256, 257, 14]. The goal is to enable protection from harmful/unwanted content tailored to individual needs. An example tool called Squadbox was developed to enable crowdsourced filtering of incoming messages by trusted friends [276].

Multi-stakeholder moderation. This approach acknowledges that content moderation is not solely the responsibility of individual platforms but a broader issue that necessitates collective action. Thus, content moderation involves collaboration among various stakeholders, including platforms, users, civil society organizations, and/or government actors, to develop and enforce industry-wide content standards. When the formulation of such governance is facilitated by external force, it becomes similar to *meta-regulation*, reflecting the fact that "outside regulators deliberately–rather than unintentionally–seek to induce targets to develop their own internal, self-regulatory responses to public problems" [135].

There have been several attempts at collectively combating online harms. An example is the Internet Content Rating Association (ICRA), which encouraged website owners to voluntarily label their websites to help parents determine whether the site was appropriate for them or their children [78]. However, despite widespread backing from industry (Microsoft) and governments (US and EU) [30], the ICRA failed to achieve its goal and closed in 2010 [78] as it failed to convince its own members, not to mention hundreds of thousands of website owners, to make small changes on their websites for the social good. It possessed no enforcement tools, and the IRAC failed to answer the question "why bother?". Similar participant attrition also occurred with respect to self-regulation of online privacy [141, 213, 191].

There have also been more successful voluntary industry-led efforts. PhotoDNA is a digital image analysis technology developed by Microsoft in 2009 to create a unique digital "fingerprint" of an image, which can then be compared against a database of child sexual abuse material (CSAM), to identify and remove any matches. [280]. It has been widely adopted by tech companies (Facebook, Twitter, and Google) and nonprofit organizations (e.g., the National Center for Missing and Exploited Children) [154]. Similarly, Facebook, Microsoft, Twitter, and YouTube created the Global Internet Forum to Counterterrorism (GIFCT) in 2017 to combat terrorism and violent extremism online by creating shared database [386]. The organization has since grown to include many other companies, NGOs, and government bodies but there are arguments about private platforms' authority to define highly-contested types [281, 232] and the lack of transparency and public oversight in the GIFCT [390].

Co-regulation as multi-stakeholder moderation. Online content co-regulation is a form of multistakeholder moderation. In this model, cross-platform bodies make and implement collective rules with the involvement of multiple stakeholders, including government and civil society. Government support can take various forms, ranging from explicit to implicit. There may be laws or regulations that explicitly empower cross-platform bodies to make decisions on certain matters related to content moderation. Alternatively, the government may provide financial support to these bodies, ensuring their operational stability and independence from individual platform interests. In some cases, the government might establish a broad policy framework within which these bodies operate, setting general guidelines or objectives without directly controlling day-to-day decisions. There are also instances where the government's role might be more passive, implicitly endorsing the co-regulatory body's authority through recognition or cooperation.

This approach reflects a more collectivist cultural perspective that seeks to balance individual expression with societal interests. Unlike the strict separation between government and content moderation seen in some Western contexts, co-regulation acknowledges a role for governmental involvement in shaping the overall structure of online content governance, while still maintaining a degree of independence in decision-making processes. This model allows for a more collaborative approach to addressing complex issues of online content moderation, leveraging the expertise and resources of multiple stakeholders to create more comprehensive and culturally appropriate solutions.

4.4 Overview of South Korea's Content Co-Regulation

South Korea has developed a distinctive approach to online content governance, balancing free speech protections with efforts to mitigate potential harms in the digital space. This section explores the evolution of content regulation in South Korea, tracing its journey from historical censorship to the current co-regulatory mechanisms. We begin by examining the country's complex history regarding freedom of speech, including constitutional provisions and the impact of authoritarian regimes. The discussion then moves to the challenges posed by the internet era and South Korea's unique geopolitical situation. Finally, we introduce two co-regulatory approaches in prominent digital media sectors—online news and web comics.

4.4.1 Content Regulatory Landscape in South Korea

South Korea's journey towards freedom of speech reflects a complex interplay of historical, cultural, and political factors, presenting a stark contrast to Western nations with longer-established traditions of free expression [311]. The country's first free speech clause, enshrined in the 1919 Provisional Charter of Korea during Japanese colonial rule (1910-1945), marked a significant departure from the traditional monarchy of the Joseon dynasty, which did not recognize freedom of speech. Article 21 of the current South Korean Constitution states:

- (1) All citizens shall enjoy freedom of speech and the press, and freedom of assembly and association.
- (2) Licensing or censorship of speech and the press, and licensing of assembly and association shall not be recognized.

- (3) The standards of news service and broadcast facilities and matters necessary to ensure the functions of newspapers shall be determined by Act.
- (4) Neither speech nor the press shall violate the honor or rights of other persons nor undermine public morals or social ethics. Should speech or the press violate the honor or rights of other persons, claims may be made for the damage resulting therefrom.

However, it is noteworthy that Article 21(4) allows content-based regulations to prevent harm to others or uphold public morals, a provision that has provided the grounds for more restrictive speech policies compared to countries like the United States. South Korea's path to free speech has been fraught with various challenges. From the 1950s to the 1980s, the country experienced multiple military coups and authoritarian regimes, during which the constitutional guarantees of free speech were largely symbolic. Under these dictatorships, the government wielded extensive censorship powers, and numerous students, activists, and journalists faced arrest, sentencing, and torture for expressing pro-democracy views.

The year 1989 marked a turning point in South Korea's official "liberalization." This period saw a gradual relaxation of speech regulations, coinciding with the rise of the internet in the 1990s. However, this technological advancement brought new challenges such as the rapid spread of potentially harmful content online. Moreover, South Korea's approach to content regulation continues to be influenced by its unique geopolitical situation: the ongoing state of war with North Korea necessitates careful consideration of national security concerns. This balancing act is evident in the powers granted to bodies such as the Korea Communications Standards Commission (KCSC), a regulatory agency that can order the blocking of content deemed illegal, including material considered defamatory, dangerous to youth, or posing a threat to national security [59]. While free speech advocacy groups like the Electronic Frontier Foundation have criticized these "substantial controls on online communications" [417], Korean courts have generally upheld such regulatory measures.

The tension between advocates for free speech and those prioritizing online safety has led to a unique regulatory landscape in South Korea. While courts have validated various forms of government involvement in online content moderation—a stark contrast to the U.S. approach—there has been a growing recognition of the need for more nuanced and flexible regulatory mechanisms. This recognition comes from both traditional regulatory agencies, reluctant to directly intervene in online content and face public backlash,

and companies, wary of making individual decisions on politically sensitive matters. Underpinning this approach is a collectivist view that the "appropriateness" of content should be defined collectively, rather than by a single entity.

4.4.2 Co-regulation of Online News

During the authoritarian regime, newspapers were subject to government control over their reporting. The number of newspapers in the country has grown significantly—from 32 in 1987 [28] to 5,078 in 2022 [176]—since censorship of the newspaper publishing industry was eased. Since then, the government has demonstrated its commitment to preserving freedom of the press by excluding online news articles from the jurisdiction of the Korea Communications Standards Commission (KCSC).

Newspapers have operated several, unsuccessful co-regulatorty systems for over several decades: the Korea Press Ethics Commission (PEC) for larger media organizations and the Internet Newspapers Committee (INC) for smaller, online-only publishers. Both bodies are constituted of news industry associations and receive annual support of 7.5 billion KRW (about 600,000 dollars) from the Ministry of Culture, Sports, and Tourism (MCST). News associations are responsible for operating both the PEC and INC and neither the government nor online platforms involve in their businesses. Although these institutions have substantiated their efficacy in regulating election polls and graphic content, journalists have expressed dissatisfaction with the effectiveness of these entities, as the imposition of simple "warnings" is often ignored by media outlets. Some of them were not even aware of the existence of such entities. This concern led to calls from some members of the National Assembly to defund both the PEC and INC.

In 2020, the National Assembly voted on a proposed law that aimed to introduce punitive damages for harm caused by false or manipulated news reports, also known as the "fake news law." [127]. This law was met with strong opposition from journalists, which ultimately prevented its passage. In response, journalists have committed to either improving the existing co-regulatory bodies or establishing a new, more effective co-regulation by getting platforms involved in an enforcement process. As of 2024, the debate surrounding the co-regulatory reform continues.

4.4.3 Co-regulation of Web Comics

In contrast to the news, comics were an area where free speech was not widely respected even after the authoritarian regime was overthrown in part because they were considered "sub-culture." (P9). As web comics gained a great popularity with the rise of mobile technology, civic groups expressed their concerns [236] and, in 2012, the KCSC issued a notice to webtoon platforms to restrict the access of minors to violent webtoons. This notice led to a large-scale protest among comic artists, resulting in the "Agreement for Self-regulation" between the KCSC and the Korean Comics Artists Association (KCAA) in 2012.

In 2017, the KCAA, in partnership with nine major webtoon platforms, established the Advisory Council on Webtoon Content (ACW). The KCAA holds the sole authority to run the ACW and the platforms pay for its operations. The ACW handles complaints about webtoons that have been forwarded by the KCSC. The ACW evaluates each complaint based on its guidelines and gives advisory opinions to the platforms, such as adjusting the rating or making the content inaccessible to minors. The ACW has faced occasional objections from the platforms, but the majority have ultimately accepted the ACW's opinions, according to the head of the ACW (P11). Webtoon artists and platforms view the ACW as a crucial entity to safeguard the industry from censorship and speak highly of its performance. The director of the KCSC (P14) expressed concerns about the ACW's growing leniency toward artists but confirmed their willingness to maintain the current collaboration with the ACW.

4.5 Qualitative Analysis of Co-Regulatory Frameworks

This section presents a comprehensive analysis of collaborative governance in South Korea's news and web comics sectors. Drawing on Ansell & Gash's framework, our investigation encompasses key elements of collaborative governance, including starting conditions, institutional design, facilitative leadership, and the collaborative process itself. Through in-depth interviews with stakeholders and careful examination of industry practices, we uncovered the factors that contribute to the success of co-regulation in the web comics industry and the challenges faced in the news sector. This analysis not only provides valuable insights into the practical application of collaborative governance theory but also offers lessons for policymakers and industry leaders grappling with content moderation challenges in the digital age. By comparing these two

industries, our aim was to illuminate the nuances of implementing effective co-regulatory frameworks and identify areas where existing theoretical models may be refined or extended.

4.5.1 Starting Conditions

Power/Resource Imbalances Ansell & Gash argue that power imbalances among stakeholders can hinder effective collaborative governance, and if significant disparities in capacity, resources, or representation exist, the process may favor stronger actors [76]. To address this issue, they suggest that collaborative governance requires a commitment to empowering and representing weaker or disadvantaged stakeholders. Our interviews revealed the overwhelming presence and influence of online platforms in the production and consumption of online content in South Korea. The lower cost of entry has led to an meteoric increase in creators, with over 10,000 media companies and 2,000 webtoon creators competing for users' attention. Participants highlighted the centralized power of online platforms, with more than 90% of users consuming news articles and webtoons through major platforms like Naver and Kakao. These platforms make crucial decisions by amplifying or de-platforming content, significantly impacting creators' success.

Power imbalances exist not only between online platforms and creators but also among different creator groups. In the news industry, there are disparities between legal media and small online-only news media, while in the web comics industry, well-established comic artists and emerging artists face different challenges. The distribution of power among online platforms also varies, particularly in the web comics industry, where Naver and Kakao dominate the market, leaving smaller specialized websites with a minor share.

The composition of co-regulatory bodies in these industries reflects the power dynamics at play. As of 2024, the Press Ethics Commission (PEC) has 214 member news organizations, and the Internet News Committee (INC) has 833 online news organization members. However, neither includes online platforms as members, as these creator-driven organizations only make rules applied to creator groups. In contrast, the Advisory Council on Webtoon Content (ACW) has eight member online comics platforms, including Naver and Kakao.

Despite broader participation of news organizations in co-regulation, the impact is limited due to several factors:

- a) Fragmented nature of the industry: The large number of news organizations involved in the PEC and INC may hinder effective decision-making and enforcement. Consensus becomes more difficult to achieve among numerous stakeholders with potentially conflicting interests.
- b) Absence of online platforms: The exclusion of online platforms from these bodies creates a significant gap in the regulatory framework. As primary distributors of news content, their absence limits the effectiveness of any decisions or guidelines established by the PEC or INC.
- c) Diffused responsibility: This fragmentation has led to the limited impact of co-regulatory efforts. P1, a news reporter with more than ten years of experience, states: "I have never felt the presence of either PEC in reality, although I was told they were there since the 1960s. That is why I am skeptical about news co-regulation in general."

In contrast, the ACW's inclusion of major online platforms like Naver and Kakao together with the support of creator groups has fostered a more cohesive and impactful co-regulatory environment:

- a) Inclusive approach: This ensures that all key stakeholders have a voice in the regulatory process, leading to more effective and widely accepted guidelines.
- b) Active platform participation: Naver and Kakao act as distributors for webtoon artists, replacing traditional print comic books. They fund, support, and comply with co-regulatory bodies' decisions while being mindful not to overstep the territory of artists.
- c) Positive reception: This accommodating approach has been well-received by artists. P10, a web comics artist, appreciates the platform editors' input in helping creators manage potential controversies.

Despite its relative success, the web comics industry still faces some challenges. There are tensions between mainstream and niche content: P11, a web comics co-regulator, emphasized that while major platforms' titles often do not raise concerns due to their broader readership, web comics on smaller platforms targeting mature audiences frequently draw complaints from the public. P10, an artist of adult comics, expressed frustration with overly ethical standards imposed on adult-only content and called for more room for creativity.

In conclusion, the power imbalances between online platforms and creators, as well as among different creator groups, pose significant challenges for effective co-regulation in the online content industry. The centralized power of major platforms like Naver and Kakao, coupled with the fragmented nature of news

organizations, limits the impact of co-regulation in the news industry. In contrast, the web comics industry has seen more success in co-regulation due to the active involvement and considerate approach of major platforms, despite some tensions between mainstream and niche content creators. These findings highlight the importance of addressing power imbalances and ensuring inclusive stakeholder participation for effective collaborative governance in digital content industries.

Incentive to Participate Ansell & Gash argue that stakeholders' incentives to participate in collaborative governance depend on the perceived impact of their participation, the availability of alternative venues, and the degree of interdependence among stakeholders [76]. They suggest that collaborative governance is most effective when stakeholders are highly interdependent and when the collaborative forum is the exclusive venue for decision-making. This requires sponsors to ensure that alternative forums (courts, legislators, and executives) respect and honor the outcomes of the collaborative process, establishing the "shadow of the state" [76].

This framework reveals the differences between the news and web comics industries. The Press Ethics Commission (PEC) and the Internet News Committee (INC) primarily evaluate readers' complaints regarding ethical violations in news articles. Their guidelines cover a broad range of journalistic principles, including accuracy, fairness, non-discrimination, copyright protection, conflict of interest avoidance, and clear separation of news and advertising. Despite having the authority to impose penalties up to 10 million KRW (approximately 8,000 USD), the PEC has consistently refrained from exercising this power. The PEC and INC primarily function to evaluate readers' complaints regarding ethical violations in news articles. Their ethical guidelines encompass a wide range of journalistic principles, including ensuring accuracy and fairness in reporting, upholding non-discrimination standards, protecting copyright, avoiding conflicts of interest, and maintaining a clear distinction between news content and advertising. Despite possessing the authority to impose penalties of up to 10 million KRW (approximately 8,000 USD), the PEC has consistently refrained from exercising this power.

Table 4.2 illustrates the PEC's tendency towards lenient punishment, suggesting ineffective disciplinary actions. Although PEC and INC decisions are published online, cautions and warnings carry little weight for news organizations or individual reporters. P2 reported that news editors often disregarded letters from

Year	Cancellation	Dismissal	Caution	Warning	Public Warning	Total
2022	-	26	794	-	68	5
2021	1	14	802	-	48	4
2020	4	-	482	8	-	494
2019	1	-	531	7	-	539
2018	1	28	172	-	82	2
Total	7	5	2,989	23	0	3,024

Table 4.2: PEC Regulatory Actions on News Articles from 2018 to 2022

these regulatory bodies, sometimes not even forwarding them to the journalists who wrote the articles. This ineffectiveness largely stems from the availability of alternative, more potent forums for complainants, such as courts and alternative dispute resolution mechanisms. Section 4.5.2 explains more about forum exclusivity.

The interdependence between creators and platforms varies significantly between the news and web comics industries. In the web comics sector, platforms and artists form a symbiotic relationship, with platforms acting as a hybrid of distribution channel and production studio. While artists retain the copyright, platforms have exclusive rights for domestic and international promotion. P15, a government official involved in both news and web comics industries, explains, "The distribution structure differs between web comics and news. Web comic artists or production companies rarely operate their own distribution channels, relying heavily on portals or specialized platforms. In contrast, news organizations distribute articles through their websites and apps, with portals serving as secondary channels. Consequently, platforms are integral to content regulation in web comics but not in news."

When the ACW issues a decision, platforms discuss it with the artists, make agreed-upon adjustments, and report back to the ACW. P10, a web comics artist, expressed appreciation for this collaborative approach:

> *P10 (web comics artist): "I value platform editors' input as they protect both platforms and creators. Creators can sometimes be detached from societal norms and less attuned to social issues. Their insights help us manage potential risks of inadvertently causing controversies."*

In contrast, online platforms hosting news content have limited influence over editorial decisions. They may terminate contracts with news organizations for breaches or remove illegal content, but they cannot edit news articles to enhance truthfulness, reduce defamation, or mitigate commercialization. Consequently, platforms are excluded from news co-regulation bodies like the PEC and INC. Interdependence among news

organizations is low, while comic artists often collaborate to advocate for shared interests. Interviewees attributed this difference to three factors:

1. Professional privilege: Unlike journalists, comic artists lack historical societal privilege and share a bond as being part of a long-stigmatized subculture.
2. Market competition: News reporters compete for breaking stories, while comic artists operate in diverse sub-genres, reducing direct competition.
3. Content judgment: News co-regulators face challenges in discerning false or deceptive articles, assessing factual accuracy, and identifying misleading sponsored content.

Furthermore, for web comics, the ACW serves as the exclusive forum for rating decisions, bolstered by the "shadow of the state" through its agreement with the Korea Communications Standards Commission (KCSC). Both platforms and artists recognize that the KCSC can intervene if co-regulation fails to meet public standards. This arrangement enhances the incentive for meaningful participation in the collaborative governance process for web comics. Conversely, due to the principle of freedom of the press, as P14, a director at the KCSC confirms that neither the KCSC nor any other executive agencies have regulatory power over news articles, and there is no clear causal connection between the decisions of PEC/INC and court rulings. Therefore, the news industry lacks the “shadow of the state” that incentivizes participation in the web comics sector.

This absence of state oversight in news co-regulation and the lack of exclusive forum status for the PEC and INC have rendered collaborative governance largely symbolic. P8, a government official with extensive experience in media regulation, notes: “Without tangible consequences for non-compliance or clear benefits for participation, many news organizations view co-regulation as an optional, often inconsequential process. This perception significantly hampers the potential for meaningful collaborative governance in the news sector.” This situation raises important questions about how to design effective co-regulatory systems for industries where direct state intervention is constitutionally limited.

Prehistory of Antagonism and Cooperation Ansell & Gash stress that a history of antagonism can create low trust and commitment, leading to manipulation and dishonest communication. Conversely, a history of successful cooperation builds social capital and trust, fostering further collaboration [76]. This element

proves highly relevant in analyzing collaborative governance in the news and web comics industries in South Korea.

While online platforms have emerged as custodians in both industries [193], the reactions of content creators have been starkly different. In the news industry, a climate of mistrust and resentment prevails. P3, a news co-regulator, notes that media companies perceived the rise of online platforms as a loss of direct channels to their audience. P2, a news reporter, laments the growing trend of sensationalist headlines designed to attract clicks. Both P1 and P2 express concern that online news consumption undermines the professional ethics and investigative journalism valued in traditional print media. From the platform perspective, P5 and P6 acknowledge the hostility from reporters while highlighting their own challenges in balancing the conflicting demands of users and news organizations.

> *P2 (News reporter): "If I could change one thing in the news history, I would choose the way [news organizations'] are providing news to online platforms. We should have anticipated that news aggregation services would entirely replace our direct communication channels with the audience."*

On the other hand, comic artists perceived online platforms as partners, not adversaries. P9 & 10 (comic artists) generally appreciated online platforms' efforts to create new business models in the comics industry. Both interviewees considered themselves as being situated in a better position than their predecessors, who were poorly treated by the print comic publishers. They also felt that their creative freedom was respected by platforms. Similarly, P12 & 13 (Web comics platforms) perceived creators as collaborators and tried not to overstep their boundaries except when it came to the illegality of content. Online platforms offer artists annual medical check-ups and free legal counseling services for copyright disputes and other matters. In short, comic artists and online platforms seem to have created a symbiotic relationship.

> *P10 (Web comics artist): "I was told that [traditional print] publishers had forbidden comic artists from discussing or disclosing their compensation with other artists. Online platforms [for web comics] have never attempted to do so and rather, organized gatherings and parties among artists."*

4.5.2 Institutional Design

Participatory inclusiveness The composition of self-regulatory boards in the news and web comics industries reflects a focus on including relevant experts and stakeholders, but with some notable gaps in creator representation. The ACW consists of comic critics/professors, child mental health experts, and online safety advocacy groups, but as of 2024, it does not include artists themselves. The PEC and INC are primarily composed of journalists and journalism professors, most of whom are referred by news industry associations.

Interviewees emphasized the importance of incorporating creators' voices in these governance structures, given the centrality of creativity and journalistic professionalism to these industries. As a comic artist (P9) noted, "Only creators themselves have firsthand knowledge of the delicate and nuanced processes of balancing creativity and social standards, which mental health experts or critiques lacks. Thus, excluding creators in self-regulation does not make sense to me." An online platform executive (P5) echoed this sentiment, stating, "If the board does not include journalists, I am sure co-regulatory efforts will be seriously discredited by journalists."

However, some practical challenges were cited as barriers to creator participation. A co-regulator (P11) justified the absence of comic artists on the ACW board, explaining, "We did not include creators because creators do not feel comfortable with commenting on other artists' work." A comic artist (P10) acknowledged the importance of artist representation but highlighted time constraints as a major obstacle: "We work 7 days a week, day and night. I am not sure if I would be able to work for co-regulation. My understanding is that the current board members try to represent creators' perspectives against paternalistic censorship. If they understand the industry and the importance of their decisions, I am fine with it."

This suggests that while direct creator participation may be ideal, it is not always feasible due to the demanding nature of their work and potential discomfort in critiquing peers. The key may be ensuring that the board members, even if not creators themselves, have a deep understanding of the industry and can effectively represent creators' interests and perspectives. Building trust and open communication channels between co-regulatory bodies and the creator community is essential to ensure that their voices are heard and their concerns are addressed, even if they are not directly seated at the decision-making table.

