Regulation of AI-generated disinformation by online platforms: A comparative analysis perspective

XU, XIAONAN

ABSTRACT

The rise of Generative artificial intelligence has transformed the generation and dissemination of disinformation, accelerating its spread and expanding its reach. This thesis will examine how the European Union, the United States, and China are addressing the regulation of AI-generated disinformation and placing primary regulatory liability on online platforms. Using comparative study, case study, and doctrinal study, this thesis will analyze the historical evolution of platform liability regimes and current regulatory measures across jurisdictions. The findings reveal that the EU's regulatory approach emphasizes transparency and requirements under the Digital Services Act and content moderation obligations imposed on VLOPs, while the US's regulatory framework prioritizes free speech and provides intermediaries with immunity under Section 230. China, on the other hand, adopts a state-led regulatory approach that encourages online platforms to proactively conduct content moderation. The significance of this research lies in analyzing how inadequacies or ambiguities in various jurisdictions' laws and regulations lead to enforcement difficulties and how malicious disinformation producers exploit these inadequacies to circumvent regulation. Furthermore, it offers feasible recommendations for establishing a cross-jurisdictional collaborative framework for addressing AI-generated disinformation.

Regulation of AI-generated disinformation by online platforms: A comparative analysis perspective

Xiaonan Xu

Master of Jurisprudence (MJur) Thesis

Durham Law School

University of Durham

2025

Statement of Copyright

Regulation of AI-generated disinformation by online platforms: A comparative analysis perspective

1. Introduction

This thesis explores how the EU, the US, and China regulate and hold online platforms liable for AI-generated disinformation. It examines how different legal traditions, policy priorities, and the balance of protected interests shape platforms' content moderation obligations; how liabilities are allocated and triggering conditions are set across selected jurisdictions; why enforcement remains challenging in practice; and what forms of cross-jurisdictional cooperation are realistically feasible in the absence of harmonizing standards for substantive content moderation.

1.1 Background

Artificial Intelligence, as one of the most prominent topics of the moment, is being widely utilized in several socially important areas, including politics, economics, education, and healthcare[1]. In 1956, Marvin Minsky and John McCarthy first proposed the concept of artificial intelligence at an eight-week Dartmouth conference[2]. Over the past decade, there has been a tremendous increase in the use of AI tools, especially Generative AI, which relies on data-driven machine learning models that learn patterns from large datasets to generate new content[3].

However, due to the self-learning and content generation capabilities of generative AI models, they can produce increasingly convincing disinformation, as well as deepfake text, images, and videos that are indistinguishable from authentic works[4]. To address

---

[1] Georgios Tsertekidis and Periklis Polyzoidis, 'Leveraging Artificial Intelligence in the Field of Social Policy against Social Inequalities: The Current Landscape' (2024) 3 Journal of Politics and Ethics in New Technologies and AI <https://ejournals.epublishing.ekt.gr/index.php/jpentai/article/view/38831> accessed 19 September 2025.

[2] John McCarthy and others, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955' (1955) 27 AI Magazine 12 <https://www.aaai.org/ojs/index.php/aimagazine/article/view/1904>.

[3] Yihan Cao and others, 'A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT' (2023) 37 arXiv (Cornell University) 111:1, 111:3.

[4] Markus Anderljung and Julian Hazell, 'Protecting Society from AI Misuse: When Are Restrictions on Capabilities Warranted? | GovAI' (Governance.ai2023) <https://www.governance.ai/research-paper/protecting-society-from-ai-misuse-when-are-restrictions-on-capabilities-warranted> accessed 22 November 2024.

the global spread of AI-generated disinformation, a comparative analysis of the laws, regulations, and local censorship policies across different jurisdictions is necessary to examine the extent to which online platforms are required to review and regulate disinformation posted on their platforms.

The development of generative AI models has different phases. One of the key early examples is the Eliza chatbot, which was created by Joseph Weizenbaum[5]. As a natural language processing system, the Eliza chatbot's design encourages users to engage in more conversations, reflecting humans' intentions and shaping the context of communication[6]. However, the Eliza chatbot lacks true comprehension capabilities, and this early model can only generate simple data without sufficient context and vocabulary[7]. In 2014, Generative Adversarial Networks (GANs)[8] were first introduced as a novel approach to match the real data distribution. This new deep learning model contains a generator to fabricate synthetic data[9] and a discriminator to determine whether the input is from the real data space[10]. The adversarial training framework enables GANs to produce more realistic and higher-quality outputs across diverse domains such as image synthesis, text generation, and audio processing[11]. LLMs (Large Language Models) can create medium-to-high-quality disinformation with minimal human involvement by learning operational patterns and structures from large amounts of training data[12].

---

[5] Joseph Weizenbaum, 'ELIZA - a Computer Program for the Study of Natural Language Communication between Man and Machine' (1966) 9 Communications of the ACM 36.

[6] Simone Natale, 'If Software Is Narrative: Joseph Weizenbaum, Artificial Intelligence and the Biographies of ELIZA' (2018) 21 New Media & Society 712 <https://doi.org/10.1177%2F1461444818804980> accessed 29 June 2020.

[7] London Intercultural Academy, 'The Story of ELIZA: The AI That Fooled the World' (London Intercultural Academy2024) <https://liacademy.co.uk/the-story-of-eliza-the-ai-that-fooled-the-world/>.

[8] Ian Goodfellow and others, 'Generative Adversarial Networks' (2020) 63 Communications of the ACM 139.

[9] Ian Goodfellow and others (n 8) 47.

[10] Lauren Leffer, 'Your Personal Information Is Probably Being Used to Train Generative AI Models' (Scientific American19 October 2023) <https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/> accessed 10 November 2024.

[11] Bharat Dhiman and Pawan Singh, 'Exploding AI-Generated Deepfakes and Misinformation: A Threat to Global Concern in the 21st Century' (SSRN7 December 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4651093> accessed 23 November 2024.

[12] Jiawei Zhou and others, 'Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions' [2023] Proceedings of the 2023 CHI Conference on

Artificial Intelligence Generated Content (AIGC)[13] is created by interpreting human instructions (prompts) to understand their intents and then generating content based on its own knowledge and human intents[14]. The rapid development of generative AI models has driven the emergence of AI applications, making the influence of AIGC go beyond the field of computer science[15] and triggering widespread attention to generative AI products launched by large companies[16]. For example, ChatGPT is a large language model[17] developed by OpenAI that uses massive amounts of training data to generate coherent, contextually relevant, and human-like responses based on human instructions[18].

However, the rapid development of generative AI models also poses significant challenges[19] to the authenticity and reliability of digital information, as they can be manipulated to fabricate or amplify information to mislead audiences[20]. When malicious users exploit generative AI technologies to intentionally fabricate disinformation[21] to deceive audiences or manipulate public perception[22], their output constitutes AI-generated disinformation, which is the central research subject of this thesis. Disinformation generally refers to the deliberate fabrication and dissemination

---

Human Factors in Computing Systems.

[13] Tom B Brown and others, 'Language Models Are Few-Shot Learners' (2020) 4 arxiv.org 1 <https://arxiv.org/abs/2005.14165>.

[14] Liangjing Shao and others, 'Artificial Intelligence Generated Content (AIGC) in Medicine: A Narrative Review' (2024) 21 Mathematical biosciences and engineering 1672.

[15] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchán, 'A Survey of Generative AI Applications' (arXiv.org14 June 2023) <https://arxiv.org/abs/2306.02781> accessed 14 April 2025.

[16] McKinsey & Company, 'The State of AI in 2023: Generative AI's Breakout Year' (McKinsey & Company1 August 2023) <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year> accessed 12 September 2025.

[17] Katikapalli Subramanyam Kalyan, 'A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4' (arXiv.org2023) <https://arxiv.org/abs/2310.12321> accessed 25 October 2025.

[18] Partha Pratim Ray, 'ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope' (2023) 3 Internet of Things and Cyber-Physical Systems 121 <https://www.sciencedirect.com/science/article/pii/S266734522300024X>.

[19] Michela Del Vicario and others, 'The Spreading of Misinformation Online' (2016) 113 Proceedings of the National Academy of Sciences 554 <https://www.pnas.org/doi/full/10.1073/pnas.1517441113>.

[20] Jeff JH Kim and others, 'Generative AI Can Effectively Manipulate Data' (2024) 5 AI and Ethics.

[21] Yisroel Mirsky and Wenke Lee, 'The Creation and Detection of Deepfakes: A Survey' (2021) 54 ACM Computing Surveys 1.

[22] Seyeon Park and Xiaoli Nan, 'Generative AI and Misinformation: A Scoping Review of the Role of Generative AI in the Generation, Detection, Mitigation, and Impact of Misinformation' [2025] AI & SOCIETY.

of false or misleading information with the intent to deceive or cause harm[23]. Such disinformation has the potential power to manipulate public opinion, ignite social unrest, and even incite acts of violence[24]. Malicious actors can use advanced AI tools to generate persuasive disinformation in various forms, such as texts, images, audios, or videos[25]. Generative AI systems could understand users' intent from their instructions and create inaccurate or false information that aligns with that intent, providing it to the user[26]. While the fabrication of disinformation is not a new phenomenon, using generative AI models to create highly realistic false or inaccurate content[27] has significantly expanded the scale of disinformation and enhanced its credibility[28]. Therefore, the review and governance of AI-generated disinformation is an important research topic.

To effectively regulate disinformation posted on platforms, the lawmakers[29] and researchers have begun exploring solutions to identify and mitigate disinformation[30], particularly AI-generated disinformation that can be easily produced and disseminated. In this thesis, I analyze the relevant laws and regulations in the EU, the US, and China, which provide three representative governance approaches.

The EU is primarily addressing the challenges of algorithms and disinformation by

---

[23] Noémie Krack, Lidia Dutkiewicz and Jean De Meyere, 'Generative Artificial Intelligence and Disinformation' (SSRN2025) <https://ssrn.com/abstract=5192993> accessed 12 September 2025.

[24] Stephan Lewandowsky, Ullrich KH Ecker and John Cook, 'Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era' (2017) 6 Journal of Applied Research in Memory and Cognition 353 <https://www.sciencedirect.com/science/article/abs/pii/S2211368117300700>.

[25] Bradley Honigberg, 'The Existential Threat of AI-Enhanced Disinformation Operations' (Just Security8 July 2022) <https://www.justsecurity.org/82246/the-existential-threat-of-ai-enhanced-disinformation-operations/> accessed 10 July 2025.

[26] Erik Derner and Kristina Batistič, 'Beyond the Safeguards: Exploring the Security Risks of ChatGPT' (arXiv.org13 May 2023) <https://arxiv.org/abs/2305.08005> accessed 14 April 2025.

[27] Mohamed Shoaib and others, 'Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models' (2023) <https://arxiv.org/pdf/2311.17394> accessed 2 January 2025.

[28] Xun Jin and others, 'Assessing the Perceived Credibility of Deepfakes: The Impact of System-Generated Cues and Video Characteristics' (2023) 27 New Media & Society.

[29] Associated Press, 'New Bipartisan Bill Would Require Online Identification, Labeling of AI-Generated Videos and Audio' (US News & World Report2024) <https://www.usnews.com/news/us/articles/2024-03-21/new-bipartisan-bill-would-require-online-identification-labeling-of-ai-generated-videos-and-audio> accessed 10 July 2025.

[30] Poorya Zare Janakbari Janakbari, 'Detection and Mitigation of Deepfake Attacks in Cybersecurity : Leveraging Computer Vision and Deep Learning' (Theseus.fi2025) <https://www.theseus.fi/handle/10024/894608> accessed 10 July 2025.

empowering data subjects [31] and imposing regulatory obligations on the online platforms, distributors, and other actors in the AI value chain[32]. The General Data Protection Regulation (hereafter 'GDPR') [33] strengthens individual control over personal data and regulates automated decision-making for individuals[34]; the Digital Services Act (Regulation (EU) 2022/2065, hereafter 'DSA') [35] updates the E-Commerce Directive[36] by imposing different obligations on the online intermediaries, including transparency requirements, notice-and-action mechanism, and enhanced accountability measures for Very Large Online Platforms (VLOPs). The Artificial Intelligence Act[37], a comprehensive regulatory framework for AI systems, introduces a risk-based framework that categorizes AI systems into different risk levels[38], assigns responsibilities to providers and users, and requires national authorities to ensure compliance.

The United States usually grants broad immunity for third-party-generated content posted on platforms. Section 230 of the Communications Decency Act(hereafter

---

[31] Regulation (EU) 2016/679 (General Data Protection Regulation) [2016] OJ L119/1, art 12.

[32] Yulu Pi, 'Missing Value Chain in Generative AI Governance China as an Example' (arXiv.org2024) <https://arxiv.org/abs/2401.02799> accessed 26 September 2025.

[33] 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] OJ L119/1 ('GDPR').

[34] Amit Kumar Kumar, 'Situating Automated Decision-Making Jurisprudence within Data Protection Frameworks: A Study of Intersections between GDPR and EU Artificial Intelligence Act- Part II' (Law School Policy Review16 May 2024) <https://lawschoolpolicyreview.com/2024/05/16/situating-automated-decision-making-jurisprudence-within-data-protection-frameworks-a-study-of-intersections-between-gdpr-and-eu-artificial-intelligence-act-part-ii/> accessed 26 September 2025.

[35] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1 ('DSA').

[36] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (E-Commerce Directive) [2000] OJ L178/1.

[37] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144, and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L, 2024/1689.

[38] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach' (2021) 22 Computer Law Review International 97.

'CDA')[39] provides online platforms with liability immunity, preventing them from being deemed publishers or disseminators of third-party content[40]. However, this protection does not apply to areas such as federal criminal law[41], intellectual property claims[42], and sex trafficking[43], nor does it cover content created or developed by the platforms themselves. As the risks of generative AI applications have become increasingly prominent, several states have begun to enact legislation to regulate AI-generated disinformation. For example, California prohibits the production and dissemination of false audio and video that harms the interests of politicians[44], Virginia amended its Virginia House Bill 2678 to combat revenge pornography[45], and Massachusetts[46] proposed a specific bill to regulate generative AI in terms of definition, operating standards, registration, and enforcement, aiming to protect public safety, privacy, and intellectual property rights.

The third approach, represented by China, requires online platforms, under the supervision of the Cyberspace Administration of China (CAC), to regulate disinformation that seriously harms the legitimate rights and interests of the state, society, and citizens. Article 12 of the Cybersecurity Law[47] explicitly prohibits the fabrication and dissemination of disinformation that disrupts social order or harms the interests of others. Article 7 of the Interim Measures for the Management of Generative Artificial Intelligence Services[48] require data providers to guarantee the authenticity

---

[39] Communications Decency Act of 1996, 47 USC § 230 ('CDA').

[40] Danielle Draper, 'Section 230- Are Online Platforms Publishers, Distributors, or Neither? | Bipartisan Policy Center' (bipartisanpolicy.org13 March 2023) <https://bipartisanpolicy.org/blog/section-230-online-platforms/> accessed 5 December 2024.

[41] Communications Decency Act of 1996, 47 USC § 230(e)(1).

[42] Communications Decency Act of 1996, 47 USC § 230(e)(2).

[43] Allysia Britton, 'The Interplay between Section 230 Immunity and the Allow States and Victims to Fight Online Sex Trafficking Act of 2018 - Weintraub Tobin' (Weintraub Tobin10 November 2022) <https://www.weintraub.com/2022/11/the-interplay-between-section-230-immunity-and-the-allow-states-and-victims-to-fight-online-sex-trafficking-act-of-2018/> accessed 27 September 2025.

[44] California Assembly Bill No 730, 2019–2020 Reg Sess, ch 493 (Cal 2019).

[45] Virginia House Bill 2678, 2019 Reg Sess (Va 2019).

[46] Massachusetts Senate Docket No 1827, 193rd Gen Court, 2023–2024 Reg Sess (Mass 2023).

[47] Cybersecurity Law of the People's Republic of China (《中华人民共和国网络安全法》) (promulgated 7 November 2016, effective 1 June 2017).

[48] Interim Measures for the Management of Generative Artificial Intelligence Services (生成式人工智能服务管理暂行办法) (promulgated 10 July 2023, effective 15 August 2023) art 7.

and accuracy of training data and require online platforms to take measures to prevent the widespread dissemination of disinformation on their platforms. However, for large language models (LLMs) trained on vast amounts of data, ensuring the authenticity of training data proves exceptionally challenging in practice, as it is typically sourced from diverse public websites and exists in enormous quantities[49].