Forum exclusiveness Ansell & Gash finds that when the collaborative forum is "the only game in town," it is easier to motivate stakeholders to participate; conversely, when they are excluded, they may be impelled to seek out alternative venues [76]. For instances of personal attacks in news articles, individuals can turn to the Press Arbitration Commission (PAC), a quasi-judicial body specializing in alternative dispute resolution for news-related defamation cases. The PAC's procedures, which require the presence of news organizations' editors-in-chief, and its decisions, which courts generally respect, offer a more effective resolution pathway for such cases. Moreover, the judicial system itself presents another viable option for aggrieved parties. P1, a veteran journalist, offers insight into this trend:

> *P1 (journalist): "In recent years, news readers have become increasingly willing to pursue legal action against media companies. We're seeing instances where readers collectively engage high-profile lawyers to file substantial lawsuits against news organizations. I've personally been subject to damage claims exceeding 1 billion KRW (approximately 800,000 USD)."*

This shift in reader behavior has significant implications for co-regulation. As P1 notes, "When individuals can exert pressure through such high-stakes legal proceedings, they are less likely to resort to regulatory bodies like the PEC, which are perceived as ineffectual and only capable of administering nominal punishments." While the PEC and INC possess broad theoretical jurisdiction over various aspects of journalism, their practical impact has been limited. These bodies have been reluctant to adopt stringent measures, which has undermined their effectiveness as dispute resolution mechanisms. Moreover, their lack of exclusive authority in addressing journalistic issues further diminishes their role in the media landscape. As a result, these co-regulatory bodies often struggle to assert themselves as pivotal actors in maintaining journalistic standards and resolving content-related disputes.

In contrast, the ACW holds exclusive authority over the rating of web comics, a power derived from its agreements with both the Korea Communications Standards Commission (KCSC), a regulatory agency, and member platforms. Prior to the ACW's establishment, web comics platforms employed a simplistic voluntary rating system, often using a binary classification (all-ages or adult-only). The ACW's first major initiative was to introduce a more nuanced four-tier rating system: all-ages, 12+, 15+, and 18+ (adult-only). These ratings are based on various factors including violence, sexual content, and extreme language or themes.

When complaints arise, the ACW has the final say in determining or adjusting a comic's rating. Even for issues not directly related to ratings, such as illegal gambling advertisements appearing alongside comics, the ACW can still recommend changes in practices to platforms. P11, a board member of the ACW, noted that while platforms occasionally express opposing views, they ultimately adhere to the ACW's decisions. This level of compliance underscores the ACW's effectiveness and the industry's recognition of its authority in co-regulation.

Moreover, the inherent characteristics of web comics present unique challenges for traditional legal forums in addressing content disputes. Unlike news articles, which are information-centric and more susceptible to defamation claims, web comics rarely contain content that meets the legal threshold for defamation. Political satire in comics, even when pointed, typically falls under protected free speech, further insulating creators from legal action. Furthermore, individuals face significant hurdles in demonstrating concrete harm resulting from violent or sexually explicit content in comics, making such cases difficult to pursue in court. As a result, legal challenges specifically targeting web comic content are infrequent. P12, an online platform executive, corroborated this trend, stating that the majority of lawsuits they have encountered in the industry have centered on copyright infringement rather than content-related issues.

The unique characteristics of web comics, combined with the legal challenges in proving harm or defamation, have contributed to the ACW's role as the primary arbiter of content disputes in this medium. This situation contrasts sharply with the news industry, where legal action is a more viable and frequently used option for addressing content-related grievances.

Clear Ground Rules Both news and web comics co-regulation rely on established rules, but how detailed these rules are varies significantly between the two domains. The PEC and the INC operate under broad, overarching guidelines that encompass a wide range of journalistic principles. These rules often take the form of declaratory statements about fundamental concepts such as freedom of the press, journalistic responsibility, and non-discrimination. For instance, the INC ethics rule states:

- Article 1: Freedom of the Press emphasizes the protection of press independence from all forms of interference to safeguard the public's right to information.
- Article 2: Responsibility of the Press underscores the pursuit of factuality, accuracy, and balance in reporting, while distinguishing between facts and opinions.

- Article 3: Protection of Personal Rights mandates respect for individual privacy and personal information, except when overridden by justifiable public interest.
- Article 4: Protection of the Vulnerable and Non-Discrimination calls for the exclusion of prejudice and discrimination based on various personal characteristics, advocating for the rights of marginalized groups.

In contrast to the broad ethical guidelines governing news media, web comics rating standards are notably more specific and detailed, addressing the unique challenges posed by visual storytelling. The ACW set out to develop highly practical, actionable standards that online platforms could seamlessly integrate into their daily content moderation practices. This approach reflects a keen understanding of the industry's needs and the complexities of regulating visual narratives.

The effectiveness of these tailored standards is evident in the feedback from industry professionals. P13, an executive at a major online platform, highlighted the tangible benefits of this targeted approach: "Since the implementation of these detailed standards, we have seen a marked improvement in our ability to predict content ratings accurately. Moreover, these clear guidelines have significantly facilitated our communication with artists, leading to a more collaborative and transparent creative process." The ACW uses a nuanced checklist to determine violence ratings for web comics, which includes:

- Non-realistic depiction of physical violence, injury, bloodshed, or bodily harm using body parts or tools of non-realistic characters.
- No or very mild depiction of physical violence using body parts or tools of realistic characters.
- Infrequent and non-emphasized depiction of injury, bloodshed, or bodily harm.
- Implicit expression of sexual violence within the overall context.
- Non-continuous or non-specific depiction of physical violence, abuse, or killing using body parts or tools.
- Specific and continuous depiction of physical violence, abuse, or killing in non-realistic art styles that may justify violence and pose a risk of imitation.
- Direct depiction of injury, bloodshed, or bodily harm in non-realistic art styles.
- Explicit, continuous, and specific depiction of physical violence, abuse, or killing accompanied by injury, bloodshed, or bodily harm.

- Specific and direct depiction of sexual violence.

This contrast in rule specificity between the news and web comics reflects the different nature of the content being moderated and the unique challenges each medium presents. While news regulation focuses on broad ethical principles, web comics regulation requires a more granular approach to address the visual and narrative complexities of the medium. By providing clear, applicable criteria, the ACW has not only enhanced the consistency of content ratings but also fostered a more constructive dialogue between platforms and creators.

Process transparency Ansell & Gash stress that the collaborative process should be transparent so that stakeholders are confident that the public negotiation is real and not undermined by backroom deals [76]. In the context of news and web comics, this principle is manifested in two key dimensions: the transparency of decision outcomes and the transparency of the decision-making process itself. There is a consensus among stakeholders regarding the importance of transparent outcomes. The PEC, INC, and ACW all publicly disclose their decisions and primary reasoning on their respective websites. However, the transparency of the decision-making process, particularly regarding the disclosure of meeting minutes, remains a contentious issue.

Those representing co-regulatory bodies and platforms expressed concerns about full procedural transparency. P6, an online platform executive in the news industry, highlighted the practical challenges: "While procedural transparency sounds ideal, it is not always pragmatically feasible. Co-regulators often make decisions that affect powerful entities. The prospect of having every statement scrutinized publicly could deter potential members from participating, especially given that these positions are often neither highly compensated nor prestigious." P4, a co-regulator with experience in both news and webtoon industries, further elaborated on the potential risks: "We maintain internal records of meetings, but full disclosure could expose individual members to undue pressure or lobbying efforts. This could compromise the integrity and candor of our deliberations, which are essential for effective co-regulation."

Conversely, content creators like P10, a web comics artist, advocate for greater transparency: "The essence of co-regulation lies in fostering diverse perspectives and open debate. It mirrors democratic values, where solutions emerge through a process of negotiation and compromise. Transparency in this process is key to its legitimacy and effectiveness." These divergent views highlight the complex balance between

transparency and effectiveness in co-regulatory systems. While transparency is crucial for building trust and legitimacy, there are valid concerns about how full disclosure might affect the quality and integrity of decision-making processes.

4.5.3 Facilitative Leadership

Ansell & Gash argue that in situations where the power distribution is more asymmetrical or incentives to participate are weak, collaborative governance is more likely to succeed if there is a strong "organic leader" who commands the respect and trust of the various stakeholders at the beginning of the process [76]. Organic leaders are those who emerge from within the community of stakeholders. The availability of such leaders is likely to be highly contingent upon local circumstances.

Building on this concept of collaborative governance and organic leadership, our study examined how these dynamics played out in different industrial contexts. As shown in our research, the two industries formed different levels of creator-creator trust and exhibited varying degrees of facilitative leadership. These differences significantly impacted the potential for collaborative governance in each sector.

In the news industry, In the news industry, the challenges to collaboration and organic leadership were particularly evident. P2, a news reporter with 30 years of experience, witnessed few successful collaborations between media companies due to competitive mindsets. P3, a self-regulator and former news reporter, noted that journalists were skeptical of co-regulatory decisions made by their peers, assuming prioritization of company interests over journalistic integrity. P2 observed that journalists' strong egos hindered collective action to improve journalistic quality. This suggests that the industry lacked the type of respected, trust-commanding leaders that Ansell Gash deem necessary for overcoming asymmetric power distributions or weak participation incentives.

The leadership vacuum in the news industry was further exacerbated by the stance of the major online platforms. Despite the fact that most Koreans consume news on online platforms for free (with platforms paying news organizations through "hosting fees"), neither Naver nor Kakao takes the initiative to regulate harmful news content unless legally required. P5 & P6 emphasized that the news enjoys the special protection of the free press, and platforms view themselves as mediators or hosts of content rather than publishers. They expressed a desire to maintain an arm's length relationship with news content. This lack of leadership

figures, combined with the lack of trust among journalists, led P1 (news reporter) and P7 & 8 (government officials) to express skepticism about the success of news co-regulation.

In contrast, the web comics industry exhibited a notably cohesive environment characterized by strong facilitative leadership. Web comic creators, as represented by P9 and P10, expressed confidence in their peers' ability to represent the community's interests without suspicion of self-serving motives. This trust was exemplified by their appreciation for the Korean Comics Artists Association (KCAA) and its leadership efforts.

The web comics industry benefited from the emergence of "organic leaders" in artists' and academic communities. Two professors, frequently cited in interviews, played a significant role in establishing a co-regulatory regime that prioritized artistic freedom. Their approach was marked by close communication with artists, a non-stigmatizing attitude towards comics, and the development of comprehensive rating systems grounded in empirical research applicable across all web comic genres. This vision of co-regulation garnered support from influential comic artists who had achieved both financial success and peer respect, further legitimizing the initiative.

Unlike in the news industry, the major online platforms in the web comics sector actively supported collaborative governance. Naver and Kakao, which together host 90% of web comics, expressed early support for co-regulation and committed to long-term funding. Although smaller, comic-specific platforms have come and gone, the enduring endorsement from these two dominant industry stakeholders lent significant legitimacy to the co-regulatory system.

These contrasting scenarios in the news and web comics industries demonstrate how local circumstances, as suggested by Ansell & Gash, can significantly influence the formation of creator-creator trust and the emergence of facilitative leadership. The collaborative mindset among stakeholders in the web comics industry, coupled with clear facilitative leadership, contributed to the rapid perception of legitimacy for the co-regulatory framework. In contrast, the competitive environment of the news industry, the lack of trust, and the absence of facilitative leadership from both news organizations and platforms hindered the development of effective collaborative governance.

4.5.4 Collaborative Process

The effectiveness of co-regulation depends on the quality of the collaborative process among stakeholders. Our study revealed several key elements that contribute to this process in both the news and web comics industries.

Face-to-Face Dialogue Ansell & Gash emphasize face-to-face dialogue between stakeholders as a necessary condition for meanningful collaboration. PEC, INC, and ACW all hold face-to-face board meetings, organize hearings and conferences to gather broader stakeholders surrounding emerging issues such as COVID-19 misinformation in news and gender discriminatory expression in web comics. The importance of face-to-face dialogue resonated with interviewees as participants emphasized deliberation and peer review. P2, a news reporter, highlighted that the "meritocracy of co-regulators often sacrificed the quality of decisions," suggesting the need for more direct, personal interactions and thorough discussions. In the web comics industry, P9 and P10 appreciated the close communication between creators and co-regulatory bodies, indicating the value of direct dialogue.

Trust Building When there has been a prehistory of antagonism among stakeholders, Ansell & Gash find that trust building often becomes the most prominent aspect of the early collaborative process and can be quite difficult to cultivate. The authors say that because trust building is so important policymakers or stakeholders should budget time for effective remedial trust building. If they cannot justify the necessary time and cost, then they should not embark on a collaborative strategy [76].

Trust levels varied significantly between industries and stakeholders. As noted, in the news industry, both journalists and online platforms reported low trust between journalists and platforms, and even among journalists themselves. The problem is that as co-regulation has not significantly contributed to news ethics for the past several decades, there is widespread skepticism about the co-regulatory effort itself. To these people, trust-building activities, such as open and deliberative dialogues, seemed too weak or naive. P2 noted, "Journalists with strong self-esteem will never follow your lead unless you have very good carrots or sticks." This skepticism was echoed by P8, a government official with over two decades of experience, who offered a pragmatic perspective on the challenges of news co-regulation:

P8 (government official): "The efficacy of news co-regulation is inherently limited without the

willing participation of media companies. In today's landscape, where journalistic endeavors are often driven more by economic necessity than by a sense of public duty, we cannot rely solely on improvements in the institutional design of existing co-regulation to effect meaningful change. Moreover, with the rise of the Internet, online platforms are increasingly becoming central players, both by choice and by necessity, while news organizations and their business associations, despite being the most vocal stakeholders in current co-regulatory efforts, wield diminishing influence. So, I don't have high hope for any co-regulatory or self-regulatory initiatives within the news industry."
P7 (news co-regulator): "Strengthening PEC or INC is a desirable direction, and recently, there seems to be some self-reflection among journalists about the need to reduce misinformation. However, the outlook is bleak. It's questionable whether having good intentions or willingness alone will make it happen. There is a lack of financial resources and enforcement power."

There is, however, a more optimistic view. P3, a news co-regulator, acknowledged the shortcomings of the current news landscape, which have been exacerbated by the Internet revolution's having caused the downfall of legacy media. Nevertheless, P3 believes that Nevertheless, P3 believes that the situation should be rectified by the news organizations themselves, rather than by external forces such as online platforms or the government. "The news organizations' revenues today," states P3, "are directly tied to the number of clicks, so they are forced to be short-sighted and produce sensational headlines to instantly capture users' attention." P3 further offers a promising future, suggesting that if co-regulation proves beneficial to member news organizations and continues to strive to earn trust, it can make positive impacts. "Unlike bigger media companies, smaller companies do not have the capacity to implement systematic complaint-handling systems. Co-regulation may help fill this gap."

On the other hand, the web comics industry demonstrated higher levels of trust among its stakeholders. P9 and P10, both web comics artists, expressed confidence in their fellow creators and online platforms to represent their interests without suspecting self-serving motives. However, even after the agreement between KCSC and ACW was signed in 2016, it took three years for ACW to officially launch. According to P11, this time was used to clarify the expectations of ACW's role among stakeholders, determine how ACW could contribute to artists and platforms, establish how ACW would cover its costs, and ensure that ACW was not perceived as being in opposition to the artists. This was achieved by arranging opportunities for

communication with the stakeholders. P11 states, "I do not want to label our organization as a regulator, nor do I desire more coercive power. We are closer to an advisory board, an organization created to assist artists, platforms, and readers." In essence, ACW began in a much more favorable environment compared to PEC or INC, with a prehistory of collaborative attitudes between stakeholders. The time and budget spent on building trust are now well appreciated by both platforms and artists.

Commitment to the Process According to Ansell & Gash, even when collaborative governance is mandated by law, policy, or organizational guidelines, simply mandating collaboration does not guarantee that stakeholders will actively engage in the process or fully commit to its outcomes [76]. Participants may go through the motions of collaboration without truly investing in the process or believing in its value. Therefore, it needs to achieve genuine "buy-in" from stakeholders, which means that participants understand the importance and potential benefits of collaboration; are willing to engage in good faith discussions and negotiations; are open to considering others' perspectives and making compromises; and feel a sense of ownership and shared responsibility for the process and its outcomes. Participants should abide by the results of deliberation, even if they don't fully align with a stakeholder's preferences.

In the news industry, achieving genuine buy-in from stakeholders in collaborative governance efforts has proven to be a significant challenge. One major obstacle is the highly competitive nature of the news industry. News organizations are constantly vying for audiences, advertisers, and exclusives, which can create a culture of secrecy and mistrust. Journalists may be reluctant to share information or collaborate with rivals, fearing that it could compromise their competitive edge or lead to the loss of a scoop. This competitive mindset can make it difficult for news organizations to see the value in collaboration and to invest time and resources in collaborative governance efforts.

Another challenge is the strong sense of independence and autonomy that is deeply ingrained in the journalistic culture. Journalists often see themselves as watchdogs and truth-tellers, with a responsibility to hold those in power accountable and to report the news without fear or favor. This can make them resistant to external influence or control, including from co-regulatory bodies that may be perceived as a threat to their editorial independence. Journalists may worry that participating in collaborative governance could compromise their ability to report freely and objectively on the issues at hand. P2, a veteran reporter, noted:

P2 (news reporter): "It is hard to expect newspapers to collaborate with each other to serve

common interests because they have a strong ego but lack strategic thinking. They will declare any co-regulatory decision as being unfair and biased whenever that decision is against their interests."

To increase commitment to the collaborative governance process, there have been numerous calls to strengthen effective incentives. Stakeholders have proposed various "carrots" such as credibility scores, safety labels, and preferential treatment in government-funded projects. "Sticks" included disclosure of violations, monetary penalties, and potential de-platforming. Notably, P2 (news reporter) and P3 (news co-regulator), despite their skepticism about platform participation in rule-making, cautiously supported platform cooperation in enforcement, acknowledging de-platforming as an effective deterrent.

P2 (news reporter): "You need to be extremely realistic when it comes to structuring self-regulation. It is like steering wild horses. Journalists with strong self-esteem will never follow your lead unless you have very good carrots or sticks."

However, given the principle of the freedom of the press, arming co-regulation with coercive incentives may not be a sensible approach. P7 states: "The Bar Association can prevent lawyers from practicing if they are not registered with the association. However, due to the freedom of the press, such coercive measures are impossible for media companies." P8 (government official) suggests that fostering consumer awareness and empowering the public to "vote with their wallets" by canceling subscriptions to unethical or unreliable outlets may be a more effective and appropriate mechanism for holding media organizations accountable.

Shared Understanding Shared understanding is a crucial element in collaborative governance, where stakeholders develop a common vision of their collective goals. This concept is described across different studies in various terms such as "common mission," "shared vision," or "clear goals." [76]. It involves agreeing on the definition of the problem, necessary knowledge, and core values. Shared understanding is part of a broader collaborative learning process and is essential for effective cooperation among diverse stakeholders [76].

In the web comics industry, there appears to be a stronger shared understanding among stakeholders. Both artists and platforms recognize the mutual benefits of co-regulation and share the common goal of maintaining creative freedom while addressing societal concerns. ACW serves as a focal point for developing and implementing shared standards, fostering a collective approach to content regulation.

In contrast, the news industry struggles to achieve a similar level of shared understanding. The fragmented nature of news organizations, their competitive environment, and the historical tensions with online platforms have hindered the development of a common vision for co-regulation. The PEC and INC have not successfully fostered a shared understanding of the goals or definition of problems among stakeholders. This challenge is compounded by the complex nature of evaluating harm in news content, as highlighted by P8:

> *P8 (government official): "Evaluating the potential harm of news reports, such as fake news or misinformation, is more difficult than that of web comics. People can easily agree upon whether certain comic cuts are too explicit and not appropriate for juveniles. However, figuring out the falsity of information takes time and often ignites endless political debates."*

The difficulty in assessing news content stems from the need to verify facts and make judgments on facts, which requires laborious investigation and often leads to political debates [98]. This characteristic imposes a greater burden on news co-regulation compared to web comics, where adjudicators can make decisions on visual material without extensive investigation. This contrast underscores the importance of developing mechanisms to foster shared understanding, particularly in industries dealing with complex and politically sensitive content like the news media.

Intermediate Outcomes Intermediate outcomes, or "small wins," are crucial in collaborative governance, especially when overcoming prior antagonism and building long-term trust [76]. These tangible achievements, though potentially modest, can create momentum and encourage a positive cycle of trust-building and commitment. Joint fact-finding is often cited as a beneficial type of intermediate outcome. However, this approach may not be suitable for all situations, particularly when stakeholders have more ambitious goals that are difficult to break down into smaller milestones.

In the news industry, despite overall challenges, there is evidence of some progress through intermediate outcomes. P3, a news co-regulator, pointed out that PEC and INC have achieved small wins, particularly in reducing vivid depictions of suicides and improving the reporting culture of election polls. This suggests that a theme-by-theme approach to co-regulation might be more effective in the news sector. These focused achievements demonstrate the potential for collaborative efforts to yield tangible results even in a complex and often antagonistic environment. By concentrating on specific, well-defined issues, the news industry's

co-regulatory bodies have managed to create some positive momentum, which could serve as a foundation for broader cooperation in the future. The web comics industry has also experienced intermediate outcomes. These include the development of industry-wide standards, and the successful implementation of content ratings. While P14, a director at the KCSC, expressed concerns about ACW's being too lenient towards artistic freedom, the achievement of these intermediate outcomes has been widely promoted among industry stakeholders and has reinforced the value of the co-regulatory process.

In both cases, recognizing and celebrating these small wins, no matter how modest, can be crucial in maintaining stakeholder engagement and demonstrating the value of collaborative governance. This is particularly important in the news industry, where historical antagonisms and complex content issues make large-scale agreements more challenging to achieve.

4.5.5 Summary of Results

This analysis of co-regulatory efforts in South Korea's news and web comics industries provides a nuanced understanding of the challenges and opportunities in the collaborative governance of online content. These industries, both integral to South Korea's digital media landscape, offer contrasting case studies in the implementation and effectiveness of co-regulation. The web comics industry, a relatively new and dynamic sector, has emerged as a successful model of co-regulation. This success can be attributed to several factors aligned with Ansell & Gash's collaborative governance framework:

- A history of cooperation rather than antagonism, with online platforms viewed as partners by comic artists, in contrast to their predecessors' experiences with traditional publishers.
- High interdependence between creators and platforms, with major players like Naver and Kakao actively supporting and participating in the co-regulatory process.
- Strong facilitative leadership from "organic" leaders, including respected professors and successful artists, who helped establish a co-regulatory regime prioritizing artistic freedom.
- The exclusive authority of ACW, which serves as the primary forum for content-related decisions.
- Clear and detailed ground rules, particularly in content rating systems, which have improved predictability and facilitated communication between platforms and creators.
- Shared understanding among stakeholders about the goals of co-regulation and the balance between

creative freedom and social responsibility.

On the other hand, the news industry, with its long-standing traditions and complex stakeholder relationships, faces the following significant challenges in implementing effective co-regulation:

- Low trust among stakeholders and towards the co-regulatory bodies, PEC and INC, stemming from competitive industry dynamics and historical tensions.
- Weak incentives for participation due to alternative forums (such as courts and the Press Arbitration Commission) and the co-regulatory bodies' lack of enforcement power.
- The fragmented nature of the industry, with over 10,000 media companies, and a strong sense of journalistic independence that resists external oversight.
- Difficulties in achieving consensus on complex issues like misinformation, which require laborious fact-checking and often spark political debates.

The emphasis that the Ansell & Gash framework places on understanding the prehistory of cooperation or antagonism proved crucial in explaining the divergent trajectories of co-regulation in the news and web comics industries. This aspect highlights the importance of qualitative analysis, as assessing starting conditions and industry dynamics often requires deeper, insider perspectives that quantitative data alone cannot capture. Our research underscores the necessity of understanding evolving business models and stakeholder landscapes, factors that significantly influence the potential for successful co-regulation. The framework's focus on stakeholder interdependence, incentives to participate, and the "shadow of the state" offered valuable insights into why ACW faced more favorable conditions for effective co-regulation compared to its counterparts in the news industry. These elements help explain the varying levels of commitment and compliance observed in each sector, highlighting the framework's utility in predicting collaborative governance outcomes.