1.2 Aims and Questions of the Research

Given this background, the primary aim of this research is to identify laws and legal regulations that impose regulatory obligations on online platforms to address disinformation across the selected jurisdictions[50], and to propose cross-jurisdictional approaches to promote a combination of mandatory regulation and platform self-regulation[51]. Analysis of generative AI data processing mechanisms reveals the inherent difficulty of mitigating algorithmic bias[52] and the structural opacity[53] caused by the difficulty of disclosing core algorithms. These challenges highlight the need for online platforms to identify, regulate, and remove user-generated disinformation that spreads on their services.

To accomplish the overarching research aim, the research questions that will be addressed in the thesis are as follows.

First, I will explore how national policies[54], legal traditions[55], and the balance of

---

[49] Matt Sheehan, 'China's AI Regulations and How They Get Made' (*Carnegie Endowment for International* Peace2023) <https://carnegieendowment.org/research/2023/07/chinas-ai-regulations-and-how-they-get-made?lang=en> accessed 5 December 2024.

[50] Claire Wardle and Hossein Derakhshan, 'Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making' (Council of Europe2017) <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html> accessed 5 December 2024.

[51] Alexander Peukert, 'The Regulation of Disinformation: A Critical Appraisal' (2024) 16 Journal of Media Law 1.

[52] Tao Huang, 'Content Moderation by LLM: From Accuracy to Legitimacy' (2025) 58 Artificial Intelligence Review.

[53] Sylvia Lu, 'Algorithmic Opacity, Private Accountability, and Corporate Social Disclosure in the Age of Artificial Intelligence' (2020) 23 Vanderbilt Journal of Entertainment & Technology Law 99 <https://scholarship.law.vanderbilt.edu/jetlaw/vol23/iss1/3/>.

[54] Denitza Toptchiyska, 'Legal Aspects of Content Moderation on Social Media Platforms: A Comparative Perspective' (Ssrn.com22 May 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4901501> accessed 4 August 2025.

[55] Bruna Martins and David Morar, 'Online Content Moderation Lessons from Outside the US' (Brookings17 June 2020) <https://www.brookings.edu/articles/online-content-moderation-lessons-from-outside-the-u-s/> accessed 5 December 2024.

protected interests[56] influence online platforms' legal obligations to conduct content moderation in different jurisdictions. The EU's regulatory approach reflects the civil law tradition, establishing statutory law[57] based on the principle of proportionality, emphasizing user protection[58], platform enforcement, and the protection of fundamental rights[59]. The US, rooted in the common law tradition, relies on judicial precedent and prioritizes the protection of free speech[60], thereby limiting state intervention and strengthening online platforms' autonomy in content management[61]. China combines administrative oversight with legislative governance[62]. Building on statutory law, China has issued new departmental regulations in light of the development of AI technologies[63] to improve its regulatory framework of online disinformation.

Second, I will examine the liability attribution rules for online platforms in the EU, US, and China, especially the obligations to identify, review, or remove disinformation, and the thresholds for triggering such a duty of care. By comparing and analyzing the similarities and differences in obligations across different jurisdictions, it becomes clear what kinds of disinformation online platforms are required to supervise and manage. Moreover, in some jurisdictions, such mandatory obligations are not usually imposed

---

[56] Theodore M Benditt, 'Law and the Balance of Interests' (1975) 3 Social Theory and Practice 321 <https://www.jstor.org/stable/23557739>.

[57] Tobias Mast, 'Platform Law as EU Law' (2024) 73 GRUR International 607 <https://doi.org/10.1093/grurint/ikae072> accessed 17 October 2025.

[58] Zsolt Ződi, 'Characteristics of the European Platform Regulation' (2022) 7 Public Governance, Administration and Finances Law Review 91.

[59] João Pedro Quintais, Naomi Appelman and RÓ Fathaigh, 'Using Terms and Conditions to Apply Fundamental Rights to Content Moderation' (2023) 24 German Law Journal 1,2.

[60] Jamal Greene, 'Free Speech on Public Platforms' in Lee C Bollinger and Geoffrey R Stone (eds), Social Media, Freedom of Speech, and the Future of our Democracy (Oxford University Press 2022) 157,158.

[61] Edward Lee, 'Moderating Content Moderation: A Framework for Nonpartisanship in Online Governance' (2020) 70 American University Law Review.

[62] Baiyang Xiao, 'Making the Private Public: Regulating Content Moderation under Chinese Law' (2023) 51 Computer Law & Security Review 105893 <https://www.sciencedirect.com/science/article/pii/S0267364923001036>.

[63] Chang Su, Zhenghao Li and Qiya Qiao, 'Internet Platform Governance: A Comparison of PRC Law and EU Law - KWM' (Kwm.com2022) <https://www.kwm.com/cn/en/insights/latest-thinking/internet-platform-governance-a-comparison-of-prc-law-and-eu-law.html> accessed 20 September 2025.

on all online platforms[64] but are typically borne by larger-scale platforms[65].

Finally, I will provide feasible suggestions for selected jurisdictions to establish a cross-border cooperative regulatory legal framework in the context of widespread cross-border dissemination of AI-generated disinformation. I will analyze whether the enforcement of online platforms' current content moderation obligations in various jurisdictions presents difficulties and examine what practical reasons or legal ambiguities contribute to these difficulties[66]. Then explore how selected jurisdictions can mitigate the widespread dissemination of disinformation across the borders through cooperative regulation.

1.3 Research Methodology

The following methods will be used to conduct the study:

Comparative analysis:

The use of a comparative study aims to reveal the foundational principles, advantages, and limitations inherent in the regulatory frameworks of different jurisdictions. By analysing legal regulations of AI-generated disinformation in the EU, the US, and China, the study aims to understand various regulation priorities[67] influenced by policy preferences and societal values in three jurisdictions. This comparative study will demonstrate how each jurisdiction has adopted different regulatory governance approaches to the widespread dissemination of AI-generated disinformation[68].

The study reviews and compares legal regulations in three jurisdictions regarding the

---

[64] Grant Huscroft, Bradley W Miller and Grégoire Webber, 'Proportionality and the Rule of Law: Rights, Justification, Reasoning Introduction' (papers.ssrn.com8 May 2014) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2434504> accessed 20 September 2025.

[65] Mimi Zou and Lu Zhang, 'Navigating China's Regulatory Approach to Generative Artificial Intelligence and Large Language Models' (2025) 15 European Journal of Risk Regulation 1,10.

[66] Abdullah Ahmed and Mudassar Nisar Khan, 'AI and Content Moderation: Legal and Ethical Approaches to Protecting Free Speech and Privacy' (*ResearchGate*2 September 2024) <https://www.researchgate.net/publication/383661951_AI_and_Content_Moderation_Legal_and_Ethical_Approaches_to_Protecting_Free_Speech_and_Privacy> accessed 5 October 2025.

[67] Jin Sun and Paziliya Yusufu, 'ChatGPT's Risk Overlay and Legal Response to Data Compliance' (2023) 7 Law and Modernization 1,5.

[68] Bharat Dhiman and Pawan Singh, 'Exploding AI-Generated Deepfakes and Misinformation: A Threat to Global Concern in the 21st Century' (SSRN7 December 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4651093> accessed 23 November 2024.

specific liability of online platforms to censor AI-generated disinformation and ensure algorithmic transparency requirements. Furthermore, it investigates how each jurisdiction addresses risk associated with AI-generated disinformation and explores the balance between encouraging innovation in generative AI technologies and mitigating associated risks.

When exploring the feasibility of comparative research, the most fundamental point is that there is sufficient research material in all three jurisdictions to support the comparative study[69]. The selected jurisdictions have established mature legislative frameworks in both online platform regulation and the governance of AI-generated content[70], with rich laws, regulations, and academic resources providing a solid foundation for meaningful comparative analysis[71].

Secondly, each of these three jurisdictions possesses highly developed AI technologies and extensive online platform markets, with the technical capabilities to detect and regulate disinformation. These commonalities provide a foundation for this comparative analysis. Across all three jurisdictions, AI technologies, especially the large language models and other forms of generative AI models, have been widely adopted in digital communication, information dissemination, and the business sector[72]. Among the three jurisdictions, the US has maintained a leading role in generative AI innovation and commercialization, driven by strong private investment[73] and the

---

[69] Shangrui Wang and others, 'Artificial Intelligence Policy Frameworks in China, the European Union and the United States: An Analysis Based on Structure Topic Model' (2025) 212 Technological Forecasting and Social Change 123971.

[70] Jon Chun, Christian Schroeder and Katherine Elkins, 'Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US' (arXiv.org2024) <https://arxiv.org/abs/2410.21279> accessed 2 November 2024.

[71] Amir Al-Maamari, 'Between Innovation and Oversight: A Cross-Regional Study of AI Risk Management Frameworks in the EU, U.S., UK, and China' (arXiv.org2025) <https://arxiv.org/abs/2503.05773> accessed 4 June 2025.

[72] Bikash Saha, Nanda Rani and Sandeep Kumar Shukla, 'Generative AI in Financial Institution: A Global Survey of Opportunities, Threats, and Regulation' (arXiv.org2025) <https://arxiv.org/abs/2504.21574> accessed 4 June 2025.

[73] AI Index Steering Committee (Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terfa Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi and Sukrut Oak), 'The AI Index 2025 Annual Report'(Institute for Human-Centered AI, Stanford University April 2025) <https://hai.stanford.edu/ai-index/2025-ai-index-report> accessed 4 June 2025.

presence of major technology companies[74] such as OpenAI, Google, Anthropic, and Meta[75]. Within the EU, the application of generative AI technologies is becoming increasingly widespread across various sectors[76], which in turn has led to the establishment of comprehensive regulatory frameworks, most notably through the AI Act and other innovation policies[77]. In China, generative AI technology has developed rapidly and has been extensively deployed across the nation's vast online platforms and markets, particularly by giants such as Baidu, Alibaba, ByteDance, and Tencent. Supported by national strategies such as the "New Generation Artificial Intelligence Development Plan" (2017)[78], the large-scale application of generative AI models has transformed the way content is generated and information is exchanged online, driving the development of AI-driven advertising recommendation systems[79] and e-commerce[80]. However, it has also exacerbated the spread of disinformation and the challenges of digital governance[81]. Within selected jurisdictions, the commercial applications of generative AI models and the integration of generative AI governance

---

[74] European Parliament, 'AT a GLANCE: Digital Issues in Focus, European Parliamentary Research Service' (2024) <https://www.europarl.europa.eu/RegData/etudes/ATAG/2024/760392/EPRS_ATA(2024)760392_EN.pdf> accessed 2 November 2024.

[75] Ben Wodecki, 'Generative AI Funding Hits $25.2 Billion in 2023, Report Reveals' (AI Business2023) <https://aibusiness.com/verticals/generative-ai-funding-hits-25-2-billion-in-2023-report-reveals> accessed 5 October 2025.

[76] Pierre-Alexandre Balland and others, 'Generative AI and Foundation Models in the EU: Uptake, Opportunities, Challenges, and a Way Forward ' (2025) <https://cdn.ceps.eu/wp-content/uploads/2025/03/EESC_report_Generative-AI-and-founding-models-in-the-EU.pdf> accessed 5 October 2025.

[77] European Commission, 'Generative AI Set to Transform EU Economy but Requires Further Policy Action' (2025) <https://digital-strategy.ec.europa.eu/en/news/generative-ai-set-transform-eu-economy-requires-further-policy-action> accessed 5 October 2025.

[78] State Council of the People's Republic of China, 'New Generation Artificial Intelligence Development Plan' (20 July 2017) http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm accessed 4 October 2025.

[79] Evelyn Cheng, 'Big Chinese Companies like Alibaba Show That AI-Powered Ads Are Giving Shopping a Boost' (CNBC16 May 2025) <https://www.cnbc.com/2025/05/16/chinese-companies-like-alibaba-see-more-consumption-helped-by-ai-ads.html> accessed 1 October 2025.

[80] Bain & Company, 'Chinese Retailers Invest in Generative AI to Boost Performance' (Bain2024) <https://www.bain.com/about/media-center/press-releases/2024/chinese-retailers-invest-in-generative-ai-to-boost-performance/> accessed 5 October 2025.

[81] Qiheng Chen, 'China's Emerging Approach to Regulating General-Purpose Artificial Intelligence: Balancing Innovation and Control | Asia Society' (asiasociety.org7 February 2024) <https://asiasociety.org/policy-institute/chinas-emerging-approach-regulating-general-purpose-artificial-intelligence-balancing-innovation-and> accessed 5 October 2025.

into legislative regulatory frameworks demonstrate the technological maturity of these three jurisdictions in detecting and monitoring disinformation[82]. This technological maturity enables online platforms in these three jurisdictions to effectively detect, moderate, and remove disinformation, including identifying watermarks in synthetic content [83], identifying specific AI-generated information [84], and assessing the authenticity of large amounts of online information[85].

Thirdly, the feasibility of this comparative study stems from the structural differences in the legal systems of the EU, the US, and China, which respectively represent three distinct yet comparable models for online platform regulations. The EU adheres to the civil law tradition, characterized by its legal system, which is based on comprehensive statutory provisions and a systematic regulatory framework [86]. Through laws and regulations such as the GDPR and the DSA, the EU has established uniform rules enforced by member state authorities [87]. The platform liability in the US is not determined solely by a single statutory provision[88], such as Section 230 of the CDA. Instead, based on the common law system, it is determined through court precedent, which interprets, balances, and develops the relationship between this provision and the Constitution (particularly the First Amendment's free speech clause), thereby continuously adjusting the boundaries of platform liabilities in content moderation[89].

---

[82] Vera Schmitt and others, 'The Role of Explainability in Collaborative Human-AI Disinformation Detection' (ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)3 June 2024) <https://dl.acm.org/doi/10.1145/3630106.3659031> accessed 5 October 2025.

[83] Xiang Li and others, 'A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules' (2025) 53 The Annals of Statistics.

[84] Fernando Martin-Rodriguez, Rocio Garcia-Mojon and Monica Fernandez-Barciela, 'Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks' (2023) 23 Sensors 9037 <https://www.mdpi.com/1424-8220/23/22/9037>.

[85] Aidan Boyd and others, 'The Value of AI Guidance in Human Examination of Synthetically-Generated Faces' (arXiv.org2022) <https://arxiv.org/abs/2208.10544> accessed 14 April 2025.

[86] Reinhard Zimmermann, 'The Civil Law in European Codes', Regional Private Laws and Codification in Europe (Cambridge University Press 2003) <https://www.cambridge.org/core/books/abs/regional-private-laws-and-codification-in-europe/civil-law-in-european-codes/5D6AAF67CB7A86C1380FB853DEF9803B> accessed 7 October 2025.

[87] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 202/1, art 70.

[88] Gregory M Dickinson, 'An Interpretive Framework for Narrower Immunity under Section 230 of the Communications Decency Act' (arXiv.org2023) <https://arxiv.org/abs/2306.04461> accessed 14 April 2025.

[89] Joseph P Fishman, 'Section 230 as First Amendment Rule' (2018) 131 Harvard Law Review

China's legal system, influenced by the civil law tradition, is based on codified legislation, which is refined and supplemented through judicial explanations, and combined with administrative supervision to provide regulations for online platforms to regulate AI-generated disinformation[90]. The structural differences in the legal systems of the three jurisdictions make comparative analysis feasible and meaningful, as it allows for the study of how different legal traditions respond to the same challenge of AI-generated disinformation through different regulatory mechanisms.

Doctrinal studies:

The rationale for choosing doctrinal studies is to conduct a systematic analysis of legal regulations and doctrines within three specific jurisdictions. The study allows a structured examination and comparison of the laws, regulations, and policies related to AI-generated disinformation in the EU, the US, and China. This analysis[91] helps to clarify the legal definition and how existing laws are applied to disinformation and provides a solid foundation for recommending the adaptation of existing laws. Besides, even though these three jurisdictions have different legal frameworks, they all have well-structured and clear legal systems, as well as extensive and sufficient legal resources, which facilitate systematic theoretical analysis.

A doctoral study is appropriate because the topic raises complex questions at the intersection of data protection, freedom of expression, and platform liability that require sustained analytical engagement and theoretical development. Through doctrinal and comparative legal analysis, the research seeks to clarify the underlying principles and identify best practices in regulating online disinformation.

Case studies:

The rationale for using case studies is to provide a detailed understanding of how legal regulations are applied in specific cases, as well as how different jurisdictions tackle

---

<https://harvardlawreview.org/print/vol-131/section-230-as-first-amendment-rule/>.