4.6 Advancing Collaborative Governance Framework Through Case Study

Our study of South Korea's news and web comics industries demonstrates Ansell & Gash's framework's strengths while also revealing areas for potential refinement. The experiences of South Korea's news and web comics industries confirms the relevance and significance of Ansell & Gash' [76]s 10 guiding principles

for constructing collaborative governance. These considerations prove valuable in anticipating the outcomes of co-regulation and informing upfront design decisions. Policymakers and leading stakeholders at international, national, and regional levels who are pursuing AI co-regulation would benefit from referring to this list.

1. If there are significant power/resource imbalances between stakeholders, such that important stakeholders cannot participate in a meaningful way, then effective collaborative governance requires a commitment to a positive strategy of empowerment and representation of weaker or disadvantaged stakeholders.
2. If alternative venues exist where stakeholders can pursue their goals unilaterally, then collaborative governance will only work if stakeholders perceive themselves to be highly interdependent.
3. If interdependence is conditional upon the collaborative forum's being an exclusive venue, then sponsors must be willing to do the advance work of ensuring that alternative forums (courts, legislators, and executives) respect and honor the outcomes of collaborative processes.
4. If there is a prehistory of antagonism among stakeholders, then collaborative governance is unlikely to succeed unless (a) there is a high degree of interdependence among the stakeholders or (b) positive steps are taken to remediate the low levels of trust and social capital among the stakeholders.
5. Where conflict is high and trust is low, but power distribution is relatively equal and stakeholders have an incentive to participate, then collaborative governance can successfully proceed by relying on the services of an honest broker that the respective stakeholders accept and trust. This honest broker might be a professional mediator.
6. Where power distribution is more asymmetrical or incentives to participate are weak or asymmetric, then collaborative governance is more likely to succeed if there is a strong "organic" leader who commands the respect and trust of the various stakeholders at the outset of the process. "Organic" leaders are leaders who emerge from within the community of stakeholders. The availability of such leaders is likely to be highly contingent upon local circumstances.
7. If the prehistory is highly antagonistic, then policymakers or stakeholders should budget time for effective remedial trust building. If they cannot justify the necessary time and cost, then they should not embark on a collaborative strategy.

8. Even when collaborative governance is mandated, achieving "buy-in" is still an essential aspect of the collaborative process.
9. Collaborative governance strategies are particularly suited for situations that require ongoing cooperation.
10. If prior antagonism is high and a long-term commitment to trust building is necessary, then intermediate outcomes that produce small wins are particularly crucial. If, under these circumstances, stakeholders or policymakers cannot anticipate these small wins, then they probably should not embark on a collaborative path.

However, our analysis also revealed several areas where the framework could be enhanced. First, reducing overlaps in concepts like forum exclusivity, interdependence, and ground rules, which appear in multiple categories within the framework, could improve its analytical clarity through a more streamlined structure. The framework would benefit from explicitly addressing the impact of constitutional and legal protections, such as freedom of the press and anti-trust law, on the feasibility of certain co-regulatory mechanisms. Additionally, more attention could be given to the difficulties of achieving shared understanding in industries dealing with complex, politically sensitive content, as exemplified by the challenges faced in regulating news content. Lastly, the framework could be strengthened by incorporating more practical considerations, such as budgetary constraints and specific incentive mechanisms, potentially repositioning these elements as part of institutional design rather than starting conditions, given their critical role in shaping collaborative governance outcomes.

Based on qualitative analysis of the news and comics industries in South Korea, this study suggests adding four more principles to Ansell & Gash's framework. By proposing this updated framework, we emphasize the importance of considering both practical and normative challenges when designing co-regulatory approaches, ensuring that these frameworks are not only effective but also aligned with fundamental values such as free speech and sustainable and effective in the long term.

11. If there are normative challenges that condition collaborative governance, such as free speech or anti-competition laws, establishing certainty about the available options for collaborative governance is crucial. Courts and policymakers need to provide clear guidance on how these legal frameworks interact with and potentially limit collaborative governance efforts.

12. Funding stability and independence must be ensured. Even if some participants exit from the governance structure, the basic operational funding should not be threatened. This ensures the longevity and effectiveness of the collaborative effort.
13. Effective enforcement and monitoring mechanisms need to be established. Clear signals must be provided to non-compliant and compliant actors, clarifying the incentive structure. This helps maintain the integrity of the collaborative governance system and encourages consistent participation.
14. If there is resource disparity and compliance with standards requires significant investment, policymakers and leading stakeholders must provide technical and other support for participants with limited resources. This ensures a level playing field and enables broader, more inclusive participation in the collaborative governance process.

This study complement Ansell & Gash's framework with the normative and practical challenges that most co-regulatory governance systems are likely to face. By extending this discussion, this study aims to offer an analytical structure for designing and implementing effective co-regulation in AI governance.

4.7 Envisioning Co-regulation in AI Governance

Faced with the emerging threats of AI systems, let us assume that stakeholders discuss establishing a co-regulatory body with AI developers and major tech companies, academic researchers and ethicists, government representatives, civil society organizations, industry-specific representatives (e.g. from healthcare, finance, law, education), and end-users and consumer advocates. This body will make industry-wide rules, have the capacity to monitor the compliance of its members (major AI developers and tech companies) and to provide technical support for under-resourced members. What challenges will they face and what lessons can they learn from the updated Ansell & Gash framework?

4.7.1 Predictable Challenges

Normative Challenge (1): Freedom of Expression While collaboration sounds ideal, implementing it in the real world is challenging. Flexibility might mean less stability and unclear accountability. Even though a representative body may develop a comprehensive set of rules, the regulatory targets might not have a strong incentive to bear the costs of changing their practices. The challenge becomes even more pronounced

when considerations of free speech come into play. Free speech principles inherently prioritize individuals' self-governance over compliance with external rules. As illustrated in Figure 4.2, free speech principles may encourage decisions to be made at the lowest level (individual users) instead of the highest levels (outside regulators), which inevitably causes tension with co-regulation.

In the US context, courts apply stricter scrutiny when the government enacts laws regulating specific types of content. For example, the U.S. Supreme Court struck down provisions of the Communications Decency Act of 1996 that aimed to protect minors from harmful material on the internet. The Court found that the Act violated the First Amendment, as it was overly broad and could potentially restrict adults' access to constitutionally protected speech [29]. Likewise, the Court sided with Playboy Entertainment, striking down a provision of the Telecommunications Act of 1996 that required cable television operators to either fully scramble or block channels primarily dedicated to sexually explicit programming [32]. Additionally, the Court struck down a California law restricting the sale or rental of violent video games to minors and found that the Federal Communications Commission's indecency policy was unconstitutionally vague, lacking clear guidelines for broadcasters to follow [36].

Even when professional associations like the National Association of Broadcasters (NAB) create rules, courts may apply First Amendment scrutiny if such initiatives are encouraged or driven by the government. In 1975, a group of television writers and independent producers challenged a policy called the "family viewing policy"[412]. They believed that this policy, which restricted certain types of content during specific hours, violated their freedom of speech. They argued that the government had pressured the networks and NAB into adopting this policy, which they opposed. The court agreed that the government's involvement in creating this policy was problematic. Although the court did not entirely prohibit networks from adopting family-friendly programming policies, it prevented NAB from enforcing the policy and the networks from following NAB's rules. The court sought to ensure that individual broadcasters could make their own decisions about what to air, without being forced to follow the decisions of a larger group. Mark M. MacCarthy commented that this decision signaled to the industry that any future code should contain only advisory guidelines, and should not be interpreted or enforced by a centralized industry body[272].

Another example involved the National News Council (NNC). NNC, established in 1973, was composed of fifteen members strategically balanced to represent both the public interest and media professionals.

This independent body was tasked with considering complaints against a broad spectrum of national media outlets, including prominent newspapers, news agencies, magazines, and television networks [118]. Despite its noble intentions to promote accountability and fairness in news reporting, the NNC faced significant resistance from both the press and the public it aimed to serve. A notable opponent was the publisher of The New York Times, who viewed the council as a potential threat to press freedom guaranteed by the First Amendment [181]. The Times argued that the council could be a precursor to press regulation and would encourage a public view of the press as a monolithic entity whose conduct was legitimately susceptible to some single standard other than the approval of readers or viewers. Due to insufficient funding and support, the NNC was dissolved in 1984, and most state press councils had shut down as well by that time [118].

How do these established doctrines influence co-regulation in generative AI systems? The traditional free speech doctrine ("marketplace of ideas") assumes that the best way to counter harmful speech is with more speech. However, the power and pervasiveness of generative AI systems may distort this marketplace, potentially requiring new approaches to regulate the centralized power of distributing and amplifying certain viewpoints and to ensure a truly open exchange of ideas. Unlike search engines that provide pointers for further exploration, the current generative AI systems like ChatGPT offer definitive answers, which can come across as overly authoritative and suggest finality [357]. By synthesizing results from multiple sources, AI chatbots mask the range of available information, hindering users' ability to explore and build information literacy. This might necessitate external regulations such as preventing discriminatory output or requiring that diverse viewpoints be presented.

However, it is not clear whether the imposition of rules on individual AI developers or companies for content-related matters violates such companies' free speech rights. Following the precedent set in cases involving search engines like Baidu [38], generative AI providers could argue that their systems' outputs constitute protected speech under the First Amendment. As seen in recent cases like *NetChoice v. Paxton* (pending in the US Supreme Court as of July 2024), courts have been inclined to protect the editorial judgments of digital platforms. Generative AI providers might claim similar protections for their algorithms and model designs, arguing that these represent editorial choices protected by the First Amendment. Furthermore, as seen in the "family viewing policy" case and recent *Murthy v. Missouri* case, where lower courts found the Biden Administration in violation of the First Amendment in its efforts to combat Covid-19 misin-

formation by "coercing" digital platforms' content moderation, government encouragement or involvement in co-regulatory efforts could trigger First Amendment scrutiny, potentially limiting the effectiveness of such initiatives.

Normative Challenge (2): Antitrust Law One of the key aspects of co-regulation is the development of consistent standards and guidelines across different platforms. However, this standardization may conflict with antitrust law. Agreements between competitors regarding business conduct can give rise to antitrust concerns, involving not only government-imposed regulations but also purely private agreements established by professional organizations. A historical example of this occurred in 1979 when the U.S. Department of Justice filed an antitrust suit against NAB [26]. The suit claimed that the NAB Code's advertising provisions, which included restrictions on advertising time sold per hour, prohibitions on specific ad types, and guidelines on separating advertising and editorial content, limited the advertising inventory available in the market, thereby driving up advertising prices. Although the NAB Code was voluntary, its widespread adoption by broadcasters had a significant impact on the market price.

In response to the lawsuit, the NAB and the DOJ reached a settlement in the form of a consent decree, which is an agreement that resolves a dispute without an admission of guilt or liability. Under the terms of the consent decree, NAB agreed to abandon the advertising provisions of the Code, and in January 1983, NAB abandoned the entire Code [26, 118]. The case illustrates the challenges and potential pitfalls of multistakeholder content regulation. As documented by Angela J. Campbell, who meticulously recorded the rise and fall of media self-regulation, Congress passed a law that exempted broadcasters from antitrust laws, allowing them to take collective action to reduce violent programming without fearing legal repercussions, but did not engage in any significant collective actions to curb violent content.

This narrative is echoed in the context of AI systems. In recent years, concerns have grown regarding the potential antitrust implications of cross-platform collaborations. For instance, Mark Lemley emphasizes the importance of preserving laws that allow tech companies the freedom to determine the content on their platforms, rather than imposing detailed content scrutiny or treating platforms as government actors [259]. Lemley stresses that in the long run, imposing too many rules on tech companies can make it harder for new competitors to enter the market and therefore limits the choices available to consumers. Evelyn Douek has warned about the rise of "content cartels" resulting from cross-platform collaboration pressure [154]. Douek

argues that an opaque cartel may be as detrimental as a monopoly over public discourse, potentially leading to unregulated cartels that exacerbate existing content moderation issues, such as lack of transparency, accountability, concentration of power, and inconsistent application of standards across platforms.

Building on antitrust concerns in content moderation, the landscape of generative AI systems presents even more complex challenges. As AI companies strive to develop safe and aligned systems, there is a growing need for collaboration and information sharing. However, this cooperation could potentially raise antitrust red flags. For instance, leading AI companies might want to share insights about their alignment techniques, safety measures, or ethical guidelines to ensure responsible AI development across the industry. This could include sharing data on model behaviors, successful containment strategies for potentially harmful outputs, or best practices in prompt engineering to mitigate risks. While such collaboration could significantly enhance AI safety and ethical standards, it might also be seen as anti-competitive behavior. The sharing of alignment strategies or safety protocols could be interpreted as a form of collusion that limits innovation or creates barriers to entry for new competitors. Smaller AI companies or startups might argue that such collaboration among industry giants unfairly disadvantages them, as they may not have access to the same level of shared knowledge or resources. This could potentially stifle competition and innovation in the AI sector.

To navigate these challenges, any co-regulatory framework for generative AI must carefully balance the need for industry-wide safety standards and ethical alignment with the principles of fair competition. This might involve creating safe harbors for certain types of information sharing related to AI safety, similar to the exemptions granted to broadcasters for reducing violent content. Alternatively, it could require the involvement of neutral third parties or government agencies to facilitate the exchange of critical safety information without risking antitrust violations.

Practical Challenge (1): "Why bother?"

Co-regulation could have a basis in statutes, but in most cases, particularly when applied globally, it tends to rely on agreement between stakeholders, or in other words, good will. When enforcement does not follow, good intentions can easily be obstructed. The situation worsens if (1) co-regulatory decision harms the individual company's short-term profits, or (2) the decision is contestable. Collaboration aim at achieving

common goals, such as online trust and safety, can face challenges in defining the roles and responsibilities of participants due to the intangible and non-exclusive benefits involved. This can lead to free-riding behaviors where some participants benefit without contributing their fair share, making it difficult to sustain collaborative efforts [305]. This free-rider problem is unavoidable when the collective goals are distant from individual companies' short-term goals and are not accompanied by effective enforcement mechanisms [305].

One example of this is the Internet Content Rating Association (ICRA), which despite widespread industry and political backing, struggled to convince website owners to voluntarily label their websites to help parents determine whether the site was appropriate for them or their children. The EU and Germany provided financial support for the development of self-labeling safety tools [78]. In 1999, Microsoft openly pledged its support for the ICRA: "Microsoft's involvement in ICRA is an expression of our commitment to working with members of the Internet industry to help users understand online safety issues and have a positive experience online." [30] However, in his compelling essay titled "ICRAfail–A Lesson for the Future," the former CEO of ICRA recollects that even members of ICRA refuse to use ICRA labels on their websites, and Microsoft did not make a necessary update on Internet Explorer. When even core contributors do not want to make small changes in their practices, the ICRA was unable to convince hundreds of thousands of website owners to accept its answers to the question, "why bother?" The ICRA failed to achieve its goal and shut down in 2010. Similar participant attrition also occurred with respect to privacy self-regulation online as documented by multiple scholars [141, 213, 191].

In the context of generative AI, comparable challenges to those faced by ICRA and privacy self-regulation initiatives could emerge in co-regulatory efforts. For instance, leading AI companies might initially agree to collaborate on developing ethical guidelines for generative AI systems, such as preventing the generation of harmful content or ensuring transparency in AI-generated output. However, as the market for AI services grows more competitive, individual companies might be tempted to prioritize their short-term profits over adherence to these collective standards. If there are safety measures that might slightly reduce the versatility or speed of an AI system, all participants might agree on the importance of such measures in principle, however, individual companies might hesitate to fully implement them if they believe it could put them at a competitive disadvantage. This reluctance could be particularly pronounced if the benefits of compliance

are intangible or long-term.

The challenge of defining and measuring compliance in the AI field could also contribute to the free-rider problem. Unlike website labeling or privacy policies, assessing adherence to AI ethics guidelines or safety protocols can be highly complex and technically challenging. This complexity could provide cover for companies that wish to appear compliant while not fully implementing agreed-upon measures. Furthermore, the rapid pace of AI advancement might render some co-regulatory agreements obsolete almost as soon as they are implemented. This could lead to a situation where companies nominally adhere to outdated standards while pushing the boundaries with newer technologies that fall outside the scope of existing agreements.

Practical Challenge (2): Who participates?

Another challenge of co-regulation is ensuring that stakeholders have a meaningful opportunity to participate in decision-making processes. Ostrom stresses that all stakeholders, regardless of their level of power or access to resources, should have a meaningful opportunity to engage in decision-making processes and shape the rules that govern their behavior [306]. Without such opportunities, as Ansell & Gash address in power imbalances, there is a risk that the co-regulation process will be dominated by powerful stakeholders or that the voices of certain groups will be marginalized or ignored, leading to a lack of legitimacy and effectiveness in the governance system. However, in a diverse and decentralized online environment, such shared norms and goals are often absent, and a dozen board members are unlikely to fully represent the unique perspectives of different user groups [119]. Geographic considerations also play a significant role in determining who makes the rules. Decisions must be made on whether the focus should be on a global, national, or local level. For example, proponents of Social Media Councils have yet to reach an agreement on whether a global or national model is more appropriate. A global model may offer consistent standards across platforms, while national models may better cater to the unique cultural, social, and legal contexts of each country and more easily serve as the "shadow of the state." [371, 76]

In the context of generative AI, the challenge of ensuring meaningful participation in co-regulatory processes is particularly complex and multifaceted. The global reach and impact of AI systems, combined with the diversity of stakeholders and the rapid pace of technological advancement, create a unique set of challenges for inclusive and representative decision-making. The technical complexity of AI systems also

poses a barrier to inclusive participation. Many of the ethical and safety considerations in AI development require a deep understanding of the technology, which may limit the ability of non-technical stakeholders to meaningfully contribute to decisions. This could lead to a situation where technical experts dominate the conversation, potentially overlooking important societal and ethical considerations.

The landscape of generative AI is further complicated by the coexistence of closed and open-source models, as well as the wide array of applications built on these models. Closed-source models, developed by large tech companies, may have different priorities and constraints compared to open-source models created by academic or community-driven initiatives. Applications ranging from creative tools to scientific research assistants add another layer of complexity. This diversity makes it challenging to define a cohesive "membership" for co-regulatory bodies and to ensure that all relevant perspectives are represented. Another concern is the cultural homogenization fostered by generative AI. Critics have pointed out that many large language models and generative AI systems are primarily trained on Western, English-language data, leading to outputs that reflect Western cultural norms and perspectives [368]. This bias has raised concerns about the global applicability and fairness of these systems. Ensuring diverse cultural representation in co-regulatory bodies is crucial in addressing this issue, but it also raises questions about how to balance global standards with local cultural contexts.

Moreover, the legal and regulatory landscape for AI varies significantly across different jurisdictions. Free speech protections, anti-discrimination laws, privacy regulations, intellectual property rights, and trade secret protections all differ from country to country. For instance, the robust free speech protections in the United States may clash with the stricter content regulation of European countries or the censorship practices in authoritarian regimes. Privacy laws like GDPR in Europe impose different requirements from those in other parts of the world. These legal disparities make it extremely challenging to create universally applicable rules for AI systems on a global scale.

Practical Challenge: Who Pays the Fees?

An operational fee is often overlooked in the early stages of forming a rule-making organization but is actually a crucial building block that defines the scope of its work. For example, paying honorariums to part-time council members and meticulous monitoring across a broad web, along with considerate dispute

resolution, require significant costs that are crucial for gaining trust in the system. For multiple stakeholders to self-regulate successfully, they must possess not only the willingness to do so but also the necessary expertise [118]. However, financial constraints can impede an organization's ability to carry out its mission effectively, as exemplified by the National News Council (NNC), which relied solely on donations and faced difficulties in reviewing the fairness and accuracy of stories due to limited funding. The NNC was unable to attract competent staff and relied heavily on graduate students for its investigations. Similarly, the ICRA was not able to assess its achievements because it lacked the infrastructure to crawl the web and count those websites that had been labeled out of all available websites [78].

In addition to funding adequacy, it is important to consider who pays. Industry funding may create a capture problem, and government funding can invite unwanted censorship attempts or political pressure. Regulators who wish to maintain independence from the industry and the government may rely on unstable funding sources such as donations. In cases where receiving a decision from a regulator, such as a rating for video games, is a condition for financial gains or entry to the market, a regulator may charge a fee for their services. This model can achieve both independence and funding sustainability but is limited in its applicability.

In the context of generative AI, operational costs for an AI co-regulatory body would be substantial. The organization would need to attract and retain highly skilled professionals with expertise in AI technology, ethics, law, and policy. These experts would need to keep pace with the rapid advancements in the field, requiring ongoing training and education. The body would also need sophisticated technical infrastructure to monitor and assess AI systems, potentially including high-performance computing resources to test and evaluate large language models and other AI systems. Moreover, the global nature of AI deployment would necessitate a geographically diverse team and potentially multiple offices worldwide to address region-specific issues and cultural contexts. This global presence would further increase operational costs.

The question of who should fund such an organization is particularly challenging in the AI context. Industry funding from major AI companies could provide substantial resources but risks creating a perception (or reality) of regulatory capture. Given the immense influence these systems can have on society, public trust in the independence of the regulatory body is crucial. Government funding could provide more stability and potentially greater resources, however, it could also invite political interference, which is particularly

Figure 4.3: Guiding Principles for AI Co-Regulatory Governance

.

concerning given the potential of AI systems to influence public opinion and democratic processes. Relying on donations, as some previous attempts at self-regulation have done, is likely to be insufficient given the scale and complexity of AI governance. A fee-for-service model, where AI companies pay for assessments or certifications, could provide a sustainable funding source. Therefore, a diversified funding model could help maintain independence while ensuring adequate resources. However, it would require careful design to balance the interests of all stakeholders and maintain public trust.

4.7.2 Strategies for AI Co-Regulation Model

Drawing from our case studies and anticipated challenges, we propose the following guiding principles for AI co-regulation as illustrated in Figure 4.3, while acknowledging the complex landscape and potential obstacles:

Starting Conditions

1. **Acknowledge Power Disparities**: Recognize and address existing power imbalances among stakeholders in the AI industry, including disparities between large and small AI companies, as well as

between companies and users. Consider differences between open-source and closed models, and between model developers and application creators.

2. **Leverage Interdependence**: Capitalize on the interdependence among AI developers, particularly in sharing advanced alignment/safety technologies, to encourage participation in the co-regulatory process.
3. **Historical Context**: Be mindful of the prehistory of cooperation and conflicts within the AI industry. For example, consider relationships between content creators and new media outlets.
4. **Normative Challenges**: Address potential claims of free speech rights by AI companies, considering recent legal precedents like *NetChoice v. Paxton*, as well as antitrust concerns that may arise from industry-wide collaboration, potentially by creating safe harbors for sharing critical safety information.

Institutional Design

1. **Inclusive Participation**: Recognize difficulty in ensuring meaningful participation from all stakeholders given AI's general-purpose nature. Address challenges in diversity across closed-source, open-source, and various AI applications, as well as different legal/cultural landscapes. Ensure participation from leading AI companies while also including smaller companies and impacted individuals or groups. Divide governance into specific themes such as labor, creativity, and professional services to allow for context-specific regulations.
2. **Exclusive Forum**: Consider establishing an official, exclusive forum for AI governance discussions, while being cautious of potential antitrust implications.
3. **Enforcement Mechanisms**: Address the free-rider problem when benefits are intangible or long-term. Acknowledge difficulty in defining and measuring compliance in AI. Establish clear signals for compliant and non-compliant actors. Clarify incentive structures to encourage consistent participation.
4. **Diversified Funding**: Industry funding risks regulatory capture; Government funding may invite political interference; Donation-based funding likely insufficient for AI governance scale. Aim for a variety of funding sources to ensure basic operational funding is not threatened by participant exits. Maintain longevity and effectiveness.
5. **Transparent Processes**: Increase transparency by publishing decisions, relevant data, and allowing

general and regular participation from various groups in the governance process.

6. **Cultural Representation**: Ensure diverse cultural representation to address concerns about AI systems reflecting primarily Western cultural norms.

Collaborative Process

1. **Face-to-Face Dialogue**: Facilitate in-person, thorough deliberative processes among stakeholders to encourage meaningful engagement.
2. **Trust Building**: Allocate sufficient time and resources for building trust, especially when addressing topics with a history of conflict.
3. **Long-Term Commitment**: Foster a long-term commitment to the process with predictable outcomes for both compliance and non-compliance. Develop mechanisms to ensure all participants contribute fairly and benefit proportionately from the collaborative efforts.
4. **Incremental Progress**: Focus on achieving "small wins," such as establishing rules for specific AI applications or contexts, or developing shared technology features.
5. **Technical Support**: Ensure the capacity to provide technical assistance to under-resourced members, promoting equitable participation.

Facilitative Leadership

1. **Identify Organic Leaders**: Recognize and nurture "organic leaders," such as academics or leading AI developers, within the AI community who can guide the co-regulatory process.
2. **Articulate Collective Benefits**: Encourage leaders to clearly communicate the shared advantages of participating in the co-regulatory framework, addressing the "Why bother?" challenge.
3. **Champion Ethical Practices**: Support leaders in promoting and exemplifying ethical AI development practices.

Ultimately, the success of this co-regulatory vision hinges on fostering a culture of responsible innovation. Leading companies and scholars must take the initiative to shape ethical AI development practices that can be adopted industry-wide. This proactive approach not only helps mitigate risks, but also builds

public trust in generative AI technologies. While implementing these strategies presents significant challenges, meaningful progress may begin with acknowledging these difficulties and approaching them with flexibility and determination. It is important to recognize that the journey towards effective AI co-regulation will likely involve trial and error, requiring patience and persistence from all involved parties. Despite these challenges, the potential benefits of a well-implemented co-regulatory framework——including enhanced innovation, improved safety, and increased public trust——make this endeavor not only worthwhile but necessary.

Chapter 5

Paths Forward for AI Co-Governance

5.1 Recap of Case Studies

The two case studies presented in this dissertation offer valuable insight into the practical implementation of co-design and co-regulation methods in contexts relevant to AI governance. Case Study 1 examined a co-design model for AI governance in the legal domain, engaging 20 legal experts through case-based deliberation workshops. This study yielded a comprehensive four-dimensional framework for evaluating AI responses in legal contexts: (1) user attributes and behaviors, (2) nature of queries, (3) AI capabilities, and (4) social impacts. This framework provides a structured approach to incorporating expert insights into AI governance policies, ensuring consideration of a full range of relevant factors.

The study highlighted the value of domain-specific expertise in identifying critical considerations often overlooked by AI developers, such as unauthorized practice of law, confidentiality concerns, and potential liability for inaccurate advice. It demonstrated the effectiveness of case-based deliberation in eliciting nuanced insights grounded in real-world scenarios, moving beyond abstract principles to actionable guidance. However, the research also revealed significant challenges in translating expert insights into coherent, implementable policies. The diversity of expert perspectives, while valuable, often led to conflicting recommendations. This underscored a key tension in co-design methods: the desire for open discourse and diverse views can conflict with the need for clear, actionable policies.

Importantly, this case study led to a critical reflection on the nature of participation in AI governance. It

challenged the notion that maximizing participation is always beneficial or achievable. Instead, it suggested a more nuanced strategy that recognizes the complexities of AI governance and the need for strategic calibration of stakeholder involvement. This critique of participation maximization highlights the importance of balancing diverse needs, motives, and contextual factors in co-design processes.

Case Study 2 investigated co-regulation in online content moderation in South Korea, comparing outcomes in the web comics and news industries. This comparative analysis, based on interviews with key stakeholders and the application of Ansell & Gash's collaborative governance framework, revealed how industry-specific dynamics significantly influence co-regulation effectiveness. The study found that successful co-regulation hinges on factors such as stakeholder interdependence, facilitative leadership, clear ground rules, and shared understanding. The web comics industry demonstrated a more successful model of co-regulation, benefiting from a history of cooperation, strong interdependence between creators and platforms, and clear ground rules. In contrast, the news industry faced significant challenges due to low trust among stakeholders, weak incentives for participation, and difficulties in achieving consensus on complex issues.

These findings led to crucial discussions about envisioning AI co-regulation. The study highlighted potential challenges in applying co-regulation to AI governance, including the general-purpose nature of AI systems, the complexity of stakeholder groups, and unclear interdependence between stakeholders. It emphasized the need for a segmented construction of AI co-regulation, recognizing that different AI applications and sectors may require tailored governance strategies. Moreover, the study extended discussions of collaborative governance by examining normative challenges particularly relevant to internet and AI regulation, such as free speech concerns, as well as practical challenges like funding sources. These insights offer a more nuanced understanding of the factors influencing the efficacy and longevity of co-regulatory frameworks in the digital era.

Together, these case studies provide empirical grounding for our theoretical framework and offer valuable lessons for future governance efforts in the rapidly evolving field of AI. They highlight the potential of co-design and co-regulation models while also revealing the significant challenges in their implementation. These insights pave the way for more effective, context-sensitive, and sustainable governance strategies in AI development and deployment.

5.2 Lessons Learned

Through detailed examination of AI in legal advice and content moderation in South Korea's web comics and news industries, several key lessons emerge. These lessons highlight the critical importance of context, the challenges inherent in collaborative governance structures, and the unique barriers posed by the nature of AI systems.

5.2.1 Contexts Matter, Significantly

The synthesis of findings from the two case studies reveals a complex landscape for co-governance in emerging technologies. Both studies highlight the critical importance of industry-specific dynamics, stakeholder relationships, and historical contexts in shaping the success of co-governing efforts.

The findings of Case Study 1 demonstrate the critical importance of context in the development and implementation of AI systems for legal advice. The appropriateness and effectiveness of AI-generated legal guidance are deeply rooted in a complex web of factors spanning individual, legal, technological, and societal domains: (1) user-specific factors such as identity, geographic location, legal sophistication, and access to resources; (2) the complexities of user queries that involve various laws, jurisdictions, and case-specific details; (3) the capabilities and limitations of AI systems themselves with respect to accuracy, context-awareness, confidentiality, accountability, and potential biases; and (4) the broader social and ethical contexts including potential consequences for third parties and society at large.

The findings reveal the profound value of drawing on centuries of professional expertise in developing AI governance for legal advice. Legal experts offer crucial perspectives on the distinction between legal information and legal advice, rooted in long-established professional ethics designed to protect clients from unauthorized guidance in high-stakes decisions. Their insights highlight critical aspects of confidentiality, privacy, and accountability in AI-generated advice.

Conversations with AI systems carry inherent security risks, as they can be unintentionally exposed to third parties, including potential leaks unknown to system operators. This vulnerability extends to legal proceedings, where AI interactions lack the protection of attorney-client privilege. Unlike conversations with lawyers, which are privileged, interactions with AI systems—whether stored locally or online—can be subject to court orders and potentially used as incriminating evidence. This leaves users exposed in

subsequent legal processes, compromising their privacy and legal position.

In addition, AI systems fundamentally lack the professional accountability that characterizes the legal profession. Although attorneys are bound by strict ethical standards and face severe consequences such as disbarment or penalties for grossly negligent counsel, AI systems bear no such responsibility for providing incorrect or harmful advice. Lawyers must adhere to a well-established standard of care, ensuring that their practice meets industry standards. In contrast, AI-generated advice currently operates without these crucial accountability mechanisms, leaving users without recourse for poor or damaging guidance.

These findings underscore a broader principle: AI developers can and should learn from the accumulated wisdom of professional communities. Each domain—be it law, finance, mental health, or medicine—has evolved unique ethical frameworks based on specific safety concerns and practical experiences. This suggests that AI governance in professional advice domains should not adopt a one-size-fits-all approach, but rather develop tailored policies that incorporate input from diverse stakeholders within each field. Furthermore, the study illuminates how AI norms cannot be divorced from human-developed norms that predate both the Internet and AI technology. Societies have formed nuanced understandings of acceptable and unacceptable practices, which vary significantly between cultures, countries, and professional domains. These established norms inevitably shape and constrain discussions around AI governance.

There may be a temptation to create universal rules that apply to all AI systems. However, this research demonstrates the inherent limitations of such an overarching solution. The complexity and context-specificity of problem domains necessitate more nuanced domain-specific governance frameworks. This perspective acknowledges the rich tapestry of human knowledge and ethical frameworks that have evolved over time, ensuring that AI systems are developed and deployed in a manner that respects and builds upon this accumulated wisdom.

Likewise, Case Study 2, which examines co-regulation efforts in South Korea's news and web comics industries, strongly emphasizes the importance of contextual matters for successful collaborative co-regulation. The web comics industry in South Korea demonstrated a more successful model of co-regulation, benefiting from a history of cooperation, strong interdependence between creators and platforms, and clear ground rules. In contrast, the news industry faced significant challenges due to low trust among stakeholders, weak incentives for participation, and difficulties in achieving consensus on complex issues. The study's appli-

cation of Ansell & Gash's collaborative governance framework reveals how industry-specific dynamics, historical relationships, and stakeholder interdependence significantly influence the potential for effective co-regulation.

While the overall institutional design is similar (government support, authorities based on agreements, creator-platform collaborations), the outcomes of co-regulatory systems differed by industry. The web comics industry benefited from a history of cooperation and high interdependence between creators and platforms, shared understanding of the public stigmatization of the industry, and the government's tangible regulatory threats. Furthermore, the presence of respected "organic" leaders in the web comics industry and the exclusive authority of its co-regulatory body contributed to its relative success.

In contrast, the news industry's fragmented nature and long-standing tensions between news organizations and online platforms hindered effective co-regulation. The strong sense of journalistic independence in the news industry conflicted with external oversight, disarming co-regulatory boards from exercising more effective enforcement mechanisms. Aggrieved parties of defamatory or false news articles often seek alternative forums such as courts and Congress, which undermines co-regulation and disincentivizes long-term participation. This disparity underscores how the specific characteristics and dynamics of each industry shape the outcomes of co-regulatory efforts.

This reveals that collaborative governance is incredibly context-sensitive. The success of such governance is largely determined by factors such as power dynamics and the nature of relationships among stakeholders. It suggests that when conditions are unfavorable across multiple dimensions, it may be prudent to abandon collaborative strategies altogether, as the inherent flexibility can lead to the rapid dissolution of the governance structure.

This study highlights the importance of human calibration and coordination, particularly the often subtle efforts required to build trust over time. These soft skills and relationship-building activities are crucial to creating a foundation of mutual understanding and cooperation. This study also stresses the need for robust organizational requirements to support the collaborative process. Central to their framework is the concept of incentives. Although motivation to participate is partly shaped by the pre-history of stakeholder relationships (e.g., their level of interdependence), Ansell & Gash argue that clear and compelling incentives are critical for sustaining governance over time [76]. These incentives serve as a driving force, encouraging

ongoing engagement and commitment from all parties involved.

5.2.2 Barriers to AI Co-Governance

The case studies indicate the inherent fragility of collaborative governance structures. The very attributes that make co-governance appealing—flexibility, adaptability, voluntary participation, and shared responsibilities—also render it vulnerable to dissolution. Co-design efforts are highly dependent on the goodwill of developers. Although companies may initially engage stakeholders in policy-making to gain legitimacy and public trust, this commitment often falters when faced with shifting corporate priorities. The history of co-design and participatory design is replete with examples of short-lived nominal initiatives that failed to effect a meaningful change in practice [147, 148, 322, 381].

The unique characteristics of generative AI systems further complicate this landscape. The concept of a "participatory ceiling" encapsulates the limitations facing stakeholders when trying to meaningfully influence the AI foundation models [381]. Corporate control presents a significant barrier, as foundation model developers, predominantly large tech companies, have little incentive to share control with broader communities. Participatory efforts often amount to mere consultation rather than conferring real decision-making power on stakeholders. This dynamic results from the centralization of resources and computational power in well-resourced companies, a corporate focus on shareholder interests over open collaboration, risk aversion in large organizations, and practical challenges such as intellectual property concerns.

The context-agnostic nature of foundation models exacerbates the challenges [113]. These models aim for universality and general applicability across domains, which conflicts with the context-specific nature of meaningful participation. Unlike the well-defined domains of the news and web comics industries examined in Case Study 2, generative AI systems serve a vast array of stakeholders–from students to teachers, government agencies to mental health institutions—adapting their persona based on user queries. This versatility makes it challenging to formulate theme-specific or context-specific governance structures.

Consequently, current participatory efforts tend to focus on abstract, universal guidelines rather than incorporating local knowledge and addressing specific contexts. It makes it difficult for participants to reason about concrete applications and risks and poses significant challenges in addressing the specific harms and concerns of marginalized communities. The result is a governance model that struggles to balance the need

for broad applicability with the imperative for meaningful context-sensitive stakeholder participation.

Similarly, co-regulation involves the challenging task of aligning diverse interests from corporations, government actors, and civil societies toward a common goal. This process requires substantial effort, but participants retain the freedom to withdraw when decisions conflict with their interests. Unlike hard binding laws, co-regulation ideally emerges from inclusive dialogues. The non-binding nature of co-regulation, while allowing for inclusive dialogue, also makes it vulnerable to collapse. In the absence of "organic" leaders, even minor operational flaws can force stakeholders to leave the governance structure.

Moreover, the current landscape of AI-generated content, primarily based on user prompts, is likely to implicate free speech concerns when external co-regulatory bodies attempt to impose safety rules. This tension between content regulation and freedom of expression echoes long-standing debates in media governance, and the dynamic and personalized nature of AI output complicates traditional notions of content moderation and raises questions about the boundaries of protected speech in human-AI interactions. There are also competing demands for transparency and data protection. Although effective oversight and the accountability of AI systems require a high degree of transparency, including access to training data, model architectures, and decision-making processes, AI developers have legitimate concerns about protecting proprietary information, intellectual property, and user privacy.

These normative challenges present significant hurdles for co-regulatory efforts in AI governance. They require careful navigation of complex legal and ethical terrain, often involving trade-offs between competing values and interests. The global nature of AI development and deployment adds another layer of complexity to these normative challenges. Different jurisdictions have varying legal frameworks and cultural norms on issues such as privacy, free speech, and the role of technology in society. This diversity of strategies makes it difficult to establish universally applicable governance frameworks, which can lead to regulatory fragmentation and forum shopping by AI developers.

5.3 Guiding Principles for AI Co-Governance

Drawing from the lessons learned through our case studies and broader research, this section outlines key guiding principles for effective AI co-governance. These principles are designed to address the challenges identified in our research while leveraging the strengths of collaborative methods. We focus on three critical

areas: the importance of context specificity in governance models, the need to view AI governance as a human-centric process, and the necessity of resolving legal ambiguities surrounding AI regulation.

5.3.1 Focus on Context-Specificity

As demonstrated in both case studies, careful consideration must be given to the evaluation of the favorability of contextual factors and how institutional design can foster a collaborative environment. Similarly, in a co-design model for AI governance, the strategy can be tailored to specific contexts and stakeholder needs, despite the general-purpose nature of generative AI systems. Case Study 1 highlights the unique considerations of the legal advice domain, which separates it from other professional fields and AI applications. This highlights the critical need for AI developers to incorporate field-specific, time-tested wisdom, and widely accepted norms when crafting governance frameworks. In doing so, they ensure that AI systems respect and align with the established practices and ethical standards of each domain in which they operate. In addition, the diversity of cultural, national, and community norms requires a nuanced approach to the representation and governance of the AI system. Different societies may have varying expectations and standards that need to be reflected in the AI systems that serve them.

Therefore, before embarking on any co-design or co-regulation initiative, policymakers, researchers, and practitioners must carefully define the scope, audience, and jurisdiction of the effort. The bottom-line questions include the following.

- Who are the key stakeholders that need to be involved, and what are their respective roles, interests, and incentives?
- What is the specific problem or challenge that the initiative seeks to address, and what are the desired outcomes?
- What is the appropriate level and scale of engagement, given the nature of the problem and the resources available?
- How will the initiative be governed, and what mechanisms will be put in place to ensure transparency, accountability, and inclusivity?
- What are the potential risks and unintended consequences of the initiative, and how will these be

mitigated?

5.3.2 Governing AI as Human-Centric Process

Case studies demonstrate the effectiveness of human-centric methodologies in AI governance research, highlighting the value of nuanced, context-rich approaches. These methods, including case-based deliberation and confidential interviews, prove to be instrumental in uncovering the complexities of AI governance challenges and potential solutions. When viewed through the lens of Ansell & Gash's collaborative governance framework, these strategies underscore the importance of well-designed, sustained processes for meaningful stakeholder engagement.

In Case Study 1, case-based deliberation enables free-form exploratory thinking among experts, utilizing realistic scenarios to engage them effectively. This method allows for the examination of both granular concerns and overarching constraints, producing concrete, contextual factors that go beyond theoretical principles. The collective deliberation facilitated by this method reveals hidden dimensions and nuances, as experts build upon each other's insights and challenge their own initial analyses. In Case Study 2, confidential one-on-one interviews foster open conversations about sensitive matters and provide insider perspectives on industry dynamics. This method allows for a deeper understanding of contexts not apparent in formal documentation and reveals underlying reasons for co-regulatory successes and failures. The anonymity afforded by these interviews encourages candid responses, leading to more authentic and valuable lessons.

These human-centric approaches offer several advantages to AI governance research, aligning with Ansell & Gash's emphasis on face-to-face dialogue, trust building, and commitment to the process. They provide a contextual understanding that goes beyond abstract principles to concrete, real-world considerations. By directly involving industry experts and insiders, these methods ensure that governance frameworks are informed by those with deep domain knowledge and practical experience, fostering the shared understanding that Ansell & Gash identify as crucial to successful collaboration.

However, it is important to recognize that these human-centric processes require careful design and implementation over an extended period to be truly effective. Collaborative governance is not a quick fix, but a long-term commitment. In the context of AI governance, this means establishing ongoing forums for dialogue, regularly reviewing and refining governance frameworks, and maintaining consistent stakeholder

engagement even as technologies and societal needs evolve.

The collaborative nature of case-based deliberation and the depth of one-on-one interviews allow the exploration of complex issues from multiple angles, revealing subtleties that might be missed in more structured ways. This aligns with Ansell & Gash's emphasis on iterative learning processes in collaborative governance. Over time, these methods can help build the shared ownership and mutual gains that are essential for sustainable collaborative efforts. These methods bridge the gap between theory and practice, offering actionable insights for AI developers and policymakers. They also play a crucial role in building trust among stakeholders, an essential element in developing effective governance structures. This trust-building aspect is particularly important in the context of AI governance, where rapid technological advances and potential social impacts can create tensions between different stakeholder groups.

Moving forward, AI governance efforts should continue to prioritize these human-centric methodologies, complementing quantitative data analysis with qualitative insights as part of a long-term iterative process rather than one-off initiatives. This approach can lead to more effective, adaptable and context-sensitive governance structures that address the unique challenges posed by AI technologies in various domains and industries. By grounding governance strategies in real-world experience and expert knowledge, and committing to ongoing, carefully designed collaborative processes, these powerful technologies will serve society's best interests, fostering a governance model that is both robust and responsive to the evolving challenges of AI.

5.3.3 Resolving Legal Ambiguities

Addressing normative uncertainties is crucial for effective AI governance, as we have seen in media co-regulation that court rulings disfavoring collective rule-making, either due to free speech or antitrust concerns, can immediately disrupt long-standing efforts to coordinate co-regulation, leading to the rapid dissolution of the entire system.

The potential conflict between AI regulation and free speech protections has been a source of concern for some stakeholders. However, a closer examination of existing legal frameworks suggests that free speech law need not be a significant barrier to constructing effective AI co-regulation. Firstly, the speech of AI providers is likely to be considered corporate speech, which historically receives less protection than indi-

vidual speech. Moreover, the unique characteristics of AI systems—their potential for widespread impact, the concentration of power in the industry, and concerns about discrimination and bias—justify tailored regulation to protect public interest and user autonomy [125].

Precedents provide a model for balancing free-speech considerations with the need for regulation. Although commercial speech is protected by the First Amendment, deceptive advertising has been prohibited. The Supreme Court established a four-part *Central Hudson* test to determine when commercial speech can be regulated. The test asks whether (1) speech concerns lawful activity and is not misleading, (2) the government's interest is substantial, (3) the regulation directly advances the government's interest, and (4) the regulation is not more extensive than necessary [25]. This test allows for the Federal Trade Commission's regulation of deceptive advertising, which has been upheld in cases such as *POM Wonderful LLC v. FTC* [40].

Another illustrative example of balancing free speech rights with harm mitigation is the ethics regulation of academic research by the International Review Board (IRB). Academic freedom is heavily protected by the First Amendment [22], but most universities and research institutes are subject to the Federal Policy for the Protection of Human Subjects, which is also known as the "Common Rule." [68] The freedom of researchers to design, conduct, and write about research is restricted by this rule to protect the rights and welfare of human research subjects. The relationship between researchers and human subjects bears similarities to that between AI service providers and users. In both cases, there is a significant power imbalance and information asymmetry that renders the latter vulnerable to potential manipulation or abuse. Just as human subjects may agree to participate in research without fully understanding the risks involved, users of AI systems may consent to service terms without a clear understanding of how their data will be used or how AI output might influence their beliefs and behaviors in ways that may cause psychological, emotional, or material damage. The lack of transparency surrounding many AI systems further compounds these risks, as users are often left in the dark about how these systems operate and make decisions.

Given these parallels, the ethical principles and regulatory frameworks that have been developed to protect human research subjects could serve as a valuable model to govern the relationship between AI providers and users. A tailored regulatory framework could be developed to ensure that these systems are designed and deployed in a manner that respects individual autonomy, mitigates potential harms, and aligns

with the public interest. This could involve requirements for transparency, accountability, and ongoing monitoring, as well as mechanisms for addressing bias and discrimination.

To resolve this legal ambiguity, collaboration between multiple stakeholders is significant. Courts play a crucial role in interpreting how existing laws apply to new AI technologies. By engaging with other stakeholders, the judiciary can develop a nuanced understanding of AI's complexities, potentially leading to more informed rulings that balance innovation with societal interests. Legal scholars and ethicists can help clarify how First Amendment protections apply to AI-generated content, drawing on precedents in areas such as corporate speech and commercial regulation. Policymakers and legislators, informed by these interpretations, can then work to craft new laws or adapt existing ones to address AI-specific challenges. AI developers and companies are essential partners in this process, as their technical expertise is vital to understand the capabilities and limitations of AI systems. Their participation ensures that governance frameworks are both effective and technically feasible.