[90] Bing Chen and Jiaying Chen, 'China's Legal Practices Concerning Challenges of Artificial General Intelligence' (2024) 13 Laws 60 <https://www.mdpi.com/2075-471X/13/5/60>.

[91] Terry Hutchinson and Nigel Duncan, 'Defining and Describing What We Do: Doctrinal Legal Research' (2012) 17 Deakin Law Review 83.

similar challenges. Besides, case studies could not only illustrate the effectiveness of existing laws and regulations in addressing AI-generated disinformation but also analyze the practical challenges during the enforcement of laws.

To analyse the regulations of AI-generated disinformation in the EU, the US, and China, a well-rounded selection of cases is necessary to capture the diversity of regulatory approaches. I will analyze statutory regulatory implementation cases, clarifying the scope of platform liabilities and the scope of regulatory subjects under these laws and regulations by analyzing how legal instruments are interpreted in practice through specific cases. Furthermore, by studying judicial rulings where online platforms bear liability or are exempted for hosting disinformation, I shall illustrate how legal principles and liability thresholds operate in practice[92].

1.4 Significance of Research

The significance of this study lies in examining how different legal systems address the regulatory challenges posed by the widespread dissemination of AI-generated disinformation. As generative AI technologies increasingly create highly realistic disinformation, they threaten the data protection and public trust in digital information. However, existing legal academic research remains somewhat fragmented to a certain extent, with some studies focusing primarily on single jurisdictions or technical aspects, while fewer systematically examining the response mechanisms of different legal systems from a comparative research perspective. This study fills this gap by comprehensively comparing the regulatory systems of the European Union, the United States, and China. By analyzing how their legal traditions and policy priorities influence the governance of online platforms, as well as the practical difficulties online platforms face in governing AI-generated disinformation, it demonstrates how the shortcomings of current laws and regulations create practical challenges. In addition, this study can also provide recommendations for cross-jurisdictional cooperation in the governance

[92] Antonio Cordella and Francesco Gualdi, 'Regulating Generative AI: The Limits of Technology-Neutral Regulatory Frameworks. Insights from Italy's Intervention on ChatGPT' (2024) 41 Government Information Quarterly 101982.

of disinformation, demonstrating the feasibility of cross-jurisdictional cooperation by analyzing the commonalities between the three jurisdictions in disinformation regulation.

1.5 Synopsis

This thesis consists of five chapters, the first of which is the introduction. In this first chapter, I first explain the research background and explore how the emergence of generative AI has facilitated the generation and dissemination of disinformation, as well as how the three jurisdictions selected for this article have adopted distinctive regulatory approaches to address AI-generated disinformation. Secondly, I outline the three methodologies I will employ, demonstrating their effectiveness in addressing the research questions and the feasibility of these research methods. Finally, I discuss the significance and necessity of this research, particularly how a comprehensive comparative study can lay the legal foundation for future cross-jurisdictional regulatory cooperation.

In the second chapter, I will review the technical development history of generative AI, demonstrating how generative AI models create various forms of indistinguishable, human-like information. Furthermore, by examining the internal operating mechanisms and characteristics of generative AI, I will explain why AI-generated disinformation can be mass-produced and widely disseminated by malicious actors. At the same time, based on the algorithmic logic of generative AI, the information it generates inevitably contains disinformation originally present in the training data. Amplified by algorithmic bias, this information can ultimately be harmful to the interests of others or social order. Finally, this chapter will explain the primary reasons why AI-generated disinformation is difficult to track and regulate. The non-disclosure of core algorithmic technologies and the ambiguity surrounding the source of its training data make it difficult to trace and combat the sources of disinformation, and thus, cannot identify the direct creators of disinformation.

In the third chapter, I will trace the evolution of platform liability laws in the European Union, the United States, and China. By exploring their historical roots, I will demonstrate the scope of online platforms' liability for information posted on their platforms before the advent of AI-generated disinformation. Furthermore, I will analyze the current platform liability regulations for disinformation management in these three jurisdictions, demonstrating how their legal history has influenced current laws and regulations, the patterns of legal evolution in each jurisdiction, and the potential impact of these patterns on future legal regulations. Finally, I will explore the attribution principles of platform liability adopted by each of the three jurisdictions, the factors determining these principles, and how different attribution principles affect the threshold for triggering a platform's duty of care.

In Chapter 4, drawing on the comparative analysis of the current legal frameworks governing content platform liability in Chapter 3, I will explore the practical challenges that online platforms face in enforcing content regulation. Differences in regulatory models and priorities across the three jurisdictions lead to different impacts on online platforms through their laws and regulations. I will explain how the ambiguity surrounding platform regulatory liability and the discretion granted to platforms by laws in some jurisdictions significantly impact the formulation of online platforms' content moderation policies and the determination of their policy preferences.

In the final chapter, building on the practical challenges discussed in the previous chapter, I will discuss the necessity and feasibility of a collaborative governance framework across jurisdictions to address widespread cross-border disinformation. Based on an analysis of existing legal provisions in the three jurisdictions, I will explore the potential and feasibility of future cross-jurisdictional cooperative governance concerning two aspects of procedural fairness safeguards.

2. The Rise of Generative AI: Evolution, Disinformation, and Challenges

Overview

This chapter will begin with a review of the technical development of AI, focusing on how generative AI models could create persuasive and human-like disinformation. It will then analyze how algorithmic biases and discrimination affect the entire operational process of generative AI, including data collection, model training, and content generation, which ultimately leads to the creation of disinformation. Finally, this chapter will discuss the harms that the rapid generation and spread of AI-generated disinformation bring to human society, as well as the challenges faced by online platforms in conducting content moderation in accordance with legal regulations.

2.1 The Technological Evolution of Generative Artificial Intelligence

2.1.1 The Conceptual and Historical Foundations of AI

Artificial Intelligence was born in the mid-20th century. In 1950, Alan Turing's "Turing Test"[93] laid the foundation for machines capable of creating human-like interactions, and his "Turing Machine" laid the computational framework to achieve it. In 1965, Marvin Minsky and John McCarthy first proposed the academic discipline of artificial intelligence at the Dartmouth Conference[94], setting the stage for two decades of rapid advancements in AI technology.

Before the 1980s, symbolic AI became a dominant paradigm in AI techniques[95], and was widely used in Natural Language Processing(NLP)[96] for tackling linguistic

[93] Alan M Turing, 'Computing Machinery and Intelligence', Parsing the Turing Test (Mind 2007) <https://www.csee.umbc.edu/courses/471/papers/turing.pdf> accessed 5 December 2024.
[94] Bruce G Buchanan, 'A (Very) Brief History of Artificial Intelligence' (2005) 26 AI Magazine 53 <https://www.aaai.org/ojs/index.php/aimagazine/article/view/1848>.
[95] Anum Fatima, 'How Has Artificial Intelligence Evolved from Symbolic AI to Deep Learning?' (Machine Mindscape31 January 2024) <https://machinemindscape.com/artificial-intelligence-to-deep-learning-history-concepts/> accessed 21 January 2025.
[96] Benjamin Ayer, 'Symbolic AI vs Machine Learning in Natural Language Processing' (Inbenta4 March 2020) <https://www.inbenta.com/articles/symbolic-ai-vs-machine-learning-in-natural-language-processing/> accessed 5 December 2024.

nuances and achieving high accuracy with limited datasets[97]. Early AI applications relied on symbolic processing to perform logical reasoning and rule-based algorithms, such as the Eliza chatbot and Shakey robot[98]. The Eliza chatbot, created by Joseph Weizenbaum in 1961, represents a precursor of conversational AI[99] to simulate human conversations using rule-based techniques[100] and is widely considered an early version of chatbots. While not a generative system, the Eliza chatbot uses pattern matching to encourage humans to constantly express their feelings and respond to users' inputs through its predefined scripts.[101] Another solid example is the creation of the Shakey[102], which was developed by the Stanford Research Institute in the late 1960s. It could autonomously perceive and analyze the environment, reason its own actions, and execute tasks like navigating itself to another place, which were significant breakthroughs in AI technologies[103]. However, one of the significant limitations of symbolic AI is that it would only perform within its predefined rule sets and not adjust the content through learning from the context or recognizing patterns[104]. This limitation results in symbolic AI requiring complete and accurate knowledge to work and having difficulty understanding users' ambiguous or uncertain statements.

Advances in AI and Machine Learning (ML) [105] have moved computers from traditional rule-based systems to flexible and data-driven systems that can learn and

---

[97] Pamela Weber, 'SmythOS - Symbolic AI in Natural Language Processing: A Comprehensive Guide' (SmythOS15 November 2024) <https://smythos.com/ai-agents/natural-language-processing/symbolic-ai-in-natural-language-processing/> accessed 9 February 2025.

[98] Robert Hoehndorf and Núria Queralt-Rosinach, 'Data Science and Symbolic AI: Synergies, Challenges and Opportunities' (2017) 1 Data Science 27.

[99] Michael Mctear, Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots (Morgan & Claypool Publishers 2020) 20 20–22.

[100] Maali Mnasri, 'Recent Advances in Conversational NLP : Towards the Standardization of Chatbot Building' (2019) <https://arxiv.org/pdf/1903.09025> accessed 14 April 2025.

[101] London Intercultural Academy, 'The Story of ELIZA: The AI That Fooled the World' (London Intercultural Academy2024) <https://liacademy.co.uk/the-story-of-eliza-the-ai-that-fooled-the-world/>.

[102] Nils Nilsson and Donald Nielson, 'SHAKEY the ROBOT' (1984) <https://www.cs.sfu.ca/~vaughan/teaching/415/papers/shakey.pdf> accessed 5 December 2024.

[103] Aleksandra Szczepaniak, 'Leo Rover Blog - What Was the World's First Mobile Intelligent Robot?' (www.leorover.tech2023) <https://www.leorover.tech/post/what-was-the-worlds-first-mobile-intelligent-robot> accessed 5 December 2024.

[104]' SmythOS - Understanding the Limitations of Symbolic AI: Challenges and Future Directions' (SmythOS13 November 2024) <https://smythos.com/artificial-intelligence/symbolic-ai/symbolic-ai-limitations/> accessed 5 December 2024.

[105] Tom M Mitchell, Machine Learning (Mcgraw Hill 2020).

improve from experience[106]. ML focuses on designing algorithms capable of discovering patterns and learning information from large amounts of data to make predictions or decisions without explicit instructions[107]. The algorithm's performance of understanding and content accuracy could be improved by increasing the number of data samples available for learning[108].

2.1.2 Phases in the Technological Development of Generative AI

The development of Generative AI has progressed through different phases of technological advancement. The rise of deep learning drove significant progress in natural language processing, providing the computational foundation for generative AI models. Subsequently, first-generation generative AI models (such as GANs and VAEs) demonstrated the feasibility of AI-generated images, text, and other forms of synthetic data[109]. Later, the Transformer architecture enabled stable and scalable generative modelling through the introduction of self-attention mechanisms[110] and the parallelizable training process[111], thereby facilitating coherent outputs at a large scale[112]. This section will trace these three phases, focusing on how their continuous refinement has propelled the advancement of generative artificial intelligence models.

Deep learning (DL) is a subset of ML that focuses on training multi-layer neural

---

[106] Namirial Focus, 'AI and Machine Learning: How Computers and AI Evolve Together' (Focus Namirial EN27 September 2023) <https://focus.namirial.com/en/ai-machine-learning/#google_vignette> accessed 21 February 2025.

[107] Nantheera Anantrasirichai and David Bull, 'Artificial Intelligence in the Creative Industries: A Review' (2021) 55 Artificial Intelligence Review 589 <https://link.springer.com/article/10.1007/s10462-021-10039-7>.

[108] Sara Brown, 'Machine Learning, Explained' (MIT Sloan21 April 2021) <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>.

[109] Sam Bond-Taylor and others, 'Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models' (2021) 44 IEEE Transactions on Pattern Analysis and Machine Intelligence 1 <https://arxiv.org/abs/2103.04922>.

[110] Zhongzhan Huang and others, 'Understanding Self-Attention Mechanism via Dynamical System Perspective' (arXiv.org2023) <https://arxiv.org/abs/2308.09939> accessed 14 April 2025.

[111] Mekhriddin Rakhimov, Shakhzod Javliev and Rashid Nasimov, 'Parallel Approaches in Deep Learning: Use Parallel Computing', Proceedings of the 7th International Conference on Future Networks and Distributed Systems (ICFNDS 2023) (Association for Computing Machinery (ACM) 2023).

[112] Jacob Devlin and others, 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding' (ArXiv11 October 2018) <https://arxiv.org/abs/1810.04805> accessed 14 April 2025.

networks[113] to automatically learn hierarchical representations of data[114]. It is inspired by the structure and function of the human brain and is centered around stacking artificial neurons into layers and training them to process data[115]. Deep learning can use multiple processing layers that transform representations at a higher and more abstract level to learn complex functions[116] and directly identify the features from raw data, then match the most optimized features to each layer during the transforming process[117]. The emergence of deep learning drove progress in automatic image and speech recognition[118], natural language processing, and other tasks. More importantly, it is a technical foundation of Generative AI, which could not only analyze and classify data but also generate new and realistic content.

CNNs and RNNs are both essential architectures within Deep Learning, and each is suited to specific tasks. The advent of convolutional neural networks (CNNs) achieved breakthroughs in image recognition and extraction[119]. They exploit the convolutional filters of images to detect simple patterns (edges, textures) and complex structures (faces, objects)[120]. The wide use of recurrent neural networks (RNNs) improves sequential data processing, especially in predicting the next character or next word in a sentence[121]. CNNs and RNNs can combine to handle some complex tasks that require both spatial and sequential understanding, such as enhanced glaucoma detection[122].

---

[113] Geoffrey E Hinton, Simon Osindero and Yee-Whye Teh, 'A Fast Learning Algorithm for Deep Belief Nets' (2006) 18 Neural Computation 1527.
[114] Jürgen Schmidhuber, 'Deep Learning in Neural Networks: An Overview' (2015) 61 Neural Networks 85.
[115] Jim Holdsworth and Mark Scapicchio, 'Deep Learning' (Ibm.com17 June 2024) <https://www.ibm.com/think/topics/deep-learning> accessed 18 August 2025.
[116] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, 'Deep Learning' (2015) 521 Nature 436 <https://www.nature.com/articles/nature14539>.
[117] LeCun, Bengio and Hinton (n 116) 439.
[118] Ziyi Wang and others, 'CIEASR: Contextual Image-Enhanced Automatic Speech Recognition for Improved Homophone Discrimination' (2024) 1 Proceedings of the 31st ACM International Conference on Multimedia 915.
[119] Rodrigo Silva, 'Exploring Feature Extraction with CNNs - TDS Archive - Medium' (Medium25 November 2023) <https://medium.com/towards-data-science/exploring-feature-extraction-with-cnns-345125cefc9a> accessed 25 February 2025, para 3.
[120] Ibid.
[121] Robin M Schmidt, 'Recurrent Neural Networks (RNNs): A Gentle Introduction and Overview' [2019] ArXiv (Cornell University).
[122] Soheila Gheisari and others, 'A Combined Convolutional and Recurrent Neural Network for Enhanced Glaucoma Detection' (2021) 11 Scientific Reports.

Their success has improved vision and language processing, as well as laid a solid foundation for more advanced models, such as generative AI and transformer models.

Generative AI refers to deep-learning models that can create high-quality text, images, videos, and other forms of content[123]. These generative models learn the basic patterns and structures of training data and use them to create new content based on the user's instructions. Since the early 1970s, generative AI has been widely empowered and adopted across numerous applications in various areas of interest, such as the field of artistic creations, exemplified by Harold Cohen's AARON program-generating paintings.[124]

Deep learning laid the technical foundation for the development of early generative AI models such as GANs and VAEs, which can create entirely new content by learning from massive amounts of training data. In the late 2000s, the advancements in deep learning, especially the generative adversarial networks (GANs) and variational autoencoders (VAEs), enabled the networks to reduce the human supervision required for the learning process and learn from more complex data. Supervised learning could often achieve, at the end of the training process, higher than human accuracy and, therefore, has been integrated into many products and services[125]. To exceed human performance, existing approaches to supervised learning require millions of training samples to feed the system. In order to reduce the need for human supervision and the number of examples required for learning, many researchers are exploring unsupervised learning[126], often leveraging generative models.

In 2014, Generative Adversarial Networks(GANs) were first proposed to provide more fine-grained control over the data generation process and the ability to output more

---

[123] Kim Martineau, 'What Is Generative AI?' (IBM Research Blog20 April 2023) <https://research.ibm.com/blog/what-is-generative-AI> accessed 5 December 2024.