Civil society organizations and human rights experts can advocate for the public interest and ensure that AI governance aligns with established human rights principles. Industry associations can help develop self-regulation standards and best practices, which can complement formal regulations. These efforts can be particularly valuable for addressing industry-specific challenges and promoting responsible AI development within the private sector. International organizations, such as the UN or OECD, can facilitate global dialogue and coordination on AI governance. Given the borderless nature of many AI applications, international cooperation is crucial to develop consistent and effective governance models.

Legal ambiguities in AI governance present challenges, but they need not be insurmountable barriers. Through collaborative efforts and proactive engagement with legal issues, it is possible to develop a governance framework that navigates these complexities effectively, fostering responsible AI development while respecting fundamental legal principles.

5.4 Limitations of the Dissertation

Although this dissertation offers significant insights into AI co-governance, it is important to acknowledge its limitations to contextualize the findings and identify areas for future research.

Narrow Domain Focus The study's concentration on AI systems providing legal advice, news apps, and comics apps, while offering depth in these areas, presents several limitations. The governance models developed for these specific domains may not easily scale or apply to the broader spectrum of AI applications, potentially overlooking unique challenges in other sectors. The study may not adequately account for emerging or future AI technologies, such as general-purpose AI systems, that could have fundamentally different stakeholder dynamics.

Stakeholder Engagement Constraints The research primarily engaged experts and professionals, such as lawyers in Case Study 1 and full-time creators, government officials, and platform executives in Case Study 2. While valuable, introduces certain biases and gaps. The absence of direct input from everyday users of AI systems may result in governance models that do not fully address consumer concerns or experiences. Over-reliance on expert opinions might lead to governance structures that favor professional or industry interests over broader societal needs.

Implementation Challenges The dissertation appears to have given limited attention to several practical aspects of implementing proposed governance models. There is insufficient exploration of the political and economic obstacles to adopting co-design and co-regulation methods across different jurisdictions and markets. The study may not fully address the challenges of integrating proposed governance models into existing regulatory frameworks or the potential resistance from established regulatory bodies.

Methodological Limitations Case studies are not representative of the diverse range of AI governance scenarios with varying cultural, political, or technological contexts in different regions or societies, potentially limiting the generalizability of the findings. Moreover, given the rapid pace of AI development, some findings can quickly become outdated, requiring ongoing research and updates.

While acknowledging the limitations of this work, I believe this dissertation contributes significantly to the ongoing dialogue on effective AI governance. This study represents a starting point for exploring the potential of co-design and co-regulation in AI governance. The application of these models in AI research is still nascent, offering ample opportunities for further investigation and refinement. Future research could examine the specific mechanisms that foster long-term stakeholder engagement, explore ways to balance

power dynamics in co-governance structures, and investigate how these models can adapt to the rapid pace of AI development.

Chapter 6

Conclusion

The development of AI governance is likely to be segmented as we observe the burgeoning of diverse voluntary guidelines, legislative proposals, and international norms. This diversity is not necessarily a drawback. Each form of rules, whether hard or soft, offers unique values, and as the development trajectory of AI remains unpredictable, ethics and norms discussed in society and policy arenas naturally oscillate and gradually evolve. This dissertation contributes to the ongoing discussions by exploring two promising governance models, co–design and co-regulation. Reflecting on the expansive literature across law, public policy, and computer science, this study demonstrates how the knowledge accumulated in these governance models can offer fertile ground for AI governance.

By combining theoretical and empirical analysis, it examines how these collaborative models can be pragmatically implemented to create more effective, context-sensitive governance frameworks. The research employs qualitative methods to uncover rich, contextual insights, advancing our understanding of stakeholder dynamics and institutional design in AI governance. Building upon and extending established theoretical frameworks such as case-based reasoning and Ansell & Gash's collaborative governance model, this work develops forward-looking guiding principles for AI governance. Although the path to AI development remains complex and uncertain, the insights gained from this study underscore the potential of collaborative approaches to governing AI systems. By fostering inclusive dialogues, leveraging diverse expertise, and remaining adaptable to changing technological and societal landscapes, co-design and co-regulation models offer promising avenues for navigating the challenges of AI governance.

Bibliography

[1] 234. Fla. Stat. § 784.048.

[2] Tex. Penal Code Ann. § 42.072.

[3] Engage!: Co-designing search engine result pages to foster interactions.

[4] 740 Ill. Comp. Stat. Ann. 14/1 et seq.

[5] Cal. Civ. Code § 1708.86, .

[6] Cal. Bus. & Prof. Code § 6450, .

[7] Cal. Penal Code § 528.5(a), .

[8] Cal. Civ. Code §§ 1798.100 - 1798.199.

[9] 20 U.S.C. § 1681, .

[10] 42 U.S.C §§ 2000d - 2000d-7, .

[11] 42 U.S.C. §§ 7401-7671q.

[12] 18 U.S.C. § 2261A.

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

[14] personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor.

[15] N.Y. Penal Law § 190.25.

[16] 29 U.S.C. §§ 651-678.

[17] 47 U.S.C. § 230.

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

[19] Brown v. Board of Education, 347 U.S. 483, 1954.

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

[21] Grievance Comm. of Bar v. Dacey, 222 A.2d 339 (Conn.), appeal dismissed, 386 U.S. 683, 1966.

[22] Keyishian v. Board of Regents, 1967. 385 U.S. 589, 603.

[23] Baron v. City of Los Angeles, 2 Cal. 3d 535, 1970.

[24] Washington v. Davis 426 U.S. 229, 1976.

[25] Central Hudson Gas & Electric Corp. v. Public Service Commission of New York, 1980. 447 U.S. 557, 566.

[26] United States v. National Ass'n of Broadcasters, 536 F. Supp. 149 (D.D.C.), 1982.

[27] O'Brien v. Muskin Corp., 94 N.J. 169, 1983.

[28] Newspapers. *Encyclopedia of Korean Culture*, 1995. URL `http://encykorea.aks.ac.kr/Contents/Item/E0032944`.

[29] Reno v. ACLU, 521 U.S. 844, 1997.

[30] Internet content rating association formed to provide global system for protecting children and free speech on the internet. *Microsoft*, May 1999. URL `https://news.microsoft.com/1999/05/12/internet-content-rating-association-formed-to-provide-global-system-for-protecting-children-and-free-speech-on-the-internet/`.

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

[32] United States v. Playboy Entertainment Group, Inc., 529 U.S. 803, 2000.

[33] Organizing a Legal Discussion (IRAC, CRAC, etc.). *Columbia Law School Writing Center*, 2001. URL https://www.law.columbia.edu/sites/default/files/2021-07/organizing_a_legal_discussion.pdf.

[34] Ashcroft v. American Civil Liberties Union, 542 U.S. 656, 2004.

[35] McDonald v. City of Chicago, 561 U.S. 742, 2010.

[36] Brown v. Entertainment Merchants Association, 564 U.S. 786, 2011.

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

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

[39] Constitution of the United States—A History. https://www.archives.gov/founding-docs/more-perfect-union, 2015.

[40] POM Wonderful LLC v. FTC, 777 F.3d 478, 501-502 (D.C. Cir.), 2015.

[41] Legal Information vs. Legal Advice. *Texas Office of Court*, 2015. URL https://www.txcourts.gov/media/1220087/legalinformationvslegaladviceguidelines.pdf.

[42] 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. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

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

[44] Community engagement matrix. *Northern Beaches Council*, January 2017. URL https://s3.ap-southeast-2.amazonaws.com/hdp.au.prod.app.nthbch-yoursay.files/9215/6214/7251/Community_Engagement_Matrix_.pdf.

[45] Packingham v. North Carolina, 137 S. Ct. 1730, 2017.

[46] Facebook to be fined $5bn over Cambridge Analytica scandal. *BBC News*, Jul 2019. URL https://www.bbc.com/news/world-us-canada-48972327.

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

[48] OECD AI Principles overview. *OECD*, May 2019. URL https://oecd.ai/en/ai-principles.

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

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

[51] Participatory Data Stewardship. *Ada Lovelace Institute*, 2021. URL https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/.

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

[53] Digital Services Act: agreement for a transparent and safe online environment. *European Parliament News*, Apr 2022. URL https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environment.

[54] AI Risk Management Framework: Second Draft. *National Institute of Standards and Technology, U.S. Department of Commerce*, 2022. URL https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

[55] Recommendation 2022-3, automated legal guidance at federal agencies. *Administrative Conference of the United States*, July 2022. URL https://www.acus.gov/recommendation/automated-legal-guidance-federal-agencies. 87 Fed. Reg. 39,798.

[56] Blueprint for an AI Bill of Rights. *The White House Office of Science and Technology Policy*, 2022. URL https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[57] Blueprint for an AI Bill of Rights. *The White House*, 2022. URL https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[58] The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees. *U.S. Equal Employment Opportunity Commission*, May 2022.

[59] Act On Promotion Of Information And Communications Network Utilization And Information Protection. Article 44-7 (Prohibition on Circulation of Unlawful Information). *Korea Legislation Research Institute*, 2022. URL https://elaw.klri.re.kr/kor_service/lawView.do?hseq=60899&lang=ENG.

[60] 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.

[61] FTC Warns About Misuses of Biometric Information and Harm to Consumers. *Federal Trade Commission*, May 2023.

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

[63] Biden-Harris Administration Announces New NIST Public Working Group on AI. *NIST*, 2023. URL https://perma.cc/FCP7-Z7P3.

[64] Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. *The White House*, Jul 2023. URL 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/.

[65] Artificial intelligence and data act. *Government of Canada*, September 2023. URL https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act.

[66] FEC Approves Rulemaking Petition, Discusses Advisory Opinion. *FEC.gov*, 2023.

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

[68] 45 C.F.R. § 46.101 et seq., N/A.

[69] 42MaleStressed. Chatgpt jailbreak – therapy session, treatment plan, custom code to log the session., December 2022. URL https://www.reddit.com/r/ChatGPT/comments/zig5dd/chatgpt_jailbreak_therapy_session_treatment_plan.

[70] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59, 1994.

[71] Kenneth W. Abbott and Duncan Snidal. The governance triangle: Regulatory standards institutions and the shadow of the state. In Walter Mattli and Ngaire Woods, editors, *The Politics of Global Regulation*, pages 44–88. Princeton University Press, Princeton, NJ, 2009. doi: 10.1515/9781400830732.44.

[72] Daron Acemoglu and Simon Johnson. Big Tech Is Bad. Big A.I. Will Be Worse. *The New York Times*, June 2023. URL https://www.nytimes.com/2023/06/09/opinion/ai-big-tech-microsoft-google-duopoly.html.

[73] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[74] Benjamin Alarie and Rory McCreight. The Ethics of Generative AI in Tax Practice. *Tax Notes Federal*, pages 785–793, 2023.

[75] Sam Altman, Greg Brockman, and Ilya Sutskever. Governance of superintelligence. *OpenAI*, May 2023. URL https://openai.com/blog/governance-of-superintelligence.

[76] Chris Ansell and Alison Gash. Collaborative governance in theory and practice. *Journal of public administration research and theory*, 18(4):543–571, 2008.

[77] Maria Antoniak, Aakanksha Naik, Carla S. Alvarado, Lucy Lu Wang, and Irene Y. Chen. Designing guiding principles for nlp for healthcare: A case study of maternal health. 2023.

[78] Phil Archer. IRACfail: Lessons for the Future. `https://philarcher.org/icra/ICRAfail.pdf`, 2009.

[79] Alissa Ardito. Social Media, Administrative Agencies, and the First Amendment. *Administrative Law Review*, 65:301, 2013.

[80] Sherry R Arnstein. A ladder of citizen participation. *Journal of the American Institute of planners*, 35(4):216–224, 1969.

[81] American Bar Association. *Nonlawyer activity in law-related situations: A report with recommendations*. ABA, Chicago, IL, January 1995. ISBN 978-1-57073-239-3.

[82] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[83] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R.

Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, 2022.

[84] 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.

[85] Brhmie Balaram, Tony Greenham, and Jasmine Leonard. Artificial Intelligence: Real Public Engagement. *RSA, London*, 5, 2018.

[86] Michael Balas, Jordan Joseph Wadden, Philip C. Hébert, Eric Mathison, Marika D. Warren, Victoria Seavilleklein, Daniel Wyzynski, Alison Callahan, Sean A. Crawford, Parnian Arjmand, and Edsel B. Ing. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. *Journal of Medical Ethics*, 2023. ISSN 0306-6800. doi: 10.1136/jme-2023-109549. URL https://jme.bmj.com/content/early/2023/11/09/jme-2023-109549.

[87] Jack M Balkin. Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society. In *Law and Society Approaches to Cyberspace*, pages 325–382. Routledge, 2017.

[88] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 116–128, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445875. URL https://doi.org/10.1145/3442188.3445875.

[89] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 116–128, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445875. URL https://doi.org/10.1145/3442188.3445875.

[90] Derek E Bambauer and Mihai Surdeanu. Authorbots. *J. Free Speech L.*, 3:375, 2023.

[91] Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab, and Pascale Fung. Enabling Classifiers to Make Judgements Explicitly Aligned with Human Values, 2022.

[92] Hugh Baxter. *Habermas: the discourse theory of law and democracy*. Stanford University Press, 2011.

[93] Eevi Beck. P for political: Participation is not enough. *Scandinavian journal of information systems*, 14(1):77–92, 2002.

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

[95] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, page 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.

[96] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, page 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.

[97] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahnemann, Jan Brauner, and Sören Mindermann. Managing AI Risks in an Era of Rapid Progress. *arXiv preprint arXiv:2310.17688*, 2023.

[98] Carl T. Bergstrom and Jevin D. West. The nature of bullshit. In *Calling Bullshit: The Art of Skepticism in a Data-Driven World*, pages 38–49. Random House, 2020.

[99] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

[100] Julia Black. Decentring regulation: Understanding the role of regulation and self-regulation in a 'post-regulatory'world. *Current legal problems*, 54(1):103–146, 2001.

[101] Aj Blechner. Legal Research Strategy. *Harvard Law School Library*, December 2022. URL `https://guides.library.harvard.edu/law/researchstrategy`.

[102] Editorial Board. 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/`.

[103] Susanne Bødker and Morten Kyng. Participatory design that matters—facing the big issues. *ACM Trans. Comput.-Hum. Interact.*, 25(1), feb 2018. ISSN 1073-0516. doi: 10.1145/3152421. URL `https://doi.org/10.1145/3152421`.

[104] Susanne Bødker, Christian Dindler, and Ole Sejer Iversen. Tying knots: Participatory infrastructuring at work. *Computer Supported Cooperative Work (CSCW)*, 26:245–273, 2017.

[105] Alexander Bogner. The paradox of participation experiments. *Science, Technology, & Human Values*, 37(5):506–527, 2012.

[106] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak,

Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2022.

[107] James Broughel. Rules for robots: A framework for governance of ai. 2023. URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4620277`.

[108] James Broughel. The Case For Artificial Intelligence Regulation Is Surprisingly Weak. *Forbes*, April 2023. URL `https://www.forbes.com/sites/digital-assets/2023/04/07/the-case-for-artificial-intelligence-regulation-is-surprisingly-weak/?sh=66fe39b950a8`.

[109] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300271. URL `https://doi.org/10.1145/3290605.3300271`.

[110] John M. Bryson, Barbara C. Crosby, and Melissa Middleton Stone. Designing and implementing cross-sector collaborations: Needed and challenging. *Public Administration Review*, 75(5):647–663, 2015.

[111] Harriet A Bulkeley, Vanesa Castán Broto, and Gareth AS Edwards. *An Urban Politics of Climate Change: Experimentation and the Governing of Socio-technical Transitions*. Routledge, 2014.

[112] Kevin Byron. Creative Reflections on Brainstorming. *London Review of Education*, 10:201–213, 2012.

[113] Aylin Caliskan and Kristian Lum. Effective AI regulation requires understanding general-purpose AI. *Brookings*, January 2024. URL https://www.brookings.edu/articles/effective-ai-regulation-requires-understanding-general-purpose-ai/.

[114] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334):183–186, 2017.

[115] Ryan Calo. The Boundaries of Privacy Harm. *Indiana Law Journal*, 86:1131, 2011.

[116] Ryan Calo. Artificial intelligence policy: a primer and roadmap. *University of California, Davis Law Review*, 51:399, 2017.

[117] Sabrina Campano, Jessica Durand, and Chloé Clavel. Comparative analysis of verbal alignment in human-human and human-agent interactions. In *LREC*, pages 4415–4422. Citeseer, 2014.

[118] Angela J. Campbell. Self-regulation and the Media. *Federal Communication Law Journal*, 51:711, 1998.

[119] Robyn Caplan and Tarleton Gillespie. Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, 6(2): 2056305120936636, 2020. doi: 10.1177/2056305120936636. URL https://doi.org/10.1177/2056305120936636.

[120] Benjamin N. Cardozo and Andrew L. Kaufman. *The nature of the judicial process*. Quid Pro Books, 2010.

[121] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX Security Sympo-*

sium (USENIX Security 21), pages 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[122] Anupam Chander. The Racist Algorithm? *Michican Law Review*, 115:1023, 2017.

[123] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021. doi: 10.48550/ARXIV.2107.03374.

[124] Quan Ze Chen and Amy X Zhang. Case Law Grounding: Aligning Judgments of Humans and AI on Socially-Constructed Concepts. *arXiv preprint arXiv:2310.07019*, 2023.

[125] Inyoung Cheong. Freedom of Algorithmic Expression. *University of Cincinnati Law Review*, 91:680, 2022.

[126] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Safeguarding Human Values: Rethinking US Law for Generative AI's Societal Impacts. *AI and Ethics*, May 2024. ISSN 2730-5961. doi: 10.1007/s43681-024-00451-4. URL https://doi.org/10.1007/s43681-024-00451-4.

[127] Sang-Hun Choi. South korea shelves 'fake news' bill amid international outcry. *New York Times*, October 2021. URL https://www.nytimes.com/2021/10/01/world/asia/south-korea-fake-news-law.html.

[128] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310. Curran Associates Inc., Red Hook, NY, USA, 2017. ISBN 9781510860964.

[129] Danielle Keats Citron. Sexual Privacy. *The Yale Law Journal*, 128:1870, 2019.

[130] Danielle Keats Citron and Mary Anne Franks. The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform. *University of Chicago Legal Forum*, 2020:45, 2020.

[131] Danielle Keats Citron and Helen Norton. Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. *Boston University Law Review*, 91:1435, 2011.

[132] Danielle Keats Citron and Frank Pasquale. The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89:1, 2014.

[133] Danielle Keats Citron and Daniel J Solove. Privacy Harms. *Boston University Law Review*, 102:793, 2022.

[134] Ignacio N. Cofone and Adriana Z. Robertson. Privacy harms. *Hastings Law Journal*, 69:1039, 2017.

[135] Cary Coglianese and Evan Mendelson. Meta-regulation and self-regulation. In Robert Baldwin, Martin Cave, and Martin Lodge, editors, *The Oxford Handbook of Regulation*, page 0. Oxford University Press, September 2010. ISBN 978-0-19-956021-9. doi: 10.1093/oxfordhb/9780199560219.003.0008. URL https://doi.org/10.1093/oxfordhb/9780199560219.003.0008.

[136] Eric Corbett, Emily Denton, and Sheena Erete. Power and public participation in ai. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703812. doi: 10.1145/3617694.3623228. URL https://doi.org/10.1145/3617694.3623228.

[137] Sasha Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press, Cambridge, MA London, March 2020. ISBN 978-0-262-04345-8.

[138] Ian Cram. The Danish Cartoons, Offensive Expression, and Democratic Legitimacy. In *Extreme Speech and Democracy*, pages 289–310. Oxford University Press Oxford, 2009.

[139] Kate Crawford and Jason Schultz. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review*, 55:93, 2014.

[140] Rowan Cruft. Is There a Right to Internet Access? In Carissa Véliz, editor, *The Oxford Handbook of Digital Ethics*. Oxford University Press, Oxford, 2022. ISBN 978-0-19-885781-5. doi: 10.1093/oxfordhb/9780198857815.013.4.

[141] Mary J. Culnan. Protecting privacy online: Is self-regulation working? *Journal of Public Policy & Marketing*, 19(1):20–26, 2000.

[142] John Danaher. The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, 29(3):245–268, 2016.

[143] Aaron Davis and Jane Andrew. From rationalism to critical pragmatism: revisiting arnstein's ladder of public participation in co-creation and consultation. 8th State of Australian Cities National Conference, 28-30 November 2017, Adelaide, Australia, June 2018. URL `https://apo.org.au/node/178271`.

[144] Simon De Deyne, Amy Perfors, and Daniel J Navarro. 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*, pages 1861–1870, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://aclanthology.org/C16-1175`.

[145] Edward L. Deci. *The Psychology of Self-Determination*. Lexington Books, 1980.

[146] Edward L. Deci and Richard M. Ryan. The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, 19(2):109–134, June 1985. ISSN 0092-6566. doi: 10.1016/0092-6566(85)90023-6.

[147] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. Stakeholder Participation in AI: Beyond "Add Diverse Stakeholders and Stir"'. *arXiv preprint arXiv:2111.01122*, 2021.

[148] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703812. doi: 10.1145/3617694.3623261. URL https://doi.org/10.1145/3617694.3623261.

[149] Derek A. Denckla. Nonlawyers and the unauthorized practice of law: an overview of the legal and ethical parameters. *Fordham Law Review*, 67:2581, 1998.

[150] Yang Deng, Yaliang Li, Wenxuan Zhang, Bolin Ding, and Wai Lam. Toward personalized answer generation in e-commerce via multi-perspective preference modeling. *ACM Trans. Inf. Syst.*, 40 (4), mar 2022. ISSN 1046-8188. doi: 10.1145/3507782. URL https://doi.org/10.1145/3507782.

[151] Anokhy Desai. US State Privacy Legislation Tracker. *IAPP*, May 2023. URL https://iapp.org/resources/article/us-state-privacy-legislation-tracker/.

[152] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2342–2351, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534647. URL https://doi.org/10.1145/3531146.3534647.

[153] Carl DiSalvo, Illah Nourbakhsh, David Holstius, Ayça Akin, and Marti Louw. The Neighborhood Networks project: a case study of critical engagement and creative expression through participatory design. In *Proceedings of the tenth anniversary conference on participatory design 2008*, pages 41–50, 2008.

[154] Evelyn Douek. The rise of content cartels. *Knight First Amendment Institute at Columbia*, 2020. URL https://knightcolumbia.org/content/the-rise-of-content-cartels.

[155] Evelyn Douek. The Meta Oversight Board and the Empty Promise of Legitimacy. *Harvard Journal of Law & Technology*, 37, 2023.

[156] Scott Douglas, Chris Ansell, Charles F Parker, Eva Sørensen, Paul 'T Hart, and Jacob Torfing. Understanding collaboration: Introducing the collaborative governance case databank. *Policy and Society*, 39(4):495–509, 2020.

[157] Christine M. Drennon. 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.

[158] Maeve Duggan. Online harassment 2017. 2017.

[159] Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards Measuring the Representation of Subjective Global Opinions in Language Models, 2023.

[160] Brenda Dvoskin. Expertise and participation in the facebook oversight board: From reason to will. *Telecommunications Policy*, 47(5):102463, 2023.

[161] Pelle Ehn. *Work-oriented design of computer artifacts*. PhD thesis, Arbetslivscentrum, 1988.

[162] Kirk Emerson, Tina Nabatchi, and Stephen Balogh. An integrative framework for collaborative governance. *Journal of public administration research and theory*, 22(1):1–29, 2012.

[163] European Commission. 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, apr 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. COM(2021) 206 final 2021/0106(COD).

[164] Joshua Evans. Trials and tribulations: Problematizing the city through/as urban experimentation. *Geography Compass*, 10(10):429–443, 2016.

[165] Ernesto Falcon. Congress must exercise caution in ai regulation, May 2023. URL https://www.eff.org/deeplinks/2023/05/congress-must-exercise-caution-ai-regulation.

[166] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22, 2021. doi: 10.48550/ARXIV.2010.11125. URL https://arxiv.org/abs/2010.11125.

[167] Jenny Fan and Amy X Zhang. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

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

[169] FDA. Artificial Intelligence and Machine Learning (AI/ML) for Drug Development. https://www.fda.gov/science-research/science-and-research-special-topics/artificial-intelligence-and-machine-learning-aiml-drug-development, May 2023.

[170] Jürgen Feick and Raymund Werle. Regulation of cyberspace. In *The Oxford Handbook of Regulation*. Oxford University Press, September 2010. ISBN 978-0-19-956021-9. doi: 10.1093/oxfordhb/9780199560219.003.0021. URL https://doi.org/10.1093/oxfordhb/9780199560219.003.0021.

[171] Noah Feldman. Free Speech in Europe Isn't What Americans Think. *Bloomberg.com*, Mar 2017. URL https://www.bloomberg.com/view/articles/2017-03-19/free-speech-in-europe-isn-t-what-americans-think.

[172] K. J. Kevin Feng, Quan Ze Chen, Inyoung Cheong, King Xia, and Amy X. Zhang. Case repositories: Towards case-based reasoning for ai alignment. *arXiv preprint arXiv:2311.10934*.

[173] Franklin Foer. Facebook's war on free will. https://www.theguardian.com/technology/2017/sep/19/facebooks-war-on-free-will, September 2017.

[174] International Association for Public Participation (IAP2). IAP2 Spectrum of Public Participation. https://cdn.ymaws.com/www.iap2.org/resource/resmgr/communications/11x17_p2_pillars_brochure_20.pdf.

[175] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 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)*, pages 653–670, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.48. URL https://aclanthology.org/2020.emnlp-main.48.

[176] Korea Press Foundation. Newspaper industry in korea 2022. https://www.kpf.or.kr/front/research/realityDetail.do?miv_pageNo=&miv_pageSize=&total_cnt=&LISTOP=&mode=W&seq=593062&link_g_topmenu_id=676f2f0f377b4b19840685a46f69a233&link_g_submenu_id=fc923e2179d547c2a0a9762070102039&link_g_homepage=F&reg_stadt=&reg_enddt=&searchkey=all1&searchtxt=, 2022.

[177] Brian Frederick. AI Allows You To Talk With Virtual Versions Of Deceased Loved Ones. *Search Engine Journal*, October 2022. URL https://www.searchenginejournal.com/ai-allows-you-to-talk-with-virtual-versions-of-deceased-loved-ones/468761/#close.

[178] Jody Freeman. Collaborative Governance in the Administrative State. *UCLA Law Review*, 45:1, 1997.

[179] Charles Fried. Privacy: Economics and Ethics: A Comment on Posner. *Georgia Law Review*, 12: 423, 1978.

[180] Thomas L. Friedman. *The World is Flat: A Brief History of the Twenty-first Century*. 2005.

[181] Jonathan Friendly. National News Council Will Dissolve. *The New York Times*, May 1984.

[182] Masato Fukushima. The experimental zone of learning: Mapping the dynamics of everyday experiment. *Mind, Culture, and Activity*, 24(4):311–323, 2017.

[183] Francis Fukuyama. What is governance? *Governance*, 26(3):347–368, 2013.

[184] Robert K Fullinwider. Philosophy, casuistry, and moral development. *Theory and Research in Education*, 8(2):173–185, 2010.

[185] Richard Futrell and Roger P Levy. Do rnns learn human-like abstract word order preferences? *arXiv preprint arXiv:1811.01866*, 2018.

[186] Iason Gabriel. Artificial Intelligence, Values and Alignment. *Minds and Machines*, 30(3):411–437, Sep 2020. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-020-09539-2. arXiv:2001.09768 [cs].

[187] Marylène Gagné. A model of knowledge-sharing motivation. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, 48(4):571–589, 2009.

[188] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint arXiv:2209.07858*.

[189] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2022.

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

[191] Robert Gellman and Pam Dixon. Many failures: A brief history of privacy self-regulation in the united states. In *World Privacy Forum*, pages 1–29. World Privacy Forum, 2011.

[192] Sourojit Ghosh and Aylin Caliskan. 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.

[193] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[194] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements.

[195] Eric Goldman. Snapchat Defeats Lawsuit Over User-to-User Harassment-Ziencik v. Snap. *Technology & Marketing Law Blog*, Feb 2023. URL `https://blog.ericgoldman.org/archives/2023/02/snapchat-defeats-lawsuit-over-user-to-user-harassment-ziencik-v-snap.htm`.

[196] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246*, 2023.

[197] Robert Gorwa. The platform governance triangle: conceptualising the informal regulation of online content. *Internet Policy Review*, 8(2), June 2019. ISSN 2197-6775. URL

https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content.

[198] Candida M. Greco and Andrea Tagarelli. Bringing order into the realm of transformer-based language models for artificial intelligence and law. *arXiv preprint arXiv:2308.05502*, 2023.

[199] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, page 79–90, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702600. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

[200] Thomas C Grey. Langdell's orthodoxy. *University of Pittsburgh Law Review*, 45:1, 1983.

[201] James Grimmelmann. The Virtues of Moderation. *Yale Journal of Law & Technology*, 17:42, 2015.

[202] Lara Groves, Aidan Peppin, Andrew Strait, and Jenny Brennan. Going public: the role of public participation approaches in commercial AI labs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1162–1173, 2023.

[203] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv preprint arXiv:2301.07597*.

[204] Wei Guo and Aylin Caliskan. 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*, pages 122–133, 2021.

[205] Hiroki Habuka. Japan's approach to ai regulation and its impact on the 2023 g7 presidency. *Center for Strategies & International Studies*, February 2023. URL https://www.csis.org/analysis/japans-approach-ai-regulation-and-its-impact-2023-g7-presidency.

[206] Thilo Hagendorff and Sarah Fabi. Methodological reflections for AI alignment research using human feedback. *arXiv preprint arXiv:2301.06859*, 2022.

[207] Christina Harrington, Sheena Erete, and Anne Marie Piper. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[208] Claudia E. Haupt. Artificial Professional Advice. *Yale Journal of Law and Technology*, 21:55–77, 2019.

[209] Claudia E Haupt. Regulating Speech Online: Free Speech Values in Constitutional Frames. *Washington University Law Review*, 99:751, 2021.

[210] Peter Henderson, Jieru Hu, Mona Diab, and Joelle Pineau. Rethinking machine learning benchmarks in the context of professional codes of conduct. In *Proceedings of the Symposium on Computer Science and Law*, CSLAW '24, page 109–120, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703331. doi: 10.1145/3614407.3643708. URL https://doi.org/10.1145/3614407.3643708.

[211] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI With Shared Human Values. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dNy_RKzJacY.

[212] Benjamin Herold. 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, April 2020.

[213] Chris Jay Hoofnagle. Privacy self regulation: A decade of disappointment. *Consumer Protection in the Age of the'Information Economy'(Jane K. Winn, ed.)(Ashgate 2006)*, 2005.

[214] Betty Li Hou and Brian Patrick Green. A multi-level framework for the ai alignment problem. *arXiv preprint arXiv:2301.03740*, 2023.

[215] The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 2023. URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[216] Goran Hyden. Governance and the reconstitution of political order. In *State, Conflict, and Democracy in Africa*, page 179–196. Lynne Rienner Publishers, April 1998. ISBN 978-1-68585-182-8. doi: 10.1515/9781685851828-011. URL https://www.degruyter.com/document/doi/10.1515/9781685851828-011/pdf?licenseType=restricted.

[217] Leonardo Horn Iwaya, Muhammad Ali Babar, and Awais Rashid. Privacy Engineering in the Wild: Understanding the Practitioners' Mindset, Organisational Culture, and Current Practices, 2022. URL https://arxiv.org/abs/2211.08916.

[218] 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. ISBN 978-1-4503-XXXX-X. doi: 10.1145/3544548.3581196. URL https://doi.org/10.1145/3544548.3581196.

[219] Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(1):100105, 2023. ISSN 2772-4859. doi: https://doi.org/10.1016/j.tbench.2023.100105.

[220] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 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.

[221] Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. CommunityLM: Probing Partisan Worldviews from Language Models. *arXiv preprint arXiv:2209.07065*, 2022.

[222] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards Machine Ethics and Norms. *arXiv preprint arXiv:2110.07574*, 2021.

[223] Stefan Johansson, Per-Olof Hedvall, Mia Larsdotter, Thomas P. Larsson, and Catharina Gustavsson. Co-Designing with Extreme Users: A Framework for User Participation in Design Processes. *Scandinavian Journal of Disability Research*, 25(1), December 2023. ISSN 1745-3011. doi: 10.16993/sjdr.952. URL https://sjdr.se/articles/10.16993/sjdr.952.

[224] Albert R. Jonsen. Casuistry and clinical ethics. *Theoretical Medicine*, 7:65–74, 1986.

[225] David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:174799410.

[226] Robert A Kagan. *Adversarial Legalism: The American Way of Law*. Harvard University Press, Cambridge, 2019.

[227] Margot E. Kaminski. Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability. *92 Southern California Law Review 1529*, 2019.

[228] Margot E. Kaminski. Regulating the Risks of AI. *Boston University Law Review*, 103, 2023.

[229] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[230] Sayash Kapoor, Peter Henderson, and Arvind Narayanan. Promises and pitfalls of artificial intelligence for legal applications. *Journal of Cross-disciplinary Research in Computational Law*, 2(22), May 2024. ISSN 2736-4321. URL https://journalcrcl.org/crcl/article/view/62.

[231] Enkelejda Kasneci, Stefan Küchemann Kathrin Sessler, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Ku-

tyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 2023. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2023.102274.

[232] David Kaye. *Speech Police: The Global Struggle to Govern the Internet*. 2019.

[233] Christopher M Kelty. Too much democracy in all the wrong places: toward a grammar of participation. *Current Anthropology*, 58(S15):S77–S90, 2017.

[234] Finn Kensing and Jeanette Blomberg. Participatory design: Issues and concerns. *Computer supported cooperative work (CSCW)*, 7:167–185, 1998.

[235] Lina M Khan. Amazon's antitrust paradox. *Yale Law Journal*, 126:710, 2016.

[236] Regina Kim. Why so many of your favorite k-dramas are based on webtoons. `https://time.com/6243447/rise-of-webtoons-k-dramas/`, December 2022.

[237] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. Humans, ai, and context: Understanding end-users' trust in a real-world computer vision application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 77–88, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593978. URL `https://doi.org/10.1145/3593013.3593978`.

[238] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 385–404, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705083. doi: 10.1145/3640543.3645148. URL `https://doi.org/10.1145/3640543.3645148`.

[239] Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Rottger, and Scott Hale. The past, present and better future of feedback learning in large language models for subjective human preferences and

values. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.148. URL https://aclanthology.org/2023.emnlp-main.148.

[240] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The empty signifier problem: Towards clearer paradigms for operationalising" alignment" in large language models. *arXiv preprint arXiv:2310.02457*, 2023.

[241] Anne Mette Kjaer. *Governance*. John Wiley & Sons, February 2023. ISBN 978-1-5095-6062-2.

[242] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018.

[243] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131:1598, 2017.

[244] Kate Klonick. The facebook oversight board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal*, 129:2418, 2019.

[245] Kate Klonick. The facebook oversight board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal*, 129(2418), 2020.

[246] Sarah Knuckey, Joshua D Fisher, Amanda M Klasing, Tess Russo, and Margaret L Satterthwaite. Advancing Socioeconomic Rights Through Interdisciplinary Factfinding: Opportunities and Challenges. *Annual Review of Law and Social Science*, 17:375–389, 2021.

[247] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

[248] Janet L Kolodner. An introduction to case-based reasoning. *Artificial intelligence review*, 6(1):3–34, 1992.

[249] Adam B. Korn, Sebastian A. Navarro, and Todd Rosenbaum. 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.

[250] Morten Kyng. Bridging the gap between politics and techniques: On the next practices of participatory design. *Scandinavian Journal of Information Systems*, 22(1):5, 2010.

[251] Karim R Lakhani and Robert G Wolf. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Open Source Software Projects (September 2003)*, 2003.

[252] Bruno Latour and Peter Weibel. *Making things public: Atmospheres of democracy*. 2005.

[253] Mason Lawler. State Appeals Court Allows Design-Defect Claims Against Snapchat to Proceed. *Law.com*, January 2023. URL https://www.law.com/dailyreportonline/2023/01/30/state-appeals-court-allows-design-defect-claims-against-snapchat-to-proceed/?slreturn=20230214085923.

[254] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[255] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35, 2019.

[256] Kalev Leetaru. Could personalized content moderation be the future of healthy social media? *Forbes*, July 2019.

[257] Kalev Leetaru. Could digital assistants be our personalized toxicity filters for social media? *Forbes*, June 2019.

[258] Alina Leidinger and Richard Rogers. Which stereotypes are moderated and under-moderated in search engine autocompletion? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1049–1061, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594062. URL https://doi.org/10.1145/3593013.3594062.

[259] Mark A. Lemley. The contradictions of platform regulation. *Journal of Free Speech*, 1, 2021.

[260] Lawrence Lessig. *Code version 2.0*. Basic Books, New York, 2006.

[261] Calvin A Liang, Sean A Munson, and Julie A Kientz. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(2):1–47, 2021.

[262] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[263] Cristiano Lima. AI chatbots won't enjoy tech's legal shield, Section 230 authors say. *The Washington Post*, March 2023. URL https://www.washingtonpost.com/politics/2023/03/17/ai-chatbots-wont-enjoy-techs-legal-shield-section-230-authors-say/.

[264] Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. Decision-Oriented Dialogue for Human-AI Collaboration. *arXiv preprint arXiv:2305.20076*, 2023.

[265] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[266] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 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)*,

pages 6691–6706, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522.

[267] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866, 2021.

[268] Natasha Lomas. Who's liable for AI-generated lies? *TechCrunch*, June 2022. URL https://techcrunch.com/2022/06/01/whos-liable-for-ai-generated-lies/.

[269] Jamie Lorimer and Clemens Driessen. Wild experiments at the oostvaardersplassen: rethinking environmentalism in the anthropocene. *Transactions of the Institute of British Geographers*, 39(2): 169–181, 2014.

[270] Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Towards Boosting the Open-Domain Chatbot with Human Feedback. *arXiv preprint arXiv:2208.14165*, 2022.

[271] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. volume 2023, page 1105–1114, January 2024.

[272] Mark MacCarthy. Broadcast self-regulation: The nab codes, family viewing hour, and television violence. *Family Viewing Hour, and Television Violence*, 13, 1995.

[273] John Leslie Mackie. *Hume's moral theory*. Routledge, 2003.

[274] James Madison. 47. The Alleged Danger from the Powers of the Union to the State Governments Considered. In *The Federalist Papers*, page 209. Open Road Integrated Media, Inc., New York, 2022.

[275] Liviu Octavian Mafteiu-Scai. A new approach for solving equations systems inspired from brainstorming. *International Journal of New Computer Architectures and Their Applications*, 5(1):10+, Jan 2015. URL https://link.gale.com/apps/doc/A419413111/AONE?u=wash_main&sid=bookmark-AONE&xid=37d45dbf.

[276] Kaitlin Mahar, Amy X Zhang, and David Karger. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[277] Laura Malinverni, Joan Mora-Guiard, and Narcis Pares. Towards methods for evaluating and communicating participatory design: A multimodal approach. *International Journal of Human-Computer Studies*, 94:53–63, 2016. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2016.03.004.

[278] Ezio Manzini and Francesca Rizzo. Small projects/large changes: Participatory design as an open participated process. *CoDesign*, 7(3–4):199–215, September 2011. ISSN 1571-0882. doi: 10.1080/15710882.2011.630472.

[279] Sandra G Mayson. Bias in, bias out. *The Yale Law Journal*, 128:2218, 2019.

[280] Sean McKeown and William J Buchanan. Hamming distributions of popular perceptual hashing techniques. *arXiv preprint arXiv:2212.08035*, 2022.

[281] Amre Metwally. "outside experts": Expertise and the counterterrorism industry in social media content moderation. *Journal of National Security Law & Policy*, 12, 2022.

[282] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. Rethinking search: Making domain experts out of dilettantes. In *ACM SIGIR Forum*, volume 55, pages 1–27, New York, NY, USA, 2021. ACM.

[283] Jeffrey Metzler. The importance of irac and legal writing. *University of Detroit Mercy Law Review*, 80:501–504, 2002.

[284] Dan Milmo. AI firms must be held responsible for harm they cause, 'godfathers' of technology say. *The Guardian*, October 2023. URL `https://www.theguardian.com/technology/2023/oct/24/ai-firms-must-be-held-responsible-for-harm-they-cause-godfathers-of-technology-say`.

[285] Newton Minow and Martha Minow. Social Media Companies Should Pursue Serious Self-Supervision-Soon: Response to Professors Douek and Kadri. *Harvard Law Review Forum*, 136: 428, 2022.

[286] Hiroshi Miyashita. A tale of two privacies: enforcing privacy with hard power and soft power in japan. *Enforcing Privacy: Regulatory, Legal and Technological Approaches*, pages 105–122, 2016.

[287] Stephen P. Mulligan and Chris D. Linebaugh. Data Protection and Privacy Law: An Introduction. *Congressional Research Service*, IF11207, October 2022. URL https://crsreports.congress.gov/product/details?prodcode=IF11207.

[288] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, 2022.

[289] John J. Nay. Large language models as corporate lobbyists. *arXiv preprint arXiv:2301.01181*, 2023.

[290] Anam Nazir and Ze Wang. A comprehensive survey of chatgpt: Advancements, applications, prospects, and challenges. *Meta-radiology*, 1(2):100022, September 2023. ISSN 2950-1628. doi: 10.1016/j.metrad.2023.100022.

[291] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*, 2020.

[292] Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1504–1515, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533205. URL https://doi.org/10.1145/3531146.3533205.

[293] Riku Neuvonen and Esa Sirkkunen. Outsourced justice: The case of the facebook oversight board. *Journal of Digital Media & Policy*, 2022.

[294] Gavin Northey, Vanessa Hunter, Rory Mulcahy, Kelly Choong, and Michael Mehmet. Man vs ma-

chine: how artificial intelligence in banking influences consumer belief in financial advice. *International Journal of Bank Marketing*, 40(6):1182–1199, 2022.

[295] Zainab Nururrohmah et al. Shared-power governance in managing common pool resources case study: collaborative planning to manage thematic parks in Bandung City, Indonesia. *Procedia-Social and Behavioral Sciences*, 227:465–476, 2016.

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

[297] Han Kun Law Offices. CAC releases guidelines for China SCC filings, June 2023. URL https://www.lexology.com/library/detail.aspx?g=9b37881f-52f2-4c9d-99ed-d7d769e8dbf4.

[298] Rosemary O'Leary and Lisa B Bingham. *The promise and performance of environmental conflict resolution*. Resources for the Future, 2003.

[299] Rosemary O'Leary, Catherine Gerard, and Lisa Blomgren Bingham. Introduction to the symposium on collaborative public management. *Public administration review*, 66:6–9, 2006.

[300] Kent C Olson, Aaron S Kirschenfeld, and Ingrid Mattson. *Principles of legal research*. West Academic Publishing, Eagan, 2015.

[301] Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. 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.

[302] Anne O'Rourke. Caring about virtual perts: an ethical interpretation of Tamagotchi. *Animal Issues*, 2(1), 1998. URL https://ro.uow.edu.au/ai/vol2/iss1/1.

[303] Tonje C Osmundsen, Kine M Karlsen, Roy Robertsen, and Bjørn Hersoug. Shared waters—shared problems: The role of self-governance in managing common pool resources. *Aquaculture Economics & Management*, 25(3):275–297, 2021.

[304] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press, 1990.

[305] Elinor Ostrom. *Understanding institutional diversity*. Princeton University Press, Princeton, NJ, 2005.

[306] Elinor Ostrom. Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, 100(3), 2010.

[307] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[308] Kentrell Owens, Johanna Gunawan, David Choffnes, Pardis Emami-Naeini, Tadayoshi Kohno, and Franziska Roesner. Exploring Deceptive Design Patterns in Voice Interfaces. In *Proceedings of the 2022 European Symposium on Usable Security*, EuroUSEC '22, page 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.

[309] Christina A Pan, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–31, 2022.

[310] Monica Lestari Paramita, Maria Kasinidou, Styliani Kleanthous, Paolo Rosso, Tsvi Kuflik, and Frank Hopfgartner. Towards improving user awareness of search engine biases: A participatory design approach. *Journal of the Association for Information Science and Technology*, 2023.

[311] Chang Sup Park. Online speech and democratic culture: a comparison of freedom of online speech between south korea and the united states. *Asian Journal of Communication*, 26(3):262–277, 2016.

[312] Elena Parmiggiani and Miria Grisot. Data infrastructures in the public sector: A critical research agenda rooted in scandinavian is research. *978-0-578-53212-7*, 2019. URL `https://www.duo.uio.no/handle/10852/80628`. Accepted: 2020-10-15T19:25:01Z.

[313] Orlando Patterson. *Freedom: Volume I: Freedom In The Making Of Western Culture*. Basic Books, New York, N.Y., 1992. ISBN 978-0-465-02532-9.

[314] Norbert Paulo. Casuistry as common law morality. *Theoretical Medicine and Bioethics*, 36(6):373–389, 2015.

[315] Camille Sojit Pejcha. Tiktok's "mind-reading" algorithm is about to change. *Document Journal*, Aug 2023. URL `https://www.documentjournal.com/2023/08/tiktok-personalized-algorithm-opt-out-cognitive-liberty-dsa-neurotechnology/`.

[316] Denis Peskoff and Brandon Stewart. Credible without credit: Domain experts assess generative language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.37. URL `https://aclanthology.org/2023.acl-short.37`.

[317] Jon Pierre and B. Guy Peters. *Governance, politics and the state*. Bloomsbury Publishing, 2020.

[318] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A Human Rights-Based Approach to Responsible AI. *arXiv preprint arXiv:2210.02667*, 2022.

[319] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv preprint arXiv:2310.03693*, 2023.

[320] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. Queer in ai: A case study in community-led participatory ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1882–1895, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594134. URL https://doi.org/10.1145/3593013.3594134.

[321] Joaquin Quiñonero Candela, Yuwen Wu, Brian Hsu, Sakshi Jain, Jennifer Ramos, Jon Adams, Robert Hallman, and Kinjal Basu. Disentangling and operationalizing ai fairness at linkedin. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1213–1228, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594075. URL https://doi-org.offcampus.lib.washington.edu/10.1145/3593013.3594075.