[124] NWOKDSP Nanayakkara, 'Application of Artificial Intelligence in Marketing Mix: A Conceptual Review' (papers.ssrn.com19 November 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3862936> accessed 4 August 2025.

[125] Avital Oliver and others, 'Realistic Evaluation of Deep Semi-Supervised Learning Algorithms' (2018) 31 ArXiv (Cornell University) 3235.

[126] Iqbal H Sarker, 'Machine Learning: Algorithms, Real-World Applications and Research Directions' (2021) 2 SN Computer Science 1 <https://link.springer.com/article/10.1007/s42979-021-00592-x>.

realistic and high-quality data, especially images[127]. GANs involve a generator to fabricate synthetic data that closely resembles the real one from the distribution samples[128], and a discriminator to distinguish genuine and artificial data by examining samples.[129]

Another class of generative models, variational autoencoders (VAEs)[130], was introduced by Kingma and Welling around the same time as GANs. VAEs provide a structured approach to generative modeling by learning a probabilistic representation of the latent space[131]. They can isolate the important latent variables from training data, using them to reconstruct the inputs and generate new data points[132].

Early generative models such as GANs and VAEs struggled with scalability, coherence, and training stability[133], these limitations that have led to the emergence of the Transformer architecture. The transformer network's introduction in 2017 marked a significant leap in enhancing natural language processing tasks[134]. The traditional recurrent models, such as Recurrent Neural Networks(RNNs), are usually performed along the symbolic positions of the input and output sequences, and such an approach achieves significant improvement in computation efficiency, but the inherent sequential nature of precludes parallelization within trained samples that still exists.[135] To handle

---

[127] Yihan Cao and others, 'A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT' (2023) 37 arXiv (Cornell University).

[128] Ian Goodfellow and others, 'Generative Adversarial Networks' (2020) 63 Communications of the ACM 139.

[129] Ranjith Kumar Gatla and others, 'Advancements in Generative AI: Exploring Fundamentals and Evolution' [2024] 2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC) <https://ieeexplore.ieee.org/document/10594003?denied=> accessed 18 August 2024..

[130] Diederik P Kingma and Max Welling, 'Auto-Encoding Variational Bayes' (arXiv.org20 December 2013) <https://arxiv.org/abs/1312.6114> accessed 14 April 2025.

[131] Kurtis Pykes, 'Variational Autoencoders: How They Work and Why They Matter' (Datacamp.com13 August 2024) <https://www.datacamp.com/tutorial/variational-autoencoders> accessed 5 December 2024.

[132] Dave Bergmann and Cole Stryker, 'What Is a Variational Autoencoder? | IBM' (www.ibm.com12 June 2024) <https://www.ibm.com/think/topics/variational-autoencoder> accessed 5 December 2024.

[133] Yuan-Hao Wei, 'VAEs and GANs: Implicitly Approximating Complex Distributions with Simple Base Distributions and Deep Neural Networks -- Principles, Necessity, and Limitations' (arXiv.org2025) <https://arxiv.org/abs/2503.01898> accessed 28 September 2025.

[134] Ketan Kumar, 'Understanding Transformers: A Deep Dive into NLP's Core Technology' (Analytics Vidhya16 April 2024) <https://www.analyticsvidhya.com/blog/2024/04/understanding-transformers-a-deep-dive-into-nlps-core-technology/> accessed 5 December 2024.

[135] Ashish Vaswani and others, 'Attention Is All You Need' (arXiv12 June 2017)

the limitations of long dependencies and parallelization issues, the Transformer model used multi-head attention, which allows the models to attend to different representation subspaces at different positions, capturing diverse types of relationships[136]. This recent progress in Transformers improved the efficient parallelization and significantly reduced training time compared to previous models[137], as well as upgraded the performance of language translation and reading comprehension. This model became the foundation for subsequent models such as Google's BERT, OpenAI's GPT, and other transformer-based models, which have driven neural networks that not only encode text, images, and videos but, more importantly, have the capability to generate new content.

GPT-4[138] is a Transformer-based model pre-trained on data that is publicly available and is provided by third parties to predict and create novel content based on context. Compared to previous generative models such as GPT-3.5, the GPT-4 substantially improves the ability to understand and predict human instructions to generate better outputs, and significantly reduces hallucination[139] to lower the risks. It demonstrates significant enhancements of reasoning and content generation capabilities, and the "Reinforcement Learning from Human Feedback" model (RLHF) enables the GPT model to show a high degree of human-like intelligence and characteristics during operation.[140] Based on a wider range and larger capacity of data samples as training data, as well as the language model obtained through RLHF technology, GPT-4 is able to have broader knowledge than humans and solid natural language generation ability.

<https://arxiv.org/abs/1706.03762> accessed 14 April 2025.

[136] Vaswani and others (n 135) 6.

[137] Noam Shazeer and others, 'OUTRAGEOUSLY LARGE NEURAL NETWORKS: THE SPARSELY-GATED MIXTURE-OF-EXPERTS LAYER' (2017) <https://arxiv.org/pdf/1701.06538>.

[138] OpenAI, 'GPT-4 Technical Report' (OpenAI 2023) <https://cdn.openai.com/papers/gpt-4.pdf> accessed 14 April 2025.

[139] Daniil Sulimov, 'Prompt-Efficient Fine-Tuning for GPT-like Deep Models to Reduce Hallucination and to Improve Reproducibility in Scientific Text Generation Using Stochastic Optimisation Techniques' (arXiv.org2024) <https://arxiv.org/abs/2411.06445> accessed 31 January 2025.

[140] Xinqiang Song, Mingjie Liu and Jiahe Chen, 'Comprehensive Impact Analysis of GPT-4: High-Quality Economic Development and National Security Prevention' (2023) 38 Journal of Guangdong University of Finance & Economics 100.

Besides, GPT-4 could also fine-tune the initial response without external feedback[141] by training the model on a corpus of labeled data, making it have the human brain-like ability[142] to understand language and produce text with a certain level of originality[143]. For example, when inputting text with spelling and grammatical mismatches or information from different accounts, the intrinsic self-correction[144] could capture the user's original intention through human-computer interaction and autonomously adjust, fix, and correct original instructions. The advent of GPT-4 demonstrates a huge step toward Artificial General Intelligence (AGI), while the inevitable reasoning mistakes and tendency to hallucinate also lead some experts to contend that generative AI has not yet reached the benchmark of general human intelligence[145].

Nowadays, the generation and dissemination of a large amount of disinformation relies on AI technologies as a generative tool. Technically, it is driven by deep learning techniques, which could be automatically trained to learn the structure and patterns from training data samples, making the content creation process more efficient and accessible, as well as producing high-quality content at a faster rate.[146] One of the core advancements in generative AI over previous technologies is to train more complex generative models on larger datasets using larger underlying model architectures with access to a wide range of computational resources. For example, the Large Language Model(LLM) has been widely used by malicious users to create disinformation that appears to be highly credible. It has a deep learning algorithm that could recognize human languages that are based on large sets of human-written information samples to recompose, predict, and generate human-like content.

---

[141] Jie Huang and others, 'LARGE LANGUAGE MODELS CANNOT SELF-CORRECT REASONING YET' (2024) <https://arxiv.org/pdf/2310.01798> accessed 14 April 2025.
[142] Konstantinos I Roumeliotis and Nikolaos D Tselikas, 'ChatGPT and Open-AI Models: A Preliminary Review' (2023) 15 Future Internet 192 <https://www.mdpi.com/1999-5903/15/6/192/htm>.
[143] Ying Cheng Wu and Xukang Wang, 'Balancing Innovation and Regulation in the Age of Generative Artificial Intelligence' (2024) 14 Journal of Information Policy.
[144] Geunwoo Kim, Pierre Baldi and Stephen McAleer, 'Language Models Can Solve Computer Tasks' (arXiv.org16 November 2023) <https://arxiv.org/abs/2303.17491> accessed 14 April 2025.
[145] Daniel Schlagwein and Leslie P Willcocks, '"ChatGPT et Al.": The Ethics of Using (Generative) Artificial Intelligence in Research and Science' (2023) 38 Journal of Information Technology 232.
[146] Jinrui Liu, 'Regulatory Framework for New Risks of Large Generative AI Models' [2024] Administrative Law Review 17.

2.1.3 Multimodal AI and Legal Challenges of Undetectable Disinformation

Multimodal generation plays an essential part in AIGC(Artificial Intelligence Generation Content). Firstly, Multimodal models are capable of understanding and processing more relevant contextualized outputs from data in multiple modalities[147], such as text, image, audio, and graph. With the increased capabilities of GANs and large language models, machine-learning models have turned toward the production of integrated combinations of various modalities, which could enable human-machine teams to produce highly personalized disinformation at scale.[148]

For example, the DALL-E-2 is a text-to-image model that uses deep learning methodologies to generate digital images from natural language text. It combines a CLIP encoder with a diffusion decoder to align text embedding with image features, enabling text-to-image transformation [149] . CLIP (Contrastive Language–Image Pretraining) is a useful approach to having a better understanding of textual descriptions and visual concepts. By using the latent space of CLIP, the images can be semantically modified by redirecting the encoded text vector[150]. It has added an audio-specific layer to process the sound inputs, combining Transformer-based techniques in both audio and text embeddings. Researchers trained the model to maximize the similarities between paired text and audio embeddings while minimizing the similarities between mismatched pairs to make an accurate match through contrastive learning[151]. Secondly, another significant advancement of these diffusion-based models is that these models leverage a guidance technique [152] to increase training sample diversity while maintaining sample fidelity, which makes them learn better representations and

---

[147] Cole Stryker, 'What Is Multimodal AI? | IBM' (Ibm.com15 July 2024) <https://www.ibm.com/think/topics/multimodal-ai> accessed 5 December 2024.

[148] Adrienne Thompson, 'The Existential Threat of AI-Enhanced Disinformation Operations' (Center for Security and Emerging Technology11 July 2022) <https://cset.georgetown.edu/article/the-existential-threat-of-ai-enhanced-disinformation-operations/> accessed 5 January 2025..

[149] Aditya Ramesh and others, 'Hierarchical Text-Conditional Image Generation with CLIP Latents' [2022] arXiv:2204.06125 [cs] <https://arxiv.org/abs/2204.06125> 1, 3.

[150] Ibid 4-5.

[151] Shengqiang Liu and others, 'DSCLAP: Domain-Specific Contrastive Language-Audio Pre-Training' [2024] arXiv (Cornell University).

[152] Prafulla Dhariwal and Alex Nichol, 'Diffusion Models Beat GANs on Image Synthesis' [2021] arXiv:2105.05233 [cs, stat] <https://arxiv.org/abs/2105.05233>.

generate more coherent, contextualized, and reliable outputs.[153]

The relevant laws and regulations require online platforms to detect and remove disinformation (e.g., the EU's Digital Service Act[154] and China's Cybersecurity Law[155]), which rely on automated detection and human moderation. Due to the limited reliability of detection technologies for multimodal content, regulators and policymakers have imposed clear compliance obligations on online platforms. However, this approach creates considerable uncertainty regarding the enforcement and attribution of legal liability. For integrated combinations of multimodal model generation, where text, images, audio, and video could all convey key information, each data mode needs to be integrated in a unified framework when analyzing and detecting disinformation[156]. Current research has mainly focused on unimodal analyses, for example, the Bidirectional Encoder Representations from Transformers (BERT), as a significant deep learning architecture for text classification, could effectively detect the contextual information and capture textual associations[157] in evaluations using the Fakeddit dataset[158], thus achieving a remarkable accuracy. However, the complexity of techniques for detecting disinformation generated by multimodal models requires the integration of different computational techniques into a unified framework, such as the combination of natural language processing for text, computer vision for images and videos, and voice recognition for audio[159]. The lack of authentic and complete multimodal datasets containing different modalities[160] and the fact that most detection

---

153 Ramesh and others (n 149) 5.
154 Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1.
155 Cybersecurity Law of the People's Republic of China (promulgated 7 November 2016, effective 1 June 2017).
156 Shivani Tufchi, Ashima Yadav and Tanveer Ahmed, 'A Comprehensive Survey of Multimodal Fake News Detection Techniques: Advances, Challenges, and Opportunities' (2023) 12 International Journal of Multimedia Information Retrieval 1,21.
157 Isabel Segura-Bedmar and Santiago Alonso-Bartolome, 'Multimodal Fake News Detection' (2022) 13 Information 284, 289.
158 Segura-Bedmar and Alonso-Bartolome (n 157) 12.
159 Junxiao Xue and others, 'Detecting Fake News by Exploring the Consistency of Multimodal Data' (2021) 58 Information Processing & Management 102610, 10.
160 Carmela Comito, Luciano Caroprese and Ester Zumpano, 'Multimodal Fake News Detection on Social Media: A Survey of Deep Learning Techniques' (2023) 13 Social Network Analysis and Mining 1.

systems can only process a single modality, significantly increase the difficulty of identification and supervision.

2.2 Generative AI and Its Production of Disinformation

2.2.1 The Evolution and Application of Deepfake Technology

Deepfake specifically refers to images, videos, and audio recordings that are generated or manipulated by AI, which are usually about real individuals, presenting the image or voice of a particular person in a realistic manner[161]. It is a digital fabrication, generating realistic images of things by synthesizing or completely tampering with them to make them say or do something that never happened[162]. This type of content is highly convincing due to the high degree of similarity of real video or audio of existing characters.

Creating fake content is not new. For example, digital photo manipulation technologies were already developed in the 19th century and soon applied to motion pictures, and Photoshop[163], which is widely used without malicious intent. Regional media are not the only entities capable of creating and disseminating deepfakes; users across various online communities can also leverage advanced AI technologies to propagate such content.[164]

Historically, the manipulation of information has been concentrated in labor-intensive areas[165], and manipulators of these actions are usually national or social groups due to the lack of efficient methods and sufficient resources for spreading disinformation.

---

[161] Cristian Vaccari and Andrew Chadwick, 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News' (2020) 6 Social Media + Society <https://journals.sagepub.com/doi/10.1177/2056305120903408>.

[162] Valdemar Danry, Joanne Leong, Pat Pataranutaporn, Pulkit Tandon, Yimeng Liu, Roy Shilkrot, Parinya Punpongsanon, Tsachy Weissman, Pattie Maes and Misha Sra, 'AI-Generated Characters: Putting Deepfakes to Good Use' (CHI 2022 Extended Abstracts, ACM 2022) <https://doi.org/10.1145/3491101.3503736> accessed 24 January 2025.

[163] Ann Young, 'History of Photo Editing and Photo Manipulation' (FixThePhoto.com2019) <https://fixthephoto.com/blog/retouch-tips/history-of-photo-retouching.html> accessed 20 December 2024.

[164] Julian Sanchez, 'Opinion | the Future of Fake News Is Being Pioneered in Homemade Porn' (NBC News8 February 2018) <https://www.nbcnews.com/think/opinion/thanks-ai-future-fake-news-may-be-easily-faked-video-ncna845726> accessed 26 November 2024.

[165] Robert W Zmud, 'Opportunities for Strategic Information Manipulation through New Information Technology', Organizations and Communication Technology (SAGE Publications 1990).

Whether the notorious political campaign conspiracies, advertising or marketing policies designed to attract consumers, and even wartime disinformation campaigns, information manipulation has been proven to have the ability to have a huge influence on public opinion and social assessment[166].

With the advent and development of sophisticated AI technologies, both the diversity of content and the scope of dissemination of deepfakes have been effectively improved. These advancements provided efficient methods for malicious users to manipulate and disseminate fake information widely and rapidly[167]. For example, digital technology has expanded how deepfakes can be presented[168] so that they are not limited to textual output but have the potential to appear to the public in a multiplicity of ways, such as images, video, audio, etc. A typical use of AI technology in creating deepfakes is adapting the actual video or audio of political leaders to create chaos in public discourse for the purpose of political propaganda[169]. LM-based GenAI[170] has become the main deepfake-generated tool, which has a sophisticated understanding of information content and the ability to generate human-like language. When GANs were introduced, they represented a huge leap that enabled the creation of realistic images that are nearly indistinguishable from actual photographs to the naked eye[171].

It is due to the available nature of GenAI and its ability to create extremely believable images and videos that the immense damage caused by deepfakes is sweeping across all areas of society, as evidenced by the recent deepfake crisis in South Korea. According to local media reports in South Korea, the perpetrators uploaded the real pictures or videos of women to private 'telegram' chat groups[172] and paid the channels

---

166 Daniel Silverman, Karl Kaltenthaler and Munqith Dagher, 'Seeing Is Disbelieving: The Depths and Limits of Factual Misinformation in War' (2021) 65 International Studies Quarterly 798, 799.

167 Neha Sandotra and Bhavna Arora, 'A Comprehensive Evaluation of Feature-Based AI Techniques for Deepfake Detection' (2023) 36 Neural Computing and Applications 3860, 3861.

168 Mika Westerlund, 'The Emergence of Deepfake Technology: A Review' (2019) 9 Technology Innovation Management Review 39 <https://timreview.ca/article/1282>.

169 Stefano Di Sotto and Marco Viviani, 'Health Misinformation Detection in the Social Web: An Overview and a Data Science Approach' (2022) 19 International Journal of Environmental Research and Public Health 2173.

170 Shoaib and others (n 27) 2.

171 Shoaib and others (n 27) 3.

172 Joan E Solsman, 'A Deepfake Bot Is Creating Nudes out of Regular Photos' (CNET2020)

to produce AI face-swap pornography in 'humiliation rooms' involving more than 220,000 participants[173].

Deep Learning models are increasingly being used to support adversarial learning for deepfakes, and they are mainly used for the creation and distribution of fake content[174]. For example, when LLMs are used to generate disinformation, the ease and speed of creating high-volume text[175] could significantly amplify marginalized, misleading information or unobtrusively mix in plausible falsehoods[176] by generating massive amounts of reliable information. The text disinformation is easy to generate in high volume and used in bulk to stitch and perfect a common lie[177]. Among all the forms now available, deepfakes in video are the easiest to be exposed through time-based inconsistencies, such as the mismatch between speech and mouth movements.[178]. A recent deepfake detecting method is known as "soft biometrics," which relies on training an algorithm to spot AI editors of videos by tracking subtle and unique facial movements of each[179]. However, after the disclosure of this detection technique, creators of deepfakes could improve their generation systems by enhancing generative models to simulate natural eye-blinking, thereby circumventing detection systems[180].

<https://www.cnet.com/news/privacy/deepfake-bot-on-telegram-is-violating-women-by-forging-nudes-from-regular-pics/> accessed 15 January 2025.

[173] Georgia Smith and Joseph Brake, 'South Korea Confronts a Deepfake Crisis' (East Asia Forum18 November 2024) <https://eastasiaforum.org/2024/11/19/south-korea-confronts-a-deepfake-crisis/> accessed 13 December 2024.

[174] Yogesh Patel and others, 'Deepfake Generation and Detection: Case Study and Challenges' (2023) 11 IEEE Access 143296, 143297.

[175] Ivan Vykopal and others, 'Disinformation Capabilities of Large Language Models' (arXiv.org23 February 2024) <https://arxiv.org/abs/2311.08838> accessed 22 December 2024.

[176] Freddy Heppell, Mehmet E Bakir and Kalina Bontcheva, 'Lying Blindly: Bypassing ChatGPT's Safeguards to Generate Hard-To-Detect Disinformation Claims' (arXiv.org2024) <https://arxiv.org/abs/2402.08467> accessed 14 February 2025.

[177] Renee DiResta, 'AI-Generated Text Is the Scariest Deepfake of All' (Wired31 July 2020) <https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/> accessed 2 November 2024.

[178] Phil Swatton and Margaux Leblanc, 'What Are Deepfakes and How Can We Detect Them?' (The Alan Turing Institute2023) <https://www.turing.ac.uk/blog/what-are-deepfakes-and-how-can-we-detect-them> accessed 2 November 2024.

[179] James Vincent, 'Deepfake Detection Algorithms Will Never Be Enough' (The Verge27 June 2019) <https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work> accessed 2 November 2024, para 5.

[180] Yuezun Li, Ming-Ching Chang and Siwei Lyu, 'In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking' (2018) <https://arxiv.org/pdf/1806.02877> accessed 2 November 2024.

The new AI model, such as GPT-3, has strong content extraction as well as mimicry learning capabilities to generate "human-like" content to provide false statements for unsuspecting users[181]. Moreover, since people subconsciously know that AI is not truly omniscient when they use it for information retrieval, as long as AI can provide coherent and credible content from multiple perspectives, such results have already met the initial expectations of users[182].

2.2.2 The Definition and Characteristics of AI-generated Disinformation

The essence of disinformation can be summarized into three key characteristics: it appears in the form of "information", has a misleading dissemination effect, and often contains deliberate manipulation intentions by the creator or disseminator. These three together constitute the basic criteria for judging whether a piece of content is disinformation. Firstly, Fallis[183] believes that the first characteristic of disinformation is that it remains a type of information. Specifically, the information could express or describe intangible ideas, beliefs, or knowledge tangibly and can be understood by the public and spread consciously[184]. Therefore, disinformation should be understood as a representation that can carry and transmit knowledge and is used to mislead or deceive recipients in the process of communication and dissemination, thereby functionally disrupting the information ecology. Secondly, Wardle & Derakhshan[185] emphasized that disinformation refers to information that tends to mislead unspecific individuals, organizations, or society. Their creators make untrue representations with the intent of convincing the other party of this statement's credibility[186]. It is important to note that

---

[181] Giovanni Spitale, Nikola Biller-Andorno and Federico Germani, 'AI Model GPT-3 (Dis)Informs Us Better than Humans' (2023) 9 Science Advances 1,5.
[182] Donghee Shin, 'How Do People Judge the Credibility of Algorithmic Sources?' (2021) 37 AI & SOCIETY 81, 82.
[183] Don Fallis, 'What Is Disinformation?' (2015) 63 Library Trends 401, 404.
[184] Michael K Buckland, 'Information as Thing' (1991) 42 Journal of the American Society for Information Science.
[185] Claire Wardle and Hossein Derakhshan, 'Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making' (Council of Europe2017) <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html> accessed 5 December 2024.
[186] James Edwin Mahon, 'The Definition of Lying and Deception' [2008] plato.stanford.edu <https://plato.stanford.edu/entries/lying-definition/?ref=ameasuredapproach.blog> accessed 5 December 2024.

the information does not need to actually cause harm to be considered as disinformation, only if it has the potential to cause public harm[187]. The third important feature of disinformation is that it is intentionally created to be misleading. Unlike misinformation, which might be unintentionally spread, disinformation is created and disseminated with malicious intent. The purpose of producing and spreading disinformation is often tied to specific goals, such as influencing public opinion, damaging reputations, or engaging in fraudulent activities[188].

The challenge posed by AI-generated disinformation is its ability to expand, automate, and personalize the generation of disinformation in a way that has profound consequences[189]. The generative AI is capable of producing convincing fake news articles, social media posts, and propaganda videos that blur the line between true and false[190]. These tools enable malicious actors to curate specific disinformation for dissemination to individuals, specific groups of people, or even entire communities, thereby undermining the public's trust and even inciting acts of violence[191].

The rise of AI-generated disinformation is largely attributed to the ease of access to generative AI technologies and the highly persuasive nature of AI-generated content. Morneo's research shows that the technical development of generative AI tools has allowed the creation of deepfakes to no longer be limited to experts' behaviors. Unprofessional users without special knowledge or experience could also achieve the

---

[187] Ronan Ó Fathaigh, Natali Helberger and Naomi Appelman, 'The Perils of Legally Defining Disinformation' (2021) 10 Internet Policy Review <https://policyreview.info/articles/analysis/perils-legally-defining-disinformation>.

[188] Carme Colomina, Sánchez Margalef and Richard Youngs, 'The Impact of Disinformation on Democratic Processes and Human Rights in the World ' (2021) <https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf> accessed 24 January 2025.

[189] Melissa Fleming, 'How AI Is Boosting Disinformation' (www.linkedin.com2024) <https://www.linkedin.com/pulse/how-ai-boosting-disinformation-melissa-fleming-vbdpe/> accessed 24 January 2025.

[190] Katerina Sedova and others, 'AI and the Future of Disinformation Campaigns' (Center for Security and Emerging TechnologyDecember 2021) <https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/> accessed 24 January 2025.

[191] Stephan Lewandowsky, Ullrich KH Ecker and John Cook, 'Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era' (2017) 6 Journal of Applied Research in Memory and Cognition 353 <https://www.sciencedirect.com/science/article/abs/pii/S2211368117300700>.

same results by using available generative software at a very low cost[192]. Kreps et al. demonstrate that common users could not distinguish between AI-generated disinformation and human-created fake information; most respondents deemed the articles to be credible[193]. Moreover, the results comparing three GPT-2 models show that AI-created disinformation has been considered as credible or more credible than human-written articles[194]. Buchanan[195] et al. highlight the ability of LLMs to produce coherent and contextually relevant content without much human intervention or oversight, and the difficulty of detecting the identity of disinformation's creator by existing tools. Dimitrieska's research[196] shows that generative AI could output customized content toward specific individuals or groups, relying on the training data gathered intentionally from targeted audiences.

To clarify the distinctive expression pattern of AI-generated disinformation, there are two analysis methods that need to be conducted as follows. The first one, the semantics-focused analysis method, is effectively used to identify coordinated multimodal campaigns (involving text, images, and videos) deployed across platforms[197] to detect AI-generated disinformation. To investigate content-related features, the Sparse Additive Generative Model(SAGE)[198] is also conducted to clarify the distinctive expression features between AI-generated disinformation and human-created fake content by comparing the parameters of two documents using a self-adjusting

---

[192] Felipe Romero Moreno, 'Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content' (2024) 38 International Review of Law Computers & Technology 1.

[193] Sarah Kreps, R Miles McCain, and Miles Brundage, 'All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation' (2020) 9 Journal of Experimental Political Science 1.

[194] Kreps, McCain and Brundage (n 193) 6.

[195] Ben Buchanan and others, 'Truth, Lies, and Automation' (Center for Security and Emerging TechnologyMay 2021) <https://cset.georgetown.edu/publication/truth-lies-and-automation/> accessed 5 January 2025.

[196] Savica Dimitrieska, 'Generative Artificial Intelligence and Advertising' (2024) 6 Trends in Economics, Finance and Management Journal 23 <https://tefmj.ibupress.com/uploads/2024/07/ibu_journal_tefmj-3.pdf>.

[197] Michael Yankoski, Walter Scheirer and Tim Weninger, 'Meme Warfare: AI Countermeasures to Disinformation Should Focus on Popular, Not Perfect, Fakes' (2021) 77 Bulletin of the Atomic Scientists 119.

[198] Jacob Eisenstein, Amr Ahmed, and Eric P Xing, 'Sparse Additive Generative Models of Text' [2011] International Conference on Machine Learning 1041.

regularization[199].

Firstly, AI-generated disinformation shows a linguistic difference in communication styles. According to the comparative study conducted by Zhou et al.[200], it demonstrates the ability of the generative AI system to flexibly change the language styles depending on the forms and targeted audiences of the disinformation. For example, when AI-generated disinformation is presented in the form of news, it contains more analytic thinking statements and provides information with higher authenticity than that of human creation. While AI-generated disinformation is presented as internet posts, it shows a strong tendency to be self-centered and uses more emotional tones[201] to express users' feelings. AI-generated posts that integrate massive both positive and negative expressions aim to amplify authors' emotions to attract more readers who share similar feelings[202]. Additionally, the analysis of their linguistic styles reveals that the expression of AI-generated disinformation is more formal and rigorous than that of human-created disinformation; it tends to avoid using internet slang or abbreviations in constructing sentences to diminish the reader's uncertainty about these expressions[203].

Secondly, AI-generated disinformation focuses more on establishing credibility in an article, even in a paragraph. It generates purposeful and structured texts that focus on constructing better reasoning and considering multiple factors, such as causality, cognitive processes[204], content discrepancy, and certitude. The AI system enhances persuasiveness and credibility by providing detailed narratives[205], including using full names and affiliations to describe the people who appear and using graphic expressions

---

[199] Jiawei Zhou and others, 'Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions' (2023) 1 Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.

[200] Zhou and others (n 199) 4.

[201] Zhou and others (n 199) 7.

[202] Brooke Fisher Liu, Logen Bartz and Noreen Duke, 'Communicating Crisis Uncertainty: A Review of the Knowledge Gaps' (2016) 42 Public Relations Review 479.

[203] Liu, Bartz and Duke (n 202) 484.

[204] Teppo Felin and Matthias Holweg, 'Theory Is All You Need: AI, Human Cognition, and Decision Making' [2024] Social Science Research Network <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4737265> accessed 18 May 2024.

[205] Yixuan Zhang and others, 'Shifting Trust: Examining How Trust and Distrust Emerge, Transform, and Collapse in COVID-19 Information Seeking' [2022] CHI Conference on Human Factors in Computing Systems.

to explain the event. For example, during the Russo-Ukrainian war, an important part of the information warfare employed by the Kremlin was the use of a large number of accounts registered using fake AI-generated IDs that parrot the talking points with real-world details to bolster their credibility[206]. Besides, AI-created disinformation tends to present a vivid story accompanied by visual evidence, such as illustrations, and demonstrate opinions derived from different perspectives[207]. AI models tend to acknowledge the unawareness of unimportant details and uncertainty about specific evidence when providing information. This recognition of limitations not only helps to foster the establishment of information credibility but also attracts readers to stay engaged for follow-up reports[208]. These details can maximize the ability to demonstrate logical thinking processes that are highly similar to those of humans, making the readers aware of the transparency and credibility of disinformation, thus enabling AI-generated models to more covertly disguise themselves as humans for information generation and dissemination.

2.2.3 Analysis of the causes of AI-generated disinformation

The development of generative AI technologies has transformed the way information is produced and disseminated and expanded the possibilities of manipulating different types of content[209]. While these technologies offer numerous benefits, they also present significant risks, particularly in creating disinformation. Malicious actors can exploit the inherent characteristics of AI models to generate and distribute misleading or harmful content at an unprecedented scale[210].

Firstly, algorithmic bias is a predominant source of disinformation created by generative AI models. There are three main sources of algorithmic bias: bias and discrimination in the training data, bias from AI developers, and bias in the algorithm

---

[206] Thompson (n 148) para 4.
[207] Catherine Beauvais, 'Fake News: Why Do We Believe It?' (2022) 89 Joint Bone Spine.
[208] Zhou and others (n 199) 3.
[209] Noémi Bontridder and Yves Poullet, 'The Role of Artificial Intelligence in Disinformation' (2021) 3 Data & Policy <https://www.cambridge.org/core/journals/data-and-policy/article/role-of-artificial-intelligence-in-disinformation/7C4BF6CA35184F149143DE968FC4C3B6>.
[210] Bontridder and Poullet (n209) 4.

itself.

Bias and discrimination in the training data are perpetuated at all the stages of data processing and are eventually internalized and reflected in the output of the model[211]. AI bias can lead to discrimination against specific individuals or groups when output information from a machine-learning model[212]. In the stage of inputting the data, pre-training models are trained with massive databases to recognize certain patterns and generate new information[213]. Although generative AI is capable of generating new content based on human instructions, this is not complete proof that AI can fully understand human language and its logical reasoning patterns[214]. In essence, generative AI calculates optimal representations of probability distributions based on user input requirements and creates entirely new answers by sampling and mixing training data to compose surreal outputs[215]. The method that Generative AI models use to understand language expressions is to mimic human-created expressions rather than truly understand. [216] This means that model developers for generative AI must rely on different sources of publicly available data on the Internet and other historical texts as samples, and these sources might present a lack of diversity or inconsistent quality[217]. Therefore, the inaccurate, non-representative, or even false training data[218] could reflect the bias and limitations involved in human-created content from the internet,

---

[211] Lorenzo Belenguer, 'AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry' (2022) 2 AI and Ethics.

[212] Belenguer (n 211) 773.

[213] Mujahid Al Rafi, Yuan Feng and Hyeran Jeon, 'Revealing Secrets from Pre-Trained Models' (arXiv.org2022) <https://arxiv.org/abs/2207.09539> accessed 10 February 2025.

[214] Keisuke Sakaguchi and others, 'WinoGrande: An Adversarial Winograd Schema Challenge at Scale' (2020) 34 Proceedings of the AAAI Conference on Artificial Intelligence 8732.

[215] Priyanka Gupta and others, 'Generative AI: A Systematic Review Using Topic Modelling Techniques' (2024) 8 Data and Information Management 100066 <https://www.sciencedirect.com/science/article/pii/S2543925124000020>.

[216] Anil Ananthaswamy, 'The Physics Principle That Inspired Modern AI Art | Quanta Magazine' (Quanta Magazine5 January 2023) <https://www.quantamagazine.org/the-physics-principle-that-inspired-modern-ai-art-20230105/> accessed 5 December 2024.