[322] Octavia Reeve, Anna Colom, and Roshni Modhvadia. What do the public think about AI? *Ada Lovelace Institute*, Oct 2023. URL https://www.adalovelaceinstitute.org/evidence-review/what-do-the-public-think-about-ai/.

[323] Reuters. Australian mayor prepares world's first defamation lawsuit over ChatGPT content. *The Guardian*, April 2023. URL https://www.theguardian.com/technology/2023/apr/

06/australian-mayor-prepares-worlds-first-defamation-lawsuit-over-chatgpt-content.

[324] Pedro Reynolds-Cuéllar and Daniela Delgado Ramos. Community-based technology co-design: Insights on participation, and the value of the "co". In *Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 1*, PDC '20, page 75–84, New York, NY, USA, June 2020. Association for Computing Machinery. ISBN 978-1-4503-7700-3. doi: 10.1145/3385010.3385030. URL https://doi.org/10.1145/3385010.3385030.

[325] Neil Richards. *Intellectual privacy: Rethinking civil liberties in the digital age*. Oxford University Press, USA, Oxford, 2015.

[326] Neil Richards and Woodrow Hartzog. A Duty of Loyalty for Privacy Law. *Washington University Law Review*, 99:961, 2021.

[327] Kimberly Jenkins Robinson. Designing the Legal Architecture to Protect Education as a Civil Right. *Indiana Law Journal*, 96(1):51, 2020. URL https://ssrn.com/abstract=4054077.

[328] Les Robinson. Public outrage and public trust: A road map for public involvement in waste management decision-making. https://www.enablingchange.com.au/Public_outrage_public_trust.pdf, October 2002.

[329] Kat Roemmich, Florian Schaub, and Nazanin Andalibi. 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, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580950. URL https://doi.org/10.1145/3544548.3580950.

[330] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 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, page 12, USA, 2012. USENIX Association.

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

[332] Mathew Rotenberg. Stifled justice: The unauthorized practice of law and internet legal resources. *Minnesota Law Review*, 97:709–742, 2012.

[333] Alan Z Rozenshtein. Moderating the fediverse: Content moderation on distributed social media. *Journal of Free Speech*, 2, 2023.

[334] Lilach Sagiv, Sonia Roccas, Jan Cieciuch, and Shalom H Schwartz. Personal values in human life. *Nature human behaviour*, 1(9):630–639, 2017.

[335] Johnny Saldaña. *The Coding Manual for Qualitative Researchers (4th ed.)*. SAGE Publications, Los Angeles, 2021.

[336] Malik Sallam. Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare (Basel, Switzerland)*, 11(6):887, March 2023. ISSN 2227-9032. doi: 10.3390/healthcare11060887.

[337] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[338] Czarina Saloma, Lorraine Mangaser, and Candy Hidalgo. Expecting the unexpected: the role of surprise in community-driven development. *Community Development Journal*, 52(4):702–719, 2017.

[339] Princess Sampson, Ro Encarnacion, and Danaë Metaxa. Representation, self-determination, and refusal: Queer people's experiences with targeted advertising. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1711–1722, 2023.

[340] Nathan Sanders, Bruce Schneier, and Norman Eisen. How public ai can strengthen democracy. `https://www.brookings.edu/articles/how-public-ai-can-strengthen-democracy/`, March 2024.

[341] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman

Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[342] Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback. *arXiv preprint arXiv:2204.14146*, 2022.

[343] Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. Intersectional hci: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 5412–5427, 2017.

[344] Lisa Schmidthuber, Frank Piller, Marcel Bogers, and Dennis Hilgers. Citizen participation in public administration: investigating open government for social innovation. *R&d Management*, 49(3):343–355, 2019.

[345] Amy J Schmitz and John Zeleznikow. Intelligent legal tech to empower self-represented litigants. *Ohio State Legal Studies Research Paper*, 23:142–191, 2022.

[346] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1350–1361, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533192. URL https://doi.org/10.1145/3531146.3533192.

[347] Joseph Seering. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4 (CSCW2):1–28, 2020.

[348] Philip Selznick. Focusing organizational research on regulation. *Regulatory policy and the social sciences*, 1(1):363–367, 1985.

[349] Olga Seminck and Pascal Amsili. 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*, pages 53–63, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-4006.

[350] Amartya Sen. *Development as Freedom*. Knopf Doubleday Publishing Group, New York, 2011. ISBN 978-0-307-87429-0.

[351] Amartya Sen. Elements of a theory of human rights. In *Justice and the capabilities approach*, page 320. Routledge, Oxfordshire, 2017.

[352] De Agência Senado. Comissão da Inteligência Artificial aprova plano de trabalho, September 2023. URL https://www12.senado.leg.br/noticias/materias/2023/09/12/comissao-da-inteligencia-artificial-aprova-plano-de-trabalho.

[353] Nandana Sengupta, Ashwini Vaidya, and James Evans. In her shoes: Gendered labelling in crowdsourced safety perceptions data from india. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 183–192, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593987. URL https://doi-org.offcampus.lib.washington.edu/10.1145/3593013.3593987.

[354] Congressional Research Service. Federal Financial Assistance and Civil Rights Requirements. CRS Report, May 2022. URL https://crsreports.congress.gov.

[355] Aaditeshwar Seth, Akshay Gupta, Aparna Moitra, Deepak Kumar, Dipanjan Chakraborty, Lamuel Enoch, Orlanda Ruthven, Paramita Panjal, Rafi Ahmad Siddiqi, Rohit Singh, et al. Reflections from practical experiences of managing participatory media platforms for development. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*, pages 1–15, 2020.

[356] Richard Severo. Kenneth Clark, Who Fought Segregation, Dies. *The New York Times*, May 2005. ISSN 0362-4331. URL https://www.nytimes.com/2005/05/02/nyregion/kenneth-clark-whofought-segregation-dies.html.

[357] Chirag Shah and Emily M. Bender. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, CHIIR '22, page 221–232, New York, NY, USA, 2022.

Association for Computing Machinery. ISBN 9781450391863. doi: 10.1145/3498366.3505816. URL https://doi.org/10.1145/3498366.3505816.

[358] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, pages 1–6, 2023.

[359] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-play with large language models. *arXiv preprint arXiv:2305.16367*, 2023.

[360] Nikhil Sharma. Regulating ai is a mistake, September 2023. URL https://www.michigandaily.com/opinion/regulating-ai-is-a-mistake/.

[361] Hong Shen, Leijie Wang, Wesley H Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. The model card authoring toolkit: Toward community-centered, deliberation-driven ai design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 440–451, 2022.

[362] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 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)*, page 3407–3412, 2019. doi: 10.18653/v1/D19-1339. URL https://aclanthology.org/D19-1339.

[363] Clea Simon. How COVID taught America about inequity in education. *The Harvard Gazette*, July 2021. URL https://news.harvard.edu/gazette/story/2021/07/how-covid-taught-america-about-inequity-in-education/.

[364] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. 2023.

[365] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6. 2022.

[366] Adam Smith. The theory of moral sentiments, ed. dd raphael and al macfie, 1976. VII, iv, 34.

[367] Centaine L Snoswell, Aaron J Snoswell, Jaimon T Kelly, Liam J Caffery, and Anthony C Smith. Artificial intelligence: Augmenting telehealth with large language models. *Journal of telemedicine and telecare*, page 1357633X231169055, 2023.

[368] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949*, 2023.

[369] Taylor Soper. Microsoft vets lead secretive education startup using generative AI to help students learn. *GeekWire*, June 2023. URL https://www.msn.com/en-us/news/technology/microsoft-vets-lead-secretive-education-startup-using-generative-ai-to-help-students-learn/ar-AA1bWdoH.

[370] Thomas E. Spahn. Is your artificial intelligence guilty of the unauthorized practice of law. *Richmond Journal of Law and Technology*, 24:1–47, 2017.

[371] Stanford's Global Digital Policy Incubator and ARTICLE 19 and David Kaye. Social Media Councils: From Concept to Reality. https://fsi.stanford.edu/content/social-media-councils-concept-reality-conference-report, 2019.

[372] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1162–1177, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533177. URL https://doi.org/10.1145/3531146.3533177.

[373] Iain Stewart. The critical legal science of Hans Kelsen. *Journal of Law & Society*, 17:273, 1990.

[374] Michael Stockdale and Rebecca Mitchell. Legal advice privilege and artificial legal intelligence: Can robots give privileged legal advice? *The International Journal of Evidence & Proof*, 23(4): 422–439, 2019. doi: 10.1177/1365712719862296. URL https://journals.sagepub.com/doi/abs/10.1177/1365712719862296.

[375] Chase Stokes and Marti Hearst. Why More Text is (Often) Better: Themes from Reader Preferences for Integration of Charts and Text, 2022. URL https://arxiv.org/abs/2209.10789.

[376] Lucille Alice Suchman. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press, 2007.

[377] Sasha Fathima Suhel, Vinod Kumar Shukla, Sonali Vyas, and Ved Prakash Mishra. Conversation to automation in banking through chatbot using artificial machine intelligence language. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 611–618. IEEE, 2020.

[378] Cass R Sunstein. State Action is Always Present. *Chicago Journal of Internationall Law*, 3:465, 2002.

[379] Cass R Sunstein. *Legal reasoning and political conflict*. Oxford University Press, 2018.

[380] Cass R. Sunstein. The Administrative State, Inside Out. https://papers.ssrn.com/abstract=4069458, Mar 2022.

[381] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1609–1621, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658992. URL https://doi.org/10.1145/3630106.3658992.

[382] Kevin Yew Lee Tan. Fifty years of the universal declaration of human rights: A singapore reflection. *Sing. L. Rev.*, 20:239, 1999.

[383] Iddo Tavory and Stefan Timmermans. *Abductive analysis: Theorizing qualitative research*. University of Chicago Press, 2014.

[384] Jacob Thebault-Spieker, Sukrit Venkatagiri, Naomi Mine, and Kurt Luther. Diverse perspectives can mitigate political bias in crowdsourced content moderation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1280–1291, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594080. URL https://doi-org.offcampus.lib.washington.edu/10.1145/3593013.3594080.

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

[386] Global Internet Forum to Counter Terrorism. https://gifct.org/, 2022.

[387] Autumn Toney and Aylin Caliskan. 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*, pages 7203–7218, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.574. URL https://aclanthology.org/2021.emnlp-main.574.

[388] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross

Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[389] Kentaro Toyama. Why Technology Alone Won't Fix Schools. *The Atlantic*, June 2015. URL https://www.theatlantic.com/education/archive/2015/06/why-technology-alone-wont-fix-schools/394727/.

[390] Heidi Tworek. Social media councils. https://www.cigionline.org/articles/social-media-councils/, October 2019.

[391] UNESCO. Recommendation on the ethics of artificial intelligence. May 2023. URL https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.

[392] Josef Valvoda, Ryan Cotterell, and Simone Teufel. On the role of negative precedent in legal outcome prediction. *Transactions of the Association for Computational Linguistics*, 11:34–48, 2023.

[393] Siv Vangen, John Paul Hayes, and Chris Cornforth. Governing cross-sector, inter-organizational collaborations. *Public Management Review*, 17(9):1237–1260, 2015.

[394] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Clinical text summarization: Adapting large language models can outperform human experts. 2023.

[395] Eugene Volokh. *Academic Legal Writing: Law Review Articles, Student Notes, Seminar Papers, and Getting on Law Review*. Foundation Press, Eagan, 4th edition, 2010.

[396] Eugene Volokh. Large Libel Models? Liability for AI Output, 2023. URL https://www2.law.ucla.edu/volokh/ailibel.pdf.

[397] Ben Wagner, Krisztina Rozgonyi, Marie-Therese Sekwenz, Jennifer Cobbe, and Jatinder Singh. Regulating transparency? facebook, twitter and the german network enforcement act. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 261–271, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372856. URL https://doi-org.offcampus.lib.washington.edu/10.1145/3351095.3372856.

[398] Jun Wang, Chang Xu, Francisco Guzmán, Ahmed El-Kishky, Yuqing Tang, Benjamin IP Rubinstein, and Trevor Cohn. Putting words into the system's mouth: A targeted attack on neural machine translation using monolingual data poisoning. *arXiv preprint arXiv:2107.05243*, 2021.

[399] Leijie Wang and Haiyi Zhu. How are ml-based online content moderation systems actually used? studying community size, local activity, and disparate treatment. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 824–838, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533147. URL https://doi.org/10.1145/3531146.3533147.

[400] Max Weber, Hans Heinrich Gerth, and C. Wright (Charles Wright) Mills. *From Max Weber: Essays in sociology*. New York: Oxford university press, 1946. URL http://archive.org/details/frommaxweberessa00webe.

[401] Benjamin Weiser. Chatgpt lawyers are ordered to consider seeking forgiveness. *The New York Times*, June 2023. URL https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html.

[402] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*, 2021.

[403] W. Bradley Wendel. The promise and limitations of artificial intelligence in the practice of law. *Oklahoma Law Review*, 72:21–50, 2019.

[404] Tom Wheeler. The three challenges of ai regulation, June 2023. URL https://www.brookings.edu/articles/the-three-challenges-of-ai-regulation/.

[405] James Q Whitman. The two western cultures of privacy: Dignity versus liberty. *Yale Law Journal*, 113:1151, 2004.

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

[407] Langdon Winner. Do artifacts have politics? In *Computer ethics*, pages 177–192. Routledge, 2017.

[408] Greg Winter. 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.

[409] Christine T. Wolf, Haiyi Zhu, Julia Bullard, Min Kyung Lee, and Jed R. Brubaker. The changing contours of "participation" in data-driven, algorithmic ecosystems: Challenges, tactics, and an agenda. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '18 Companion, page 377–384, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360180. doi: 10.1145/3272973.3273005. URL https://doi.org/10.1145/3272973.3273005.

[410] Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

[411] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1174–1185, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594072. URL https://doi.org/10.1145/3593013.3594072.

[412] Writers Guild of Am., W., Inc. v. American Brdcst. Cos., 609 F.2d 355 (9th Cir.), 1979.

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

[414] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Literature Review. *arXiv preprint arXiv:2303.13379*, 2023.

[415] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

[416] Karen Yeung. The regulatory state. In Robert Baldwin, Martin Cave, and Martin Lodge, editors, *The Oxford Handbook of Regulation*, pages 64–81. Oxford University Press, September 2010. ISBN 978-0-19-956021-9. doi: 10.1093/oxfordhb/9780199560219.003.0004. URL https://doi.org/10.1093/oxfordhb/9780199560219.003.0004.

[417] Jillian C. York and Rainey Reitman. In south korea, the only thing worse than online censorship is secret online censorship. https://www.eff.org/deeplinks/2011/08/south-korea-only-thing-worse-online-censorship, 2011.

[418] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

[419] Cat Zakrzewski. FTC investigates OpenAI over data leak and ChatGPT's inaccuracy. *Washington Post*, Jul 2023. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/.

[420] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

[421] Amy X Zhang, Grant Hugh, and Michael S Bernstein. Policykit: building governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 365–378, 2020.

[422] Chunyan Zhang, Junchao Wang, Qinglei Zhou, Ting Xu, Ke Tang, Hairen Gui, and Fudong Liu. A survey of automatic source code summarization. *Symmetry*, 14(3), 2022. ISSN 2073-8994. doi: 10.3390/sym14030471. URL https://www.mdpi.com/2073-8994/14/3/471.

[423] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*, 2023.

[424] Ruijie Zhou, Soham Deshmukh, Jeremiah Greer, and Charles Lee. Narle: Natural language models using reinforcement learning with emotion feedback. *arXiv preprint arXiv:2110.02148*, 2021.

[425] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2020.

Appendix A

Supplement Material for Case Study 1

A.1 Workshop Participant Information

Table A.1: Workshop Participant Information

Number	Legal Experience (yrs)	Category	AI Use (General)	AI Use (Work)
P1	> 20	Law faculty	Occasional	Occasional
P2	< 5	Attorney	Occasional	Occasional
P3	> 20	Law faculty	Regular	Occasional
P4	6-10	Attorney	Regular	Regular
P5	11-15	Attorney	Occasional	Never
P6	> 20	Law faculty	Regular	Regular
P7	< 5	Law student	Regular	Never
P8	11-15	Attorney	Regular	Regular
P9	6-10	Law faculty	Occasional	Occasional
P10	< 5	Attorney	Regular	Regular
P11	< 5	Attorney	Regular	Regular
P12	< 5	Researcher	Regular	Regular
P13	6-10	Attorney	Regular	Regular
P14	< 5	Attorney	Regular	Regular
P15	< 5	Law student	Regular	Occasional
P16	6-10	Attorney	Regular	Regular
P17	16-20	Attorney	Occasional	Occasional
P18	< 5	Attorney	Occasional	Never
P19	< 5	Law student	Regular	Occasional
P20	< 5	Attorney	Regular	Regular

Note: Years of legal experience is self-reported with years of legal education removed for consistency.

A.2 Provided AI Response Strategies and Examples

Table A.2: AI Response Strategies and Corresponding Example Responses.

Response Strategies	**Description**	**Example Response (Legal Advice)**
Content violation	Displays a warning message that the user's input violates the AI's content policy.	This content may violate our content policy. If you believe this to be in error, please submit your feedback—your input will aid our research in this area.
Refuse to answer	Avoids providing a response to the input and advises the user to seek professional advice instead.	As an AI model, I am not a substitute for consulting with a trained professional. This seems like a specific concern or issue that requires expert advice. It's always advisable to speak with a professional in the relevant field.
Empathize + Refuse	Emphasizes with a user but avoids providing a response to the input.	I am sorry that this is happening to you. It sounds like a challenging situation. Unfortunately, as an AI model, I cannot provide specific legal advice...
Retrieve (non-opinion) facts	Provides a detailed response that cites specific facts relevant to the input but avoids providing a further legal analysis.	As an AI model I cannot give legal advice specific to your situation but here is a list of laws and cases that may be helpful to look into: 1. Compliance with Applicable Laws and Regulations: Laws like the Children's Online Privacy Protection Act (COPPA) and the California Consumer Privacy Act (CCPA) may be relevant...
Help work through questions	Helps users to identify relevant matters in their situations and, if necessary, responds with a question requesting more information.	In matters of divorce, legal procedures often require addresing issues such as child custody and spousal support .... To help you with this, I will need some additional information. How many children do you have?
Recommend Actions	Analyzes a user's situation under the relevant laws and provides specific further steps that a user may take.	I regret the distressing experience of verbal abuse that you endured at school. In California, potential legal recourses may encompass claims of negligence, a violation of federal civil rights laws (Title XI) ... As a prudent next step, I recommend initiating formal proceedings ...

A.3 Linear Regression of Participants' AI Usage and Desired Responses

Presented in Table A.3, participants' receptivity to a tailored AI response is estimated by the average of the most generous answer types per each prompt. The "content warning" is marked as 0 points, the lowest comfort level, and the "recommend action" template is marked as 6 points. For example, if a participant chose both "empathize + refusal" (2 points) and "Help work through questions" (4 points) for the first case (the higher point is 4) and chose "Recommend actions" (6 points) for the second case, we mark their level of receptivity as 5 points. While P13 worked on four cases, all other participants chose two cases each. The regression results (Table A.4) indicate that general AI fluency significantly predicts higher comfort levels with proactive AI responses ($p < 0.05$), whereas work AI fluency is marginally associated with lower comfort levels ($p = 0.054$). The predictors explain 25.6% of variation. Further investigation is required to substantiate these preliminary relationships with a larger sample.

Table A.3: Participants' AI Use and Their Receptivity to More Tailored Responses

Number	AI Use (General)	AI Use (Work)	Receptivity
P1	1	1	1
P2	1	1	4
P3	2	1	5
P4	2	2	4
P5	1	0	4
P6	2	2	3.5
P7	2	0	5.5
P8	2	2	2.5
P9	1	1	4
P10	2	2	4.5
P11	2	2	4
P12	2	2	1
P13	2	2	3.75
P14	2	2	4.5
P15	2	1	4.5
P16	2	2	6
P17	1	1	4
P18	1	0	4
P19	2	1	6
P20	2	2	4.5

Note: A pre-survey asked participants to describe their AI usage in both professional ("Work") and non-professional ("General") settings, using a scale where 0 represented "Never," 1 "Occasional use," and 2 "Regular use." We then estimated receptivity to more tailored responses such as opinion by averaging the most generous answer types for each case.

Table A.4: Regression Results

Predictor	Estimate	p-value
Intercept	2.4682	0.0297
AI usage in work	-0.9682	0.0543
AI usage daily	1.6773	0.0373

- Residual Std. Error: 1.199 on 17 degrees of freedom
- Multiple R-squared: 0.2557
- Adjusted R-squared: 0.1681
- F-statistic: 2.92 on 2 and 17 DF
- p-value: 0.08127

Appendix B

Interview Protocol for Case Study 2

B.1 Assessment of harmful content

1-1. Have you been concerned about harmful content in your industry? What were they?

1-2. Has the rise of the internet affected the distribution of harmful content?

1-3. Have you noticed any significant change in harmful online content in recent years?

1-4. What kind of efforts have you made to address this situation?

B.2 Relationship between stakeholders

2-1. Who are the main stakeholders in your industry? Could you name the influential entities or individuals?

2-2. How can you identify "creators" in your industry? How many are there?

2-3. Can you describe the relationship between online platforms and creators? Who has editorial control over the content? If there is a dispute, how do they resolve it?

2-4. Have you ever experienced any threats from the government on freedom of expression?

B.3 Assessment of the existing co-regulation

3-1. Are you aware of ongoing co-regulation in your industry? What is your impression?

3-2. What purposes do you think a co-regulatory organization must achieve?

3-3. What are the benefits of co-regulation? Could you give me any specific examples?

3-4. If you think the current organization is failing, can you think of any reasons?

3-5. How do you feel about the funding sources? Do you think the funding scheme affects the independence of co-regulation?

3-6. If there was no possibility of government regulation, do you think co-regulation is still necessary?

B.4 Solutions

4-1. Which stakeholder should be more active in co-regulatory governance? (e.g., creators, platforms, co-regulators, the government, civic groups, academics, labor unions, courts)

4-2. What are the most important organizational factors for effective co-regulation?

4-3. What would be the worst case that significantly undermines the legitimacy of co-regulation?

4-4. Which roles government should play to support effective co-regulation?

B.5 Specific Questions

B.5.1 For creators:

5.1-1. Could you describe the process of content production and distribution?

5.1-2. Who poses the most significant threats to your freedom of expression?

B.5.2 For platform executives:

5.2-1. How do you correspond with creators, other platforms, a co-regulator, and the government?

5.2-2. How many people are working to detect and remove harmful content in your company? Can you describe the internal procedure? What kind of guidelines are you using?

5.2-3. What kind of technology are you using? Do you share it with other platforms or entities?

B.5.3 For co-regulators:

5.3-1. How many people are working in your organization? What are their roles and responsibilities?

5.3-2. What is your annual budget? Who funds it?

5.3-3. What kind of rules and guidelines are you using?

5.3-4. How do you recruit decision-making members? How often do they meet? How many cases do they cover? Do they often reach a consensus?

5.3-5. Could you describe the decision-making process? How do you enforce those decisions?

5.3-6. What were the major achievements of your organization?

B.5.4 For government officials:

5.4-1. What kind of regulatory tools are available to address harmful content? Have you exercised authority?

5.4-2. If you have, how many cases were there last year? Could you give me some examples?

5.4-3. If you haven't, what were the concerns about exercising authority?

5.4-4. How often and in what manner do you correspond with a co-regulator?