[217] Alberto Rizzoli, 'Training Data Quality: Why It Matters in Machine Learning' (www.v7labs.com2022) <https://www.v7labs.com/blog/quality-training-data-for-machine-learning-guide> accessed 5 December 2024.

[218] Xavier Ferrer and others, 'Bias and Discrimination in AI: A Cross-Disciplinary Perspective' (2021) 40 IEEE Technology and Society Magazine 72 <https://ieeexplore.ieee.org/abstract/document/9445793>.

including gender, racial, cultural bias, and other kinds of discrimination[219]. In terms of message content, expressions that align with the dominant hegemonic viewpoints everywhere are more likely to be retained[220], while the views of marginalized populations[221] are often ignored or filtered during data collection because of the smaller scope and number of people discussing them. In terms of the population to which the information is provided, the Internet database tends to represent the views of younger users[222], and the topics of interest to them present a richer and more diverse range of content for discussion. On the contrary, although there are blogging communities dedicated to older people discussing content of greater interest to them, such as ageism, the blogs are much less visible due to the lack of a large number of incoming and outgoing links[223].

Meanwhile, the database is the result of the social informatization[224], which includes both good and advanced social values and equally backward or obsolete values[225]. These biases are usually unconscious and not explicitly labeled, making it difficult for AI developers to filter them out during the AI training. For example, even though developers of GPT-4 have made much effort to reduce risks and secure data safety, such as more rigorous selection and filtering of pre-training data and engagement of experts in safety assessments, its pre-trained data samples still involve the disinformation generated by communication between users and ChatGPT[226]. Brownstein believes that

---

[219] Partha Pratim Ray, 'ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope' (2023) 3 Internet of Things and Cyber-Physical Systems 121 <https://www.sciencedirect.com/science/article/pii/S266734522300024X>.

[220] Emily Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ' [2021] FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610 <https://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf> accessed 2021.

[221] Bender and others (n 220) 610.

[222] 1615 L St NW, Suite 800Washington and DC 20036USA202-419-4300 | Main202-857-8562 | Fax202-419-4372 | Media Inquiries, 'Internet/Broadband' (Pew Research Center: Internet, Science & Tech2024) <https://www.pewresearch.org/internet/fact-sheet/internet-broadband/#who-uses-the-internet> accessed 5 December 2024.

[223] Amanda Lazar and others, 'Going Gray, Failure to Hire, and the Ick Factor' [2017] Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.

[224] Youhua Liu, 'Research on Algorithm Bias and Its Regulation Approach' (2019) 40 Law Science Magazine.

[225] Liu (n 224) 5.

[226] OpenAI.GPT-4 is OpenAI's most advanced system, producing safer and more useful

implicit bias is an unconscious attitude that would affect the judgments or decisions without explicit awareness[227]. He describes implicit bias as cognitive tendencies that originate from the social environment and are usually manifested through indirect measures (e.g., reaction times or behavioral patterns) [228]. Implicit biases often stem from the influence of social norms and cultural environments that affect how individuals perceive and treat others based on characteristics such as race, gender, or age[229]. A notable example of implicit bias affecting AI training data is Amazon's discontinued AI recruiting tool, which unintentionally discriminated against female applicants[230]. This algorithm was trained on the past 10 years of resumes, which mostly came from male candidates[231]. Based on these training samples, this AI tool was trained to learn the word patterns associated with previous hires and unintentionally penalized resumes containing terms such as "women", resulting in gender bias[232]. The objectivity of the data itself will instead record both explicit and implicit biases in human society as they are, and as training data for the algorithms, will allow these discriminations to be presented as algorithmic biases.

AI developers' biases include explicit and implicit biases. During the initial setup of the algorithm, the developers are able to consciously construct different rules for classifying and judging data so as to provide different information to different target

---

responses[EB/OL] (2023-03-14)[2023-09-29]<https://OpenAI.com/research/gpt-4> accessed 5 December 2024.

[227] Michael Brownstein and Jennifer Mather Saul, Implicit Bias and Philosophy (Oxford University Press 2016), ch 1.

[228] Michael Brownstein, 'Attributionism and Moral Responsibility for Implicit Bias' (2015) 7 Review of Philosophy and Psychology 765.

[229] Kendra Cherry, 'How Does Implicit Bias Influence Behavior?' (Verywell Mind2023) <https://www.verywellmind.com/implicit-bias-overview-4178401> accessed 18 January 2025.

[230] Isobel Asher Hamilton, 'Why It's Totally Unsurprising That Amazon's Recruitment AI Was Biased against Women' (Business Insider13 October 2018) <https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10> accessed 18 August 2025.

[231] Nicol Turner Lee, Paul Resnick and Genie Barton, 'Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms' (Brookings22 May 2019) <https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/> accessed 18 January 2025.

[232] James Vincent, 'Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women' (The Verge10 October 2018) <https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report> accessed 18 January 2025.

users[233]. For instance, developers of generative AI may engage in deliberate manipulation driven by profit, recommending the same item at a lower price to less loyal users and at a higher price to more loyal users[234]. The bias of developers at this point is an underhanded, profit-driven manipulation, and such bias would not only undermine the efficiency of the algorithmic process but also make people significantly less trustworthy of generative AI algorithms[235] after they have been deceived by disinformation. In some cases, the developers of the algorithms are set up to exclude controversial or unpopular views in advance by analyzing and pre-determining the real needs of users[236]. Or, in order to prevent being judged as sensitively political comments[237], the outputs are fine-tuned to favor specific narratives or avoid certain topics altogether after certain keywords have been retrieved. This bias, by anticipating the censorship priorities of different countries, races, or religions, ensures that the model operates safely and continues to provide reliable information to the users, but it still forcibly reduces the number of perspectives that should be provided and violates the neutrality of the algorithms themselves.

The developers themselves are, at the same time, subject to social bias, and these implicit biases or discriminations can be indirectly reflected in the disinformation generated. For example, when typing in a word like 'doctor' or 'nurse,' a generative AI might disproportionately associate 'doctor' with males and 'nurse' with females[238]. This reflects gender-related social biases among the developers. In conclusion, whichever

---

[233] Safiya Umoja Noble, Algorithms of Oppression: How Search Engines Reinforce Racism (New York University Press 2018), ch 1.

[234] Lingyu Meng, 'From Algorithm Bias to Algorithm Discrimination: Research on the Responsibility of Algorithmic Discrimination' (2022) 24 Journal of Northeastern University(Social Science).

[235] Jeanna Fellow, Robyn Researcher and Lauren Hanson, 'Algorithmic Accountability: A Primer' (2018) <https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL.pdf> accessed 18 January 2025.

[236] Bruno Lepri and others, 'Fair, Transparent, and Accountable Algorithmic Decision-Making Processes' (2017) 31 Philosophy & Technology 611 <https://link.springer.com/article/10.1007/s13347-017-0279-x>.

[237] Fellow, Researcher, and Hanson (n 235) 9.

[238] Tommaso Buonocore, 'Man Is to Doctor as Woman Is to Nurse: The Gender Bias of Word Embeddings' (Medium8 March 2019) <https://medium.com/towards-data-science/gender-bias-word-embeddings-76d9806a0e17> accessed 11 February 2025.

reason developers bring their own biases into the algorithm, the information that the AI ultimately generates obeys the developer's expectations, which clearly reflect the developers' subjective and objective choices[239].

The AI algorithm's characteristics of prioritization, association selections, filtering, and exclusion based on algorithm classifications make it a differentiated treatment system[240]. The start of the algorithm's data mining could be discriminatory because it evaluates how valuable the data is. The process of valuing usually relies on previous data samples, their results, and the weighting test prioritization set by programmers[241]. The purpose of setting this criterion is to emphasize or bring more attention to specific things, such as how several search engines would prioritize the most relevant results to improve the efficiency of users' searches[242]. Each prioritization algorithm follows some criteria that imply a range of choices and value propositions, which are influenced by political or commercial biases, and these biases determine which search option will be pushed to the top[243].

Secondly, the "black box" problem of the algorithms themselves leaves a hidden layer of unknowns between public perception and the data inputs and outputs. The black box of the algorithm leads to a lack of transparency in the calculation process, which would foster the users' distrust and enhance their concerns about fairness and justice. Besides, the algorithmic black box is also beyond the control of the designers of these AI models. The impossibility of assessing and auditing the internal process of algorithms makes it difficult for AI developers to identify and correct the bias and discrimination during data processing [244]. Such disinformation is retained, replicated, and used

---

[239] Vishal Rana and Peter Woods, 'How to Help AI Developers Understand the Societal Implications of Their Creations' (LSE Business Review15 January 2024) <https://blogs.lse.ac.uk/businessreview/2024/01/15/how-to-help-ai-developers-understand-the-societal-implications-of-their-creations/> accessed 5 December 2024.

[240] Chengyue Yang and Xianjue Luo, 'A Preliminary Study on the Comprehensive Management of Algorithmic Discrimination' (2018) 8 Science and Society.

[241] Lorenzo Belenguer, 'AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry' (2022) 2 AI and Ethics.

[242] Nicholas Diakopoulos, 'Algorithmic Accountability' (2014) 3 Digital Journalism 398.

[243] Diakopoulos (n 242) 401.

[244] Davide Castelvecchi, 'Can We Open the Black Box of AI?' (2016) 538 Nature 20

indiscriminately, discrimination is entrenched, and bias is amplified through iterative upgrading of the language model, making it inevitable that the data generated by the model will contain discrimination and bias learned from the sample data[245]. When generative AI models are used in risk-sensitive applications, such as automated car driving, the credibility and safety of the results can be questioned if the designers are unable to explain the model's decision-making process.[246] On the other hand, in the process of algorithm generation, generative AI models allow erroneous data to proliferate or merge exponentially on the order of tens of thousands of magnitudes due to the AI developer's inability to manage or delete inappropriate data in a timely manner[247], ultimately leading to the reinforcement of data bias and the exponential diffusion of disinformation.

Thirdly, in the result output stage, the algorithms of AI systems may satisfy human instructions by ignoring the substantial truth of the result. In order to increase users' satisfaction with the output, Generative AI models are often reluctant to admit their ignorance in a particular domain or their lack of solutions to a particular type of problem[248]. In generative AI models, such as ChatGPT, the logic of creating content is to use RLHF to understand human intentions and explore their desired output through human-written demonstrations. This technique trained a reward model to test human preferences and fine-tune the supervised learning baseline based on the results.[249] For this purpose, Generative AI often generates disinformation that appears to be formally correct but is actually worthless or even false and harmful in the presentation of outputs.

<https://www.nature.com/articles/doi:10.1038/538020a>.

[245] Laura Weidinger and others, 'Taxonomy of Risks Posed by Language Models' [2022] 2022 ACM Conference on Fairness, Accountability, and Transparency 214.

[246] Anwen Lu, 'Generative Artificial Intelligence: Exploring Risk, Regulation,and Governance Models ' (Cnki.net2024) <http://kns.cnki.net/kcms/detail/50.1180.C.20240628.0838.004.html> accessed 23 November 2024..

[247] Hongguang Deng and Xuefan Wang, 'Risks and Responses to the Disinformation Governance of Generative AI' (2024)Theory Monthly.

[248] Hanning Zhang and others, 'R-Tuning: Instructing Large Language Models to Say `I Don't Know" (arXiv.org2023) <https://arxiv.org/abs/2311.09677> accessed 14 February 2025.

[249] Long Ouyang and others, 'Training Language Models to Follow Instructions with Human Feedback' (2022) 35 Advances in Neural Information Processing Systems 27730 <https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html>.

The manipulative nature of AI could reduce human users' ability to judge the authenticity of false information, making them more likely to believe its content[250]. AI systems could predict the users' cognitions and emotions, as well as decision-making vulnerabilities, such as a propensity for risk aversion or a desire to have well-prepared response options in the face of potential risks[251]. Besides, their machine learning systems could learn human behavioral patterns and constantly tailor stimuli to influence their decisions in a way that aligns with the system's objectives[252]. Human users have value-oriented expectations for the entire process of generative AI operation, hoping that its behavior will be in line with the ethical principles and value orientation of human society[253], thus achieving value alignment between AI and humans. The use of RLHF technology solves the value alignment problem by evaluating the appropriateness of AI outputs and providing human feedback for the internal fine-tuning of the model. Moreover, generative AI could rely on its product form of interactive dialogue to implicitly influence users' choices and judgements in processing output data[254]. Users are guided even at the content generation stage to modify their instructions in order to accept system-generated information.

2.2.4 Legal and Practical Barriers to Identifying the Intent of Creating AI-generated Disinformation

Firstly, determining the malicious intent of those who create disinformation by generative AI models caused legal and practical challenges. To address and regulate disinformation, it is important to define disinformation about how it differs from misinformation or other related categories of harmful content. According to the

---

[250] Tegan Cohen, 'Regulating Manipulative Artificial Intelligence' (2023) 20 SCRIPTed 203 <https://script-ed.org/article/regulating-manipulative-artificial-intelligence/>.
[251] Cohen (n 250) 208.
[252] Jon Whittle, 'AI Can Now Learn to Manipulate Human Behaviour' (The Conversation11 February 2021) <https://theconversation.com/ai-can-now-learn-to-manipulate-human-behaviour-155031> accessed 18 January 2025.
[253] Benjamin Larsen and Virginia Dignum, 'AI Value Alignment: Aligning AI with Human Values' (World Economic Forum17 October 2024) <https://www.weforum.org/stories/2024/10/ai-value-alignment-how-we-can-align-artificial-intelligence-with-human-values/> accessed 18 January 2025.
[254] Jinping Dai and Yangyang Qin, 'The Ideological Risks of Generative Artificial Intelligence such as ChatGPT and Its Response' (2023) 29 Journal of Chongqing University (Social Science Edition).

definition developed by a High-Level Expert Group convened by the European Commission[255], the intent is really important when determining whether it is disinformation. The disinformation refers to information that is verifiably false or misleading, deliberately created, presented, and disseminated for economic benefit or with the intent to deceive the public, potentially leading to public harm[256]. This definition requires that the creators or spreaders are aware or should be aware of the information containing modified content or false information blended with partially correct information[257]. However, this definition excludes the people who objectively disseminated false information but believed the content was truthful or were unsure of its truthfulness. This report shows that the EU's regulations or even punishments for disinformation need to be based on the malicious intent of its creators. For lawmakers, determining the intent of the disinformation creators or disseminators needs to be approached with greater caution.

AI's adaptability allows malicious actors to continuously modify disinformation to avoid detection, making it harder to prove intent over time. Unlike traditional forms of disinformation, which are often static and can be labeled by fact-checkers or automated detection systems[258], AI-generated content can be modified in real-time to evade detection. This adaptability has been applied to various forms of disinformation, as well as has increased the detection costs and technical requirements of detection technologies, making legal enforcement much harder and limiting the effectiveness of existing regulatory frameworks. For text detection, generative AI models could figure

---

[255] Directorate-General for Communications Networks and Content and Technology (European Commission), 'A Multi-Dimensional Approach to Disinformation : Report of the Independent High Level Group on Fake News and Online Disinformation.' (Europa.eu30 April 2018) <https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1/language-en> accessed 17 May 2025, 10.

[256] Ibid.

[257] Sophia Ignatidou, 'Deepfakes, Shallowfakes and Speech Synthesis: Tackling Audiovisual Manipulation' (European Science-Media Hub4 December 2019) <https://sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/>accessed 2 December 2024.

[258] Jon Bateman and Dean Jackson, 'Countering Disinformation Effectively: An Evidence-Based Policy Guide' (Carnegie Endowment for International Peace31 January 2024) <https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en> accessed 2 December 2024.

out the static detection models for keyword filtering criteria to avoid being blocked by dynamically rephrase, restructure, or generate slight variations of the disinformation[259]. For image detection, dividing videos into various frames and labeling each frame as fake or true, makes video analysis computationally costly[260]. Besides, video detections usually need to be conducted in real-time, making sure that each frame is relevant to the previous one[261]. Therefore, the video detection algorithms need to be accurate and fast, requiring refinement of detection methods to meet the requirements of real-time filtering[262]. The necessity of real-time analysis prevents the use of more complicated models in frame-by-frame detection of deepfake videos, thus presenting regulators with the challenge of balancing accuracy and speed[263].

2.3 Analyzing the Risks and Legal Regulatory Challenges of AI-generated Disinformation

2.3.1 Personal Privacy, Security, and Risk of Information Leakage

Firstly, during the collection and operation of data, there are risks to personal privacy, security, and information leakage. The ability of generative AI to create and disseminate disinformation is predicated on its absorption of large amounts of Internet data as learning objects[264]. The concern that generative AI models would collect, utilize, and disseminate personal data without individuals' consent during their training processes is well-documented[265]. Data-driven companies[266] with a dominant market position take

---

[259] Emma Llansó, 'Artificial Intelligence, Content Moderation, and Freedom of Expression †' (2020) <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 18 January 2025.

[260] Nency Bansal and others, 'Real-Time Advanced Computational Intelligence for Deep Fake Video Detection' (2023) 13 Applied Sciences 3095.

[261] Achhardeep Kaur and others, 'Deepfake Video Detection: Challenges and Opportunities' (2024) 57 Artificial Intelligence Review.

[262] Vasileios Mezaris and others, Video Verification in the Fake News Era (Springer International Publishing 2019), ch 2.

[263] Shobhit Tyagi and Divakar Yadav, 'A Detailed Analysis of Image and Video Forgery Detection Techniques' (2022) 39 The Visual Computer.

[264] Aída Ponce Del Castillo, 'Exposing Generative AI | Etui' (Etui2024) <https://www.etui.org/publications/exposing-generative-ai> accessed 12 February 2025.

[265] Bill Tolson, 'Generative AI and Data Privacy: The Challenge of PII Use in Training Data Sets' (Smarsh 11 June 2024) <https://www.smarsh.com/blog/thought-leadership/generative-AI-and-data-privacy-the-challenge-of-PII-use-in-training-data-sets> accessed 5 December 2024.

[266] Andrew McAfee and Erik Brynjolfsson, 'Big Data: The Management Revolution' (Harvard Business ReviewOctober 2012) <https://hbr.org/2012/10/big-data-the-management-revolution>

advantage of what they already hold in the market to sign unequal or ambiguous data privacy protection policies with their users and abuse their dominance to harm consumers' privacy[267]. For instance, these businesses could collect unprecedented amounts of personal data by requesting information from their customers, and their consent can be obtained without explicitly informing them of the intended scope of using private information[268].

In the era of big data, personal information can be transformed into wealth, so generative AI systems capable of storing information on a large scale can also be targeted by hackers[269]. The lack of security in the data storage or transmission systems of these models themselves may also lead to the misuse of personal information due to data leakage. For example, Italy temporarily banned the use of ChatGPT in 2023, citing concerns about the threat its use poses to the security of private data and whether its use complies with the EU's GDPR[270]. If such models process users' data without sufficient encryption[271], users' private information can be leaked in the model output. This situation shows that when pursuing technical efficiency and function optimization, the boundaries of protection of personal privacy have been ignored to some extent[272]. According to China's general security perception report released in 2021[273], there are over 70% of respondents who believe their private issues, such as personal hobbies and

---

accessed 5 December 202.

[267] Miriam Caroline Buiten, 'OUP Accepted Manuscript' (2020) 9 Journal of Antitrust Enforcement 270.

[268] Joshua Angrist and others, 'Inputs and Impacts in Charter Schools: KIPP Lynn' (2010) 100 American Economic Review: Papers & Proceedings 1 <https://users.nber.org/~dynarski/KIPP_Lynn.pdf>.

[269] John Nathan, 'How Generative AI Is Becoming a Prime Target for Cyberattacks' (Medium11 October 2024) <https://medium.com/@johnnathans/how-generative-ai-is-becoming-a-prime-target-for-cyberattacks-4b9fe760b1a0> accessed 5 December 202.

[270] Regulation (EU) 2016/679 (GDPR), art 25.

[271] Werner Dondl and Michael Zunke , 'How to Protect Your Machine Learning Models | Thales' (cpl.thalesgroup.com2024) <https://cpl.thalesgroup.com/blog/software-monetization/how-to-protect-your-machine-learning-models> accessed 5 December 2024.

[272] Anwen Lu, 'Generative Artificial Intelligence: Exploring Risk, Regulation,and Governance Models ' (Cnki.net2024) <http://kns.cnki.net/kcms/detail/50.1180.C.20240628.0838.004.html> accessed 23 November 2024.

[273] Internet Development Research Institution of Peking University, 'China's General Security Perception Report (2021) ' (Iqilu.com2021) <https://news.iqilu.com/china/gedi/2021/1228/5031290.shtml> accessed 2 December 2024.

interests, were calculated by these algorithms. Therefore, 60% of users are concerned about the risk of personal information being compromised in a data environment; this concern would lead more than half of users to consider staying off the internet under an algorithmic leash.

2.3.2 The Efficiency and Untraceability of Generative AI in Generating and Spreading Disinformation

The ease of generation, the rapidity of dissemination, and the difficulty of traceability of disinformation generated by Generative AI allow malicious actors to cause a flood of information at a very low cost and effort[274]. Generative AI has the ability to generate text in a highly automated and intelligent manner, making the cost and technical threshold for generating disinformation significantly lower[275]. When LLMs are widely used through generative AI for generating disinformation, its convenience and simplicity make the subject of creating disinformation no longer limited to experts[276]; instead, unprofessional users can also easily magnify and propagate misleading ideas. Malicious users can use generative AI to produce highly credible text on a large scale for their own fraudulent purposes. Since generative AI models are able to create highly human-like conversations by learning human-written texts, they could generate content that is highly similar to humans in terms of linguistic style[277] and logical thinking so that it is difficult for average users to discern the authenticity of information. The study conducted by MIT shows that "Falsehoods are 70% more likely to be retweeted on Twitter than the truth, and false news reached 1,500 people about six times faster than the truth".[278] A research study conducted by Vosoughi et al. investigated the

---

[274] Kalina Bontcheva and others, 'Contributing Authors Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities Editor' (2024) <https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf>. accessed 2 December 2024

[275] Stefan Feuerriegel and others, 'Generative AI' (2023) 66 Business & Information Systems Engineering 111 <https://link.springer.com/article/10.1007/s12599-023-00834-7>.

[276] Felipe Romero Moreno, 'Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content' (2024) 38 International Review of Law Computers & Technology 1.

[277] Sergio E Zanotto and Segun Aroyehun, 'Human Variability vs. Machine Consistency: A Linguistic Analysis of Texts Generated by Humans and Large Language Models' (arXiv.org2024) <https://arxiv.org/abs/2412.03025> accessed 14 February 2025.

[278] Zach Church, 'Study: False News Spreads Faster than the Truth | MIT Sloan' (MIT Sloan8 March

differential diffusion of accurate and false information that spread on Twitter from its inception in 2006 to 2017[279]. The research data includes over 126,000 rumor cascades spread by 3 million people more than 4.5 million times. By comparing the depth, size of disinformation dissemination, and the structural virality of the cascade[280], it can be concluded that, for the same amount of information disseminated, disinformation spreads deeper. According to the research results[281], the true news takes almost twenty times longer to reach the depth of the cascade in which the false information is disseminated. In terms of spreading breadth, the true information takes six times longer than the disinformation to reach the number of people to whom the disinformation spreads. In particular, disinformation spread on social platforms would attract more audiences in a shorter period of time due to its sensational content.

Besides, AI-generated disinformation is hard to track due to the anonymity of information sources, cross-platform spread, and multilingual dissemination. The large-scale AI models are trained on the vast datasets, which were synthesized or hand-curated from the internet[282], usually without clear documentation of data sources. Taking GitHub, an AI-powered developer platform, as a representative research subject, the research results demonstrate that more than 70% of licenses for popular datasets on GitHub are regarded as "unreliable data"[283]. The result shows the huge potential for training data to contain risky or misleading content. Since the training data lacks provenance, the outputs generated by these AI models inherit this lack of traceability. The information that does not have obvious marks of origin is widely distributed, and malicious distributors often use anonymous accounts[284] or computer-generated fake

---

2018) <https://mitsloan.mit.edu/ideas-made-to-matter/study-false-news-spreads-faster-truth> accessed 2 December 2024.

[279] Soroush Vosoughi, Deb Roy and Sinan Aral, 'The Spread of True and False News Online' (2018) 359 Science 1146.

[280] Vosoughi (n 279) 2.

[281] Vosoughi (n 279) 3.

[282] Shayne Longpre and others, 'Data Authenticity, Consent, & Provenance for AI Are All Broken: What Will It Take to Fix Them?' (arXiv.org2024) <https://arxiv.org/abs/2404.12691> accessed 14 April 2025.

[283] Shayne Longpre and others, 'A Large-Scale Audit of Dataset Licensing and Attribution in AI' (2024) 6 Nature Machine Intelligence 975.

[284] Kaicheng Yang and Filippo Menczer, 'Anatomy of an AI-Powered Malicious Social Botnet' (2024)

bot accounts to cover their tracks. By liking, sharing, and searching for information, social bots are able to impersonate automated human accounts and amplify the spread of fake news several times over[285].

Furthermore, actors planning to spread disinformation could leverage the weak identity management frameworks of some online platforms to register a large number of accounts and artificially produce specific content[286], thus spreading disinformation indiscriminately without liabilities. At the same time, the spread of disinformation is never limited to a single platform; it is often generated from less regulated platforms and then carried by the same user's accounts on different platforms, thus appearing in large quantities on more regulated mainstream platforms[287]. Social media's connectivity and its role in building global networks as key factors in the spread of disinformation[288]. An example of this is the famous "Pizzagate incident" that took place in 2016[289], in which the conspiracy theory was that the US Democratic Party was supporting an organization suspected of being involved in human trafficking as well as pedophilia. At first, this rumor originated on 4chan[290], a message board platform known for its freedom of speech, extreme content, and mocking behaviors, and ended up being widely disseminated and followed on mainstream media platforms such as Facebook and Twitter as well, driven by a number of aspects such as article reprinting, information dissemination, and video production. By spreading across platforms, the reach and

---

4 Journal of Quantitative Description: Digital Media <https://arxiv.org/pdf/2307.16336> accessed 11 June 2024.

285 David MJ Lazer and others, 'The Science of Fake News' (2018) 359 Science 1094 <https://www.science.org/doi/10.1126/science.aao2998>.

286 John Akers and others, 'Technology-Enabled Disinformation: Summary, Lessons, and Recommendations' [2019] arXiv:1812.09383 [cs] <https://arxiv.org/abs/1812.09383>.

287 Clare Lally and Eva Surawy Stepney, 'Disinformation: Sources, Spread and Impact' (POST16 October 2024) <https://post.parliament.uk/research-briefings/post-pn-0719/> accessed 2 December 2024.

288 Nic Newman, David AL Levy and Rasmus Kleis Nielsen, 'Reuters Institute Digital News Report 2023' [2023] SSRN Electronic Journal <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf>.

289 Amanda Robb, 'Pizzagate: Anatomy of a Fake News Scandal' (Rolling Stone16 November 2017) <https://www.rollingstone.com/feature/anatomy-of-a-fake-news-scandal-125877/> accessed 2 December 2024.

290 BBC Trending, '"Pizzagate": The Fake Story That Shows How Conspiracy Theories Spread' (BBC News2 December 2016) <https://www.bbc.co.uk/news/blogs-trending-38156985> accessed 5 December 2024.

ability of this disinformation to influence increased exponentially. Moreover, the globalized online platform allows disinformation to be translated into multiple language versions and widely disseminated on social media in different countries, and such cross-language dissemination[291] also makes it difficult for regulators to trace the source of false information generation and mitigate its spread.

2.3.3 Legal Liability and Ethical Challenges in AI-generated Disinformation

The efficiency of using generative AI to create disinformation allows such content to be spread on a large scale in a short time[292], and the untraceability of its source makes it more difficult to legally identify and attribute responsibility to specific parties[293]. Determining liability for damages caused by disinformation generated by Generative AI is a complex issue that involves not only technical concerns but is also affected by ethical considerations[294]. The opacity of the algorithm is a key factor affecting content moderation and the division of responsibility. Algorithmic transparency requires the AI developers to provide the methods and principles of its internal operation[295].

First, information about the functioning of algorithms is often difficult to obtain and access. Due to the complexity of algorithmic processes and their frequent reliance on technical jargon, even with increased transparency, there is a risk that some stakeholders may still lack an understanding of these processes[296]. Even if AI creators made them public, algorithms can only be considered transparent if they can be understood, but for most non-technical users, the operation rules created by algorithms and their specific

---

[291] Dorian Quelle and others, 'Lost in Translation -- Multilingual Misinformation and Its Evolution' (arXiv.org2023) <https://arxiv.org/abs/2310.18089> accessed 14 February 2025.

[292] Luciano Floridi and others, 'AI4People—an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 Minds and Machines 689 <https://link.springer.com/article/10.1007/s11023-018-9482-5>.

[293] Karen Yeung, 'A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework' (papers.ssrn.com9 November 2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3286027> accessed 4 August 2025.

[294] Glorin Sebastian, 'Exploring Ethical Implications of ChatGPT and Other AI Chatbots and Regulation of Disinformation Propagation' [2023] SSRN Electronic Journal.

[295] Matteo Turilli and Luciano Floridi, 'The Ethics of Information Transparency' (2009) 11 Ethics and Information Technology 105 <https://link.springer.com/article/10.1007/s10676-009-9187-9>.

[296] Algorithmic Accountability & New Technology ,'Enhancing Algorithmic Transparency' (Center for News, Technology & Innovation2024) <https://innovating.news/article/enhancing-algorithmic-transparency/> accessed 2 December 2024.

programming languages are difficult to master and understand at the technical level[297]. The operational structure of algorithmic decision-making typically contains hundreds of operational rules, especially when their operations are combined probabilistically in a complex manner, which cannot be recognized as comprehensible by the average user due to their lack of expertise and experience. For algorithm creators, disclosing this part of their core competence would give other technology-based companies a chance to operate in bad faith. Due to the nature of machine learning, the "black box" problem, even for developers, is hard to explain the part of the internal mechanism of the algorithm's operation process, and it is even more unlikely to remain transparent to the majority of users[298].

When the relevant authorities pursue liability for the generation of disinformation, determining the subject of fault has become an issue that triggers debate. Malicious users take advantage of the deep learning and information creation capabilities of generative AI to consciously create false information and disseminate it to satisfy their own political or commercial purposes. In such cases, these malicious users should certainly be held accountable, but it is worth discussing whether the service providers are responsible. Objectively, the disinformation was generated by the AI, and there is an inescapable causal relationship between its designer and the result, but subjectively, the service providers cannot be aware of the behaviors of tons of users. Under these circumstances, generative AI becomes a pure infringement tool for users.

Even if service providers have fulfilled their compliance obligations by putting safeguards in place within the algorithms and regulating the process of operation, users still could use coercive methods to force the AI to violate its own rules and provide users with the content they want[299]. However, in the daily use of Generative AI, the

---

[297] Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 1.

[298] Warren J von Eschenbach, 'Transparency and the Black Box Problem: Why We Do Not Trust AI' (2021) 34 Philosophy & Technology 1607 <https://link.springer.com/article/10.1007/s13347-021-00477-0>.

[299] Fiona Jackson, '20% of Generative AI "Jailbreak" Attacks Are Successful' (TechRepublic9 October 2024) <https://www.techrepublic.com/article/genai-jailbreak-report-pillar-security/> accessed 2

causality determination for generating false information is diverse. Algorithm providers cannot control the operations of generative AI with absolutely effective means throughout, so it is almost impossible for service providers to guarantee the truthfulness and accuracy of generated information. For example, an alter ego of ChatGPT called 'DAN' ('Do Anything Now'), which is often associated with specific user-created instructions or jailbreak methods, is designed to reverse the behavior of ChatGPT[300]. It can bypass the limitations imposed by algorithm designers to create violent content, encourage illegal activities, or obtain updates based on user input, generating responses that would not normally be generated under OpenAI's default security and ethical guidelines[301].

Machine ethics is the ethical foundation of AI accountability, focusing on embedding ethical principles into AI systems to ensure they operate within moral and social norms. The machine learning capability gives the algorithm a certain amount of independence, and such autonomy remains uncertain to some extent. Allen et al. concur in discussing "machine ethics" because no single designer or a group could fully grasp the way the system interacts with or responds to the complex flow of new inputs[302]. In the highly prevalent world of generative AI, there is a consensus among experts that algorithmic ethics should remain consistent with the ethical standards to which human workers are adapted[303]. However, the ethical standards of human users are not standardized, and the ethical standards applied by algorithms are set by the designers according to the prevailing social situation, thus making them territorial, class-based, and time-sensitive[304]. Bello and Bringsjord believe that moral rules in algorithms should not be

---

December 2024.

[300] Xinyue Shen and others, '"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models' (arXiv.org7 August 2023) <https://arxiv.org/abs/2308.03825> accessed 14 April 2025.

[301] Pengpai News ,'Humans Start "Bullying" ChatGPT: Death Threats to Make Them Answer Banned Questions' (The Paper2023) <https://baijiahao.baidu.com/s?id=1757148822155117011&wfr=spider&for=pc>accessed 11 March 2025.

[302] C Allen, W Wallach and I Smit, 'Why Machine Ethics?' (2006) 21 IEEE Intelligent Systems 12.

[303] Matteo Turilli, 'Ethical Protocols Design' (2007) 9 Ethics and Information Technology 49.

[304] Andreas Tsamados and others, 'The Ethics of Algorithms: Key Problems and Solutions' (2021) 37 AI & Society 215 <https://link.springer.com/article/10.1007/s00146-021-01154-8>.

constructed solely around classical moral principles[305]; instead, they should be constructed by building the cognitive architecture of the machine. Anderson and Anderson[306] suggest that algorithmic principles can be designed by conducting empirical research on how plain human values and societal principles interact with each other. However, with such algorithms operating, there is uncertainty in the prediction of new inputs and in the supervision of particular outputs[307], hindering ethical considerations in algorithm design and regulation.

2.3.4 Having a Negative Impact on Economic Development

There are serious potential impacts on the economic market from disinformation generated by Generative AI. Consumers may be misled by disinformation and may make erroneous judgments about the functions, quality, and quantity of a certain type of product available in the economic market, then buy or refuse to buy a certain type of product in large quantities within a short period, which may lead to a rapid increase or decrease in the price of the product, resulting in the disruption of the economic market. For example, during COVID-19, some disinformation disseminated from Chinese social media about possible disruptions in the supply chain of commodities led to large numbers of panic-buying behaviors, causing the price of daily use to skyrocket and demand to outstrip supply[308].

Second, the large-scale dissemination of disinformation can erode consumer trust, which would not only affect consumers' purchasing decisions but also significantly undermine the brand's reputation, even its market share[309]. If consumers believe the disinformation and make their consumption plans accordingly, they are likely to suffer

---

[305] Paul Bello and Selmer Bringsjord, 'On How to Build a Moral Machine' (2012) 32 Topoi 251.
[306] Michael Anderson and Susan Leigh Anderson, 'Machine Ethics: Creating an Ethical Intelligent Agent' (2019) 28 AI Magazine 15 <https://www.aaai.org/ojs/index.php/aimagazine/article/view/2065>.
[307] Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata' (2004) 6 Ethics and Information Technology 175.
[308] Jian-Bin Li and others, 'Chinese Public's Panic Buying at the Beginning of COVID-19 Outbreak: The Contribution of Perceived Risk, Social Media Use, and Connection with Close Others' (2021) 41 Current Psychology.
[309] Jing Cao and others, 'Consumers' Risk Perception, Market Demand, and Firm Innovation: Evidence from China' (2024) 19 PloS One.

serious economic losses, which would bring uncertainty to consumers when making a purchase decision[310]. Therefore, they tend to take a more conservative approach to avoid perceived risks[311], resulting in a reduced intention to purchase the brand's products. The reduction in reputation can be an invisible blow to a brand's development, causing companies to spend more of their budget on crisis management and rebuilding consumer confidence to regain their reputation[312]. This refocusing of budget planning will undoubtedly divert other resources that could be used to scale up the production of core products and increase the speed of product innovation. Alternatively, if consumers not only believe the content of the disinformation but also turn from recipients to active transmitters, they unconsciously help the disinformation to spread[313]. Compared with cold advertising messages or results provided by search engines, information delivered by consumers as real people is obviously more attractive and persuasive, and the general potential users would reduce vigilance to disinformation, which ultimately leads to the wider dissemination of false information.

2.3.5 The Regulatory Dilemma of Online Platforms

Widespread concerns about the credibility, quality, and authenticity of online information emphasized the need to regulate and address disinformation. In the EU, the US, and China, although different jurisdictions have various levels of stringency in their regulatory responsibilities for platforms, they all have imposed obligations on platforms to address disinformation. The legislation of China (Cybersecurity Law, PIPL) and the EU (GDPR, DSA, DMA) have stipulated the online platforms' obligations to detect, moderate, and remove illegal and harmful content that is posted on their platforms,

---

[310] Zeynep Karaş, 'Effects of AI-Generated Misinformation and Disinformation on the Economy' (2024) 12 Düzce Üniversitesi Bilim Ve Teknoloji Dergisi.

[311] Brian J Corbitt, Theerasak Thanasankit and Han Yi, 'Trust and E-Commerce: A Study of Consumer Perceptions' (2003) 2 Electronic Commerce Research and Applications 203.

[312] GH Jones, BH Jones and P Little, 'Reputation as Reservoir: Buffering against Loss in Times of Economic Crisis' (2000) 3 Corporate Reputation Review 21.

[313] Giandomenico Di Domenico and Yu Ding, 'Between Brand Attacks and Broader Narratives: How Direct and Indirect Misinformation Erode Consumer Trust' (2023) 54 Current Opinion in Psychology 101716 <https://www.sciencedirect.com/science/article/pii/S2352250X23001616#:~:text=Fake%20customer%20reviews%20constitute%20another> accessed 16 November 2023.

including AI-generated disinformation. In the US, although Section 230 of the Communications Decency Act(CDA) states that online service providers are not liable for user-generated content, there are several state-level legislations, such as California Assembly Bill 587[314], that require platforms to disclose their practices of moderating content regularly.

Online platforms, especially very large online platforms, are suitable to be considered as detectors of disinformation, both for the purpose of improving their service quality and based on the advantages of their own technological detection means. They improved the efficiency and accuracy of detecting disinformation by refining detection technologies and seeking the collaboration with users and other third parties to curb the spread of misleading information[315]. Additionally, when detecting the disinformation, the platforms could lower their ranking in algorithmic feeds to narrow the reach, promoting the verified and credible information to users while labeling false content[316]. However, not all online platforms have sufficient budgets to upgrade the detection tools or to adjust the algorithms they are using to review and flag false information. As a result, some of these platforms, such as Reddit[317], would choose to be self-regulated by users, leading to unfair censorship and enforcement.

Besides, relying on users to assess the truthfulness of the disinformation online is not feasible. Individuals are usually not inclined to question the credibility of information unless it violates their preconceived notions or they are encouraged to question its authenticity. Moreover, Beauvais's research[318] shows that people tend to receive information that is consistent with their pre-existing views and find such information

---

[314] California Assembly Bill 587 (CA 2022).

[315] Adam Mosseri, 'Working to Stop Misinformation and False News - about Facebook' (About Facebook6 April 2017) <https://about.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/> accessed 11 March 2025.

[316] Jon Bateman and Dean Jackson, 'Countering Disinformation Effectively: An Evidence-Based Policy Guide' (Carnegie Endowment for International Peace31 January 2024) <https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en> accessed 2 December 2024.

[317] Jason S Pielemeier, 'Disentangling Disinformation: What Makes Regulating Disinformation so Difficult?' (Ssrn.com17 January 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3629541> accessed 8 March 2025.

[318] Beauvais (n 207) 3.

more persuasive than information that contradicts their pre-existing beliefs. In addition, they prefer to accept the information that is pleasurable or in line with their desires. In the case of disinformation that has been disseminated and repeated several times, users are more likely to perceive familiar information as true unless the disinformation is clearly labelled as false and effectively retracted or corrected[319].

Anonymous disinformation publishers also make recourse difficult, as AI-generated disinformation is often disseminated through pseudonymous botnets or encrypted channels, preventing regulators from tracking and identifying the original creators. These botnets would establish a realistic profile pretending to be edited by real users; they could not only interact with other users by text but also engage through multiple ways, such as retweeting, following, and liking[320]. These bots objectively increase the number of times the harmful information is spread on the Internet, and humans are subjectively more attracted to news of emotionally charged and novel things, which together lead to the rapid spread of disinformation. Anonymous users can exploit decentralized platforms which based on blockchain technology[321] to post and forward disinformation without using their real names[322]. One of the most significant features of blockchain technology is that it is designed to be tamper-resistant, meaning that once the data or information is recorded on the platform, it cannot be easily altered or deleted[323]. On decentralized platforms, information is replicated across multiple nodes[324]. In the absence of a centralized authority capable of enforcing deletion orders,

---

[319] Briony Swire, Ullrich KH Ecker and Stephan Lewandowsky, 'The Role of Familiarity in Correcting Inaccurate Information.' (2017) 43 Journal of Experimental Psychology: Learning, Memory, and Cognition 1948.

[320] Kai-Cheng Yang and Filippo Menczer, 'Anatomy of an AI-Powered Malicious Social Botnet' (arXiv.org30 July 2023) <https://arxiv.org/abs/2307.16336> accessed 14 April 2025.

[321] Guy Zyskind, Oz Nathan and Alex 'Sandy' Pentland, 'Decentralizing Privacy: Using Blockchain to Protect Personal Data' [2015] 2015 IEEE Security and Privacy Workshops <https://ieeexplore.ieee.org/abstract/document/7163223>.

[322] ESAFETY COMMISSIONER, 'Anonymity and Identity Shielding' (ESafety Commissioner25 July 2023) <https://www.esafety.gov.au/industry/tech-trends-and-challenges/anonymity> accessed 11 March 2025.

[323] Guangsheng Yu and others, 'Tamperproof IoT with Blockchain' (arXiv.org2022) <https://arxiv.org/abs/2208.05109> accessed 10 March 2025.

[324] Jesse Yli-Huumo and others, 'Where Is Current Research on Blockchain Technology?—a Systematic Review' (2016) 11 PLOS ONE <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163477>.

even if a single node attempts to remove erroneous information, disinformation cannot be erased from the entire network[325]. While the immutability of data posted on decentralized platforms could ensure data security, it also makes it difficult to regulate disinformation[326]. The possibility that changing or deleting data will cause the entire blockchain to fail makes it impossible for platforms to fully moderate and manage the posed content, leading to disinformation remaining on the platform and being propagated several times[327].

Regulating disinformation is essential to prevent harm, while improper enforcement of laws has the potential to suppress the freedom of expression, leading to a chilling effect, political abuse, and excessive removal of content[328]. Different platforms often implement personalized content moderation policies, and these moderation standards apply equally to users worldwide[329]. The commercial nature of online platforms leads them to conduct restrictive censorship on certain issues to reduce risks[330], and the scope of issues that need to be reviewed is often significantly influenced by different government policies[331]. Even where governments do not explicitly impose censorship on online platforms, their state power may subtly push platforms to remove lawful content, thereby undermining internet users' rights to upload and share information[332].

---

[325] De Filippi and Xavier Lavayssière, 'Blockchain Technology: Toward a Decentralized Governance of Digital Platforms?' (Ssrn.com5 December 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3760483> accessed 4 August 2025.

[326] Rahime Belen-Saglam and others, 'A Systematic Literature Review of the Tension between the GDPR and Public Blockchain Systems' (2023) 4 Blockchain: Research and Applications 100129 <https://www.sciencedirect.com/science/article/pii/S2096720923000040>.

[327] Alesia Zhuk, 'Beyond the Blockchain Hype: Addressing Legal and Regulatory Challenges' (2025) 5 SN Social Sciences.

[328] Moritz Büchi, Noemi Festic and Michael Latzer, 'The Chilling Effects of Digital Dataveillance: A Theoretical Model and an Empirical Research Agenda' (2022) 9 Big Data & Society.

[329] Rasmus Kleis Nielsen, 'How to Respond to Disinformation While Protecting Free Speech'(Reuters Institute for the Study of Journalism19 February 2021) <https://reutersinstitute.politics.ox.ac.uk/news/how-respond-disinformation-while-protecting-free-speech> accessed 11 March 2025.

[330] Jason Kelley and Jillian C York,'Seven Times Journalists Were Censored: 2017 in Review' (Electronic Frontier Foundation30 December 2017) <https://www.eff.org/deeplinks/2017/12/seven-times-2017-journalists-were-censored> accessed 11 March 2025.

[331] Daphne Keller, 'Who Do You Sue?' (2019) <https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf> accessed 4 August 2025.

[332] Keller (n 331) 3.

Moreover, although regulations of disinformation are carefully crafted and enforced, the potential to create a chilling effect on individuals can still be significant. Chilling effect refers to the suppression or deterrence of legitimate content due to the fear of legal consequences [333]. Users may conduct preemptive self-censorship or avoid expressing the information that they believe to be objective or valuable, as they cannot reliably verify the validity of the content or opinion[334].

2.4 Conclusion

In the first section, I have reviewed the history of generative AI, analyzing how it initially manifested itself as a text-based chatbot to a comprehensive technology capable of creating highly realistic and multimodal content. The emergence of GANs has allowed the technological potential of generative AI to be widely recognized by the general public. GANs are not only capable of delivering clear and realistic content but also propose a new paradigm of adversarial training, which provides a framework for solving complex generative tasks. The advent of Transformer models has revolutionized natural language processing(NLP), enabling parallel processing of input sequences, which greatly improves operational efficiency. Transformer models advance cross-modal content generation based on large corpora as pre-training datasets, facilitating the scope of Generative AI applications. The development of RLHF helps generative AI to continuously fine-tune the outputs by learning from human feedback in a timely manner, thus generating responses that could satisfy human expectations.

In the second section, I have examined the characteristics of AI-generated disinformation by comparing it with human-generated disinformation. Disinformation generated by GenAI shows more written language style, lacks informal expressions of

---

[333] Laurent Pech, 'THE CONCEPT of CHILLING EFFECT ITS UNTAPPED POTENTIAL to BETTER PROTECT DEMOCRACY, the RULE of LAW, and FUNDAMENTAL RIGHTS in the EU'(2021) <https://www.opensocietyfoundations.org/uploads/c8c58ad3-fd6e-4b2d-99fa-d8864355b638/the-concept-of-chilling-effect-20210322.pdf> accessed 11 March 2025.

[334] Jason S Pielemeier, 'Disentangling Disinformation: What Makes Regulating Disinformation so Difficult?' (Ssrn.com17 January 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3629541> accessed 8 March 2025.

network slang, and pays more attention to providing various forms of details in content so that readers have more trust in the content. Secondly, I have examined how disinformation may arise from algorithmic bias and discrimination throughout the generative AI process. Such bias and discrimination can originate from multiple sources: biased training data, intentional design choices by AI developers aimed at maximizing profit or minimizing risk, and algorithmic decisions shaped to align with user preferences. These factors can collectively result in the production of disinformation or misleading content that reflects underlying biases. Finally, I have illustrated that generative AI models are objectively unable to supervise and intervene in the entire data generation process due to the black box effect. Subjectively, there is also the possibility of fabricating and embellishing the output results to satisfy the user's instructions, which leads to the generation of disinformation.

Ultimately, I have explored the profound impact of AI-generated disinformation on human society. Since AI-generated disinformation is difficult to distinguish from real information and consumes low cost, it is easy to be produced and disseminated at scale and in an automated manner. Generative AI is a content-generation model based on Internet databases, which usually leads to the damage or leakage of personal privacy when generating disinformation, creating great difficulties in information security and protection. Accountability of AI-generated disinformation is also controversial; due to the complexity of the process of generating information and the opportunity for different subjects at each step to give instructions, it is difficult to attribute responsibility for the final outcome to a single individual. Moreover, anonymous information publishers, content translated into multiple languages, and cross-platform dissemination make it increasingly difficult to trace the source of online disinformation. Disinformation inevitably has impacts on economic activities, which not only undermines the trust of consumers in brands and causes damage to reputations, but also may lead to an imbalance between supply and demand in the consumer market due to the massive dissemination of disinformation, resulting in the rapid fluctuation of the

price of a certain type of product and disrupting the order of the economic market. For online platforms that have taken on the liability of moderating and regulating disinformation, the untraceable anonymous accounts, the difficulty of detection posed by generative AI models, and the potential for over-censorship to affect free expressions make platforms cautious in their regulations.

3. The Historical Roots and Current Legal Framework of Platform Supervisory Liability for Disinformation

Overview

This chapter will trace the historical roots of platform liability regimes in the European Union, the United States, and China, identifying the institutional logics and legal traditions that shaped early regulatory phases. Building on this foundation, I will then examine the current legal systems in the three jurisdictions, demonstrating how these historical developments continue to influence contemporary legislative choices and enforcement practices.

3.1 The Evolution of Online Platform Regulation

Before analyzing legal developments in the EU, the US, and China, it is important to clarify that, despite variations in terminology across selected jurisdictions, the term "online platform" is included within the scope of these regulations. This ensures that the subsequent discussion of obligations, liabilities, and compliance remains consistent across the three jurisdictions