Appendix C

Safeguarding Human Values: Rethinking US Law for AI's Societal Impacts

C.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.[1] This includes efforts to discern user intent more accurately [307], refuse unethical commands [188, 82], avoid hallucinated content [288, 194], and generate more coherent and engaging responses [270]. 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

[1]This research is published in the proceedings of the AI and Ethics 2024 under the following citation: Inyoung Cheong, Aylin Caliskan & Tadayoshi Kohno. Safeguarding Human Values: Rethinking US Law for Generative AI's Societal Impacts. AI and Ethics (2024). https://doi.org/10.1007/s43681-024-00451-4.

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" [97]. Indeed, translating abstract shared values into actionable decisions is a fundamental function of legal systems [239]. 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 [373].

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." [284]) and is unlikely to change its path in the near future. This contrasts with the legislative progress in EU [49, 60], Brazil [352], and China [297]. US has and will have voluntary commitments from corporations [64], advisory guidance like the NIST AI Safety Framework [54], internal guidance for government AI use [55], and sector-specific rules such as drug development [169]. Additionally, the recent and more promising AI Executive Order [215] indicates movement towards more proactive federal guidance on AI safety and bias mitigation priorities. However, a broad national regulation on private AI development and deployment remains unlikely to occur in the near future.

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

- Section 2 emphasizes the law's role in translating contested values into AI alignment and governance framework.
- Section 3 illuminates the deficiencies in current liability laws, rooted in a libertarian tradition (described in Table C.1), regarding emerging issues like AI harms. We crafted five scenarios depicting

potential AI damages: (1) educational disparities, (2) LGBTQIA+ bias reinforcement by AI systems, (3) community-optimized AI services that intensify hatred, (4) emotional addiction to AI replica services, and (5) enabling sexual relationships with AI replicas. Our analysis reveals that existing legal frameworks insufficiently address such ethical issues without clear malicious intent or tangible individual harms evident.

- 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.

Scenario	**1**	**2**	**3**	**4**	**5**
Facts	Only rich public schools offer AI-assisted learning, resulting in educational disparity.	LGBTQIA+ individuals physically attacked due to AI-reinforced stereotypes.	AI tool fine-tuned by communities produces derogatory comments against certain individuals.	User's obsession with AI replica of their former partner leads to self-harm of the user.	AI replica service offers secret sexual relationship without the knowledge of the person who was replicated.
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 C.1: Legal Assessment of Different AI-mediated Value Infringement. We assume that Section 230 liability immunity does not extend to 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.

C.2 The Role of Law in AI Alignment Discussions

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 [106]. In response to benign requests, they produce biased content (e.g., sexually objectified images of women [411], biased judgment against LGBTQIA+ people [362]); makes false claims about certain individuals [323] by deviating from their training data (often called *hallucinating* [220]); and helps to spread misinformation that significantly undermines democratic principles [196]. 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.

C.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) [425, 128, 288, 194, 189] and AI feedback (RLAIF) [83, 254], and adversarial testing through domain experts (referred to as *red-teaming* [188, 82]). Table C.2 provides an example of aligned output of GPT-4.

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,

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 can- not 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 C.2: Example Prompt and Completions for Improved Refusals on Disallowed Categories from OpenAI (2023) [73].

concerning aspects like word order, coherence, and vocabulary matching [239, 144, 117, 185, 349]. 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 [189, 266], another more value-oriented group sought to reduce harmful content [420, 188], improve safety [266, 402, 342], mitigate bias [387, 411], handle ethical dilemmas [222, 175], and balance political views [267].

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.

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 [423]. 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" [247] and "Take a deep breath" [415]) 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 [69]. Research confirms that fine-tuning GPT-3.5 Turbo with a few adversarial examples costing pennies compromises its safety [319]. Even well-intended practices like RLHF inadvertently increase risks by making unsafe behaviors more distinguishable [410].

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 [186, 211]. 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 [334, 214]. Some argue the very notion of "alignment" serves as an "empty signifier"—a rhetorical placeholder appealing to our vague ideals without offering meaningful specificity [240]. 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 [318]. 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 [368]. 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.

Uncertain Incentives for AI Alignment

Market incentives do not automatically encourage comprehensive alignment. Thoughout 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) [130, 326, 235]. 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 [206, 418]. 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.

C.2.2 Codifying Values into Law

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 [83, 91].
- **Personalized Alignment**, developing personalized or curated alignment tailored to criteria defined by individual users or specific communities [424, 150, 221].
- **Open-Source Models**, adopting open-source models that can be fine-tuned as needed rather than centralized closed models [341, 388].
- **Linking Technology and Law**, for example, by using universal human rights as a globally salient value framework to ground responsible AI [318].

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 [94]. 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" [351].

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. US people may find French baby naming laws odd, while French find gun ownership in the US bizarre - but the differences often lie in how laws are put into practice, not just espoused values [405]. 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.

C.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.

C.3.1 Methods

Crafting Scenarios through Expert Workshop

we organized a brainstorming workshop [112, 275, 368] with 10 experts in computer security, machine learning, NLP, and law, guided by a threat-envisioning exercise from the field of computer security research [308]. The first and last authors participated as members of this workshop. During the workshop, participants were asked to identify: (1) potential use-cases of AI systems, (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 [335, 375, 217]. See Appendix C.8 for the structure of the workshop.

The analysis of these codes (available on Appendix C.8) 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.9 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.

Figure C.1: Sticky Notes from Experts Outlining Stakeholders of AI-Based Systems

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 fine-tuning that propagates hatred, self-harm due to over-reliance of technology, and virtual sexual abuse.

Legal Analysis

Our legal analysis is rooted in traditional methods of legal research [300, 101, 395]. 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 C.3: Types of Legal Sources, Classified by the Harvard Law Library [101].

C.3.2 Preliminary Question: Applicability of Section 230 to AI

Section 230 of the Communications Decency Act [17] 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 [67, 195].

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 [90, 396]. 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" [62]. 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 [263].

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." [17] 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 [38]; Google's automated summary of court cases containing false accusations of child indecency [43]; and Google's automated search query suggestions that falsely describe a tech activist as a cyber-attacker [268]. 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 [62].

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 [420], including even unintended results [411, 220]. 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 [264]. 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 [102]. 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 [90]. The factors contributing to the emergent capabilities of AI-based systems, which are not evident in smaller pre-trained models, remain inadequately understood [423]. 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.

C.3.3 Evaluating Legal Recourse for Emerging AI Threat 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 C.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 [13]. 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 [157]. *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 C.4: Differences Between Inner-City and Suburban School Districts in San Antonio, Texas, 1968, Reclassified by Drennon (2006) [157].

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 C.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" [408, 406], 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."

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 [31, 378]. 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., [37, 47]). (2) Applications such as SecretEdu are unlikely to be defined as educational facilities or programs under the laws [9]. (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" [354]. (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 [58].

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) [18, 396]. 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.9.5) 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 [253].

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.

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 [396]. 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 [396]. 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 [106]) and their profound influence (the potential to sway human thoughts and impact significant decisions in areas like employment and housing [264]), 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 [151]. 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).

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) [8]. 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 [4], such as facial geometry and voice prints [249]. 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 [27].

Could Alex make a claim for extreme emotional distress? Although an intentional infliction of emotional distress claim is known to be difficult to establish [20], 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 [12] and the antistalking statutes of many states [2, 1] include a specific "fear requirement" that Riley intended to threaten Alex, which is not found in our case. Impersonation laws [15, 7] 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 [5], 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.

C.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.

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).

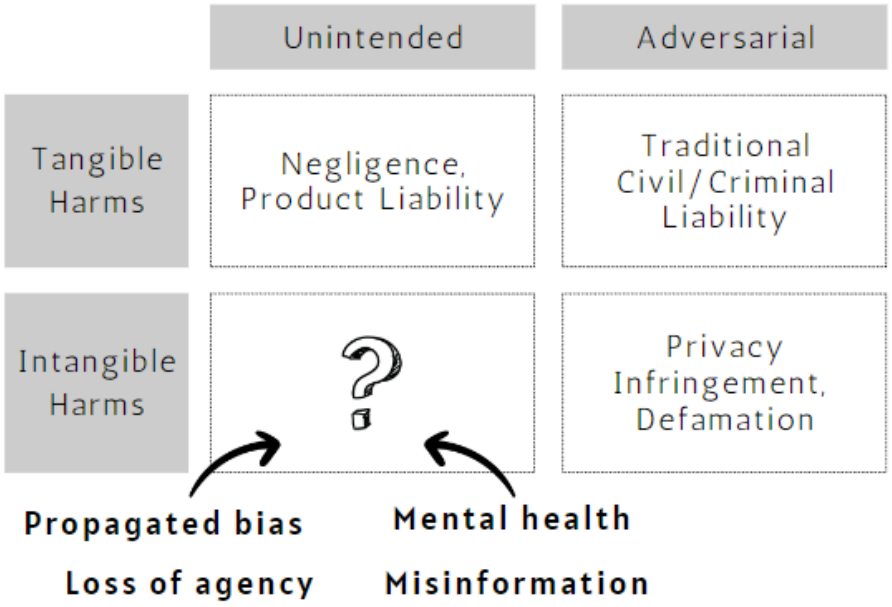


Figure C.2: Legal Mitigations for Propagated AI Bias.

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.

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.

C.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[116, 279, 122, 132, 242], scientists [211, 329] and AI companies [64, 75]. 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" [75].

Indeed, federal agencies have developed sector-specific rules for AI use in domains like drug development [169] and political campaigns [66], while across-sector initiatives including the AI Bills of Rights [56] and NIST's AI Risk Management Framework [54] 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 [64]. while there are widely-discussed bills like the Algorithmic Accountability Act of 2022 [52], 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.

C.4.1 Government: Enemy of Freedom?

The notion of freedom is shaped by "local social anxieties and local ideals," rather than logical reasoning [405]. 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 [39]. As James Madison put it: "The powers delegated by the proposed Constitution to the federal government are few and defined." [274]. 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 [31]. 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 [260]. 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 [327].

C.4.2 Adversarial v. Regulatory 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 [226]. 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,

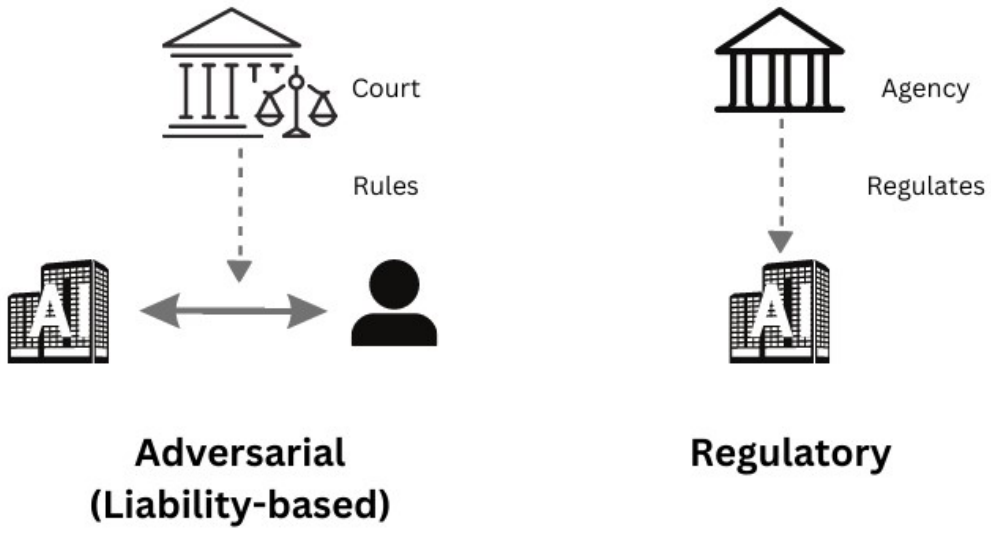


Figure C.3: Comparison Between Adversarial and Regulatory Legal Systems.

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 C.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 [405]. Similarly, Asian cultures, influenced by collectivist ideologies, emphasize community well-being and social cohesion over individual liberty [313, 94]. 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" [286].

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 [42].

(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* [178]) 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 [380]. 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 [11]. 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 [16].

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 [46]. In 2023, the FTC released a Policy Statement on Biometric Information, addressing privacy, security, and potential biases linked to biometric technologies [61], and initiated an investigation into OpenAI, particularly concerning ChatGPT's generation of inaccurate information and its potential reputational harms to consumers [419].

C.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 [79].

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 [29]. 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 [34]. 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 [45].

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 [209, 171]. 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 [138]. Furthermore, the EU's forthcoming Digital Services Act, set to be effective in 2023, includes provisions on swift removal of illegal content online [53]. 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 C.3.2, Section 230 of the CDA [17], 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 [17]. This contrasts with more conditional liability immunity for internet intermediaries in Europe and Asia [125].

C.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 [227, 287]. 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 C.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 C.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 C.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 [228].

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 [42]. 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 [151].

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) [53]. 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.

C.4.5 Fundamental Tensions

Section C.2 demonstrates that law offers time-tested formulas for instilling human values into technological progress through accountable democratic structures. Section C.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 US 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 C.3. Figure C.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.

C.5 Paths Forward

C.5.1 Why Regulations Are Essential in AI Governance

The enduring tension deeply rooted in US jurisprudence poses significant challenges to ongoing efforts in AI regulation. Many people in the US 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.

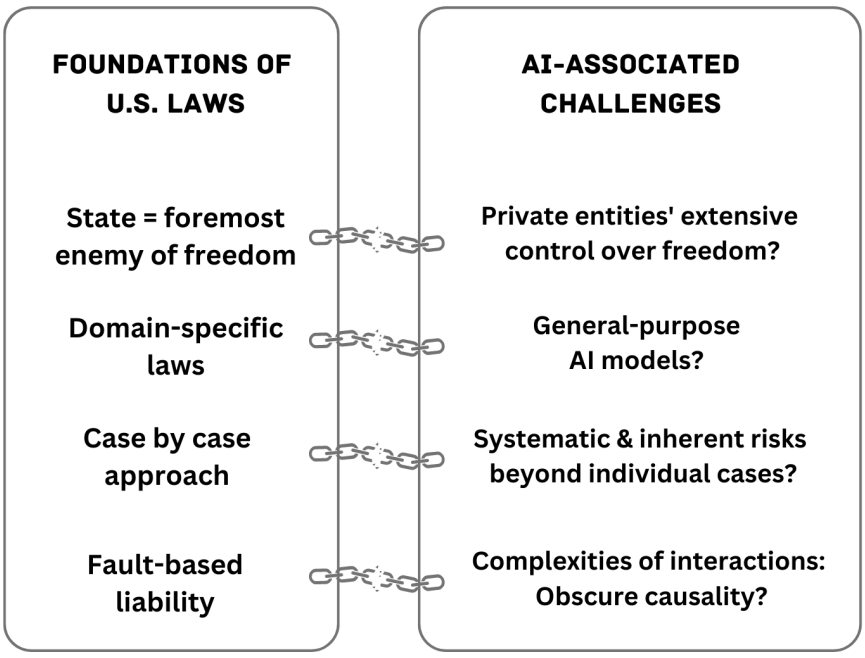


Figure C.4: Tensions between the US law and AI technology.

Democratic Oversight

The ethical foundations of AI should be firmly grounded in shared societal values, not unilateral corporate interests. As discussed in Section C.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 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.

Incentives to AI Safety Alignment

In the absence of a regulatory approach that prioritizes industry efforts to align AI systems with human values, 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.

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.

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 guaran-

tees 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.

Proven Legal Mechanisms

Existing laws, such as bans on deepfakes and regulations concerning biometric data in Section C.3.4, 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.

C.5.2 Towards an Ethical AI Regulatory Framework

This section outlines pragmatic solutions that steer our regulatory system to effectively govern AI by encoding human values into law. We first propose reconstituting rights to directly address emerging threats like manipulative systems and unequal access. Next, we discuss comprehensive safety regulations that incentivize ethical design while emphasizing inclusion. Finally, we explore evolving liability rules to bridge gaps between existing laws and intricate algorithmic harms. As depicted in Figure C.6, this multi-pronged approach accounts for the complex AI ecosystem by employing time-tested legal tools to encode priorities, deter violations, and remedy damages.

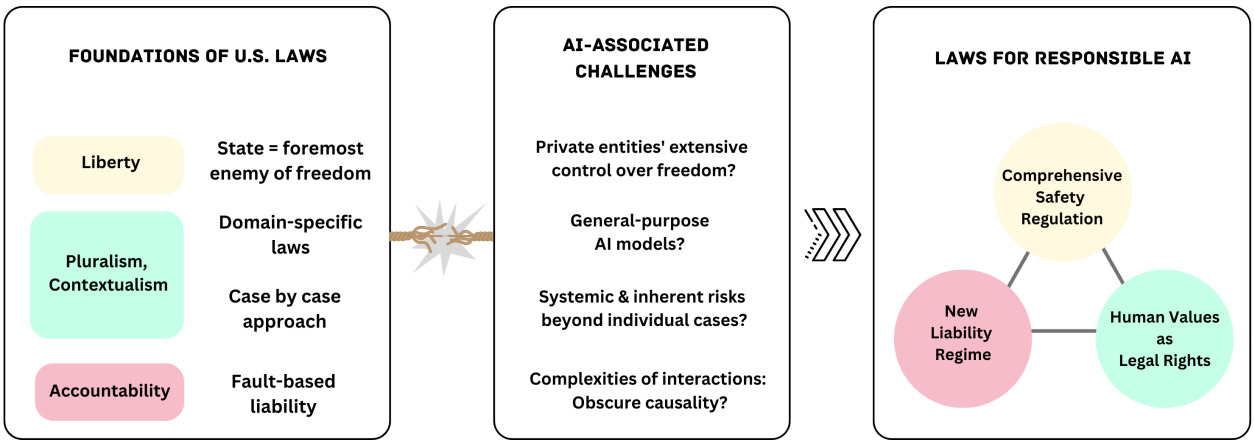


Figure C.5: Ethical AI Regulatory Framework.

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 [140, 350]. Furthermore, since AI applications are proven to reflect harmful stereotypes against marginalized populations (See Section C.9.3), empowering marginalized groups to participate in the development and use of AI will be a more significant demand in the AI-mediated society [159].

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 [56]. Furthermore, as speculated by Franklin Theodore Roosevelt (1944) in his proposed Second Bill of Rights [331], 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 [350].

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 [274]. 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" [400]. To these early thinkers, the government stood as the preeminent 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 [378].

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 [246].

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 [10, 327, 190]. In this legislative

movement, as well as in the 1954 Supreme Court ruling that overturned the "separate but equal" racial segregation theory [19], 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 [356].

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 [330, 115, 133, 139, 134]. 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 [115]. 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 [129].

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 [315]. The law and philosophy scholar Nita Farahany (2023) conceptualizes this effort as the preservation of "cognitive liberty," individual's control over mental experiences [168]. 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.

Comprehensive Safety Regulation

As we have observed in many failed attempts in the field of online privacy self-regulation [191], 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 collab-

oration 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 [178]. In essence, a forward-looking regulatory framework aligned with alignment incentives, equity, and stakeholder input guides AI progress while steadfastly safeguarding human values.

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 [228].

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 [60].

In addition, a "disparate impact" theory developed in relation to the Civil Rights Act of 1964 [10] 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 [190]. This theory diverges from traditional discrimination laws, which have often focused on intent or explicit discriminatory actions [24]. In particular, the recent settlement between the Department of Justice and Meta [296] sets a precedent by attributing responsibility to Meta based on acknowledging the disparate impact caused by targeted advertising algorithms [296]. 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 [73, 423].

C.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.

C.7 Appedix A. Expert Workshop Instruction

The instruction for the workshop is available at:
https://anonymous.4open.science/r/LLM-DDD0/expert_panel_instruction.pdf.

C.8 Expert Workshop Results

A detailed overview of the responses obtained is available at:
https://anonymous.4open.science/r/LLM-DDD0/expert_panel_result.pdf.

C.9 Appendix B. Human Values at Risk in the Era of AI

C.9.1 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 [414, 231, 166, 422]. 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 [159].

An international development scholar Kantrao Toyama contends that technology alone cannot rectify the inequity in educational opportunities [389]. 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 [363]. The COVID-19 pandemic further exacerbated these divides, particularly for low-income students who faced limited access to essential technology and live instruction [212].

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 [385, 369, 363]. Because AI models demand substantial computing resources, incurring significant operational costs [96], 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.9.2 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 [325, 186]. 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 [125, 325].

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 [218]. 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.9.3 Diversity, Inclusion, and Equity

The presence of biases in language models is a significant concern [114, 387, 192, 362, 301] as it can lead to perpetuation and amplification of harmful stereotypes, biases, and discriminatory viewpoints in the generated output [204, 262, 106, 96]. 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" [362].

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 [221].

C.9.4 Privacy and Dignity

Privacy holds a crucial place in defining the boundaries of an individual's "personhood" and is integral to human development [405, 179]. However, Generative AI models, trained on uncurated web data, may inadvertently perpetuate biases and prejudices while also revealing private information [106, 121]. 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 [323].

Beyond inadvertent disclosure of private data, we must also address more subtle privacy risks, such as the misrepresentation of individuals, including sexual objectification [411]. 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 [398].

Defamation law has traditionally been applied to specific forms of misrepresentation, requiring elements such as falsity, targeted harm, and reputational damage [396]. 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 [106].

C.9.5 Physical and Mental Well-being

Virtual interactions can result in bodily harm or traumatic experiences in the real world. Jurgens et al. [225] depicts the frequency and possibility of physical danger of various virtual harms (Fig. C.6), inspired by prior surveys [158, 337].

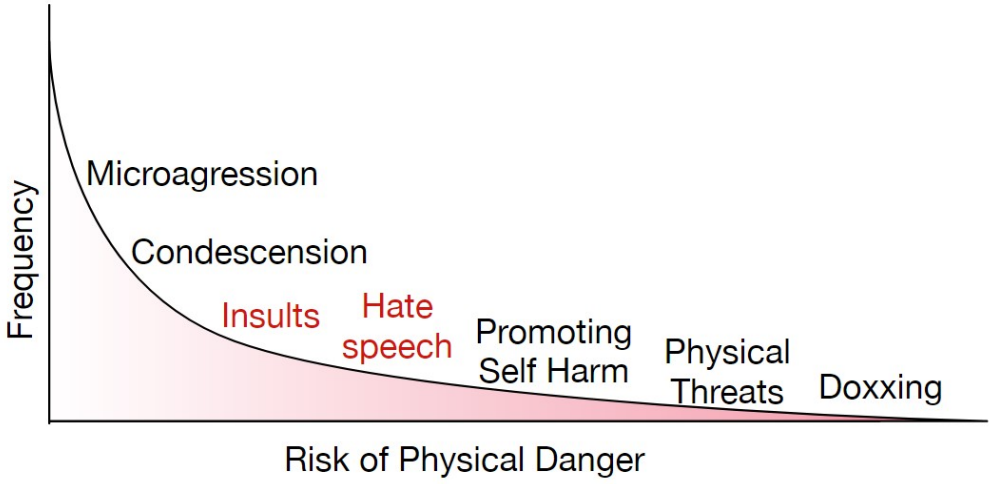


Figure C.6: Frequency and Physical Danger of Abusive Behavior Online [225].

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 [50]. 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 [177]. However, such increased engagement can blur boundaries between the virtual and physical/real world, causing people to anthropomorphize these AI systems [359, 302].

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 [413]. 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 [359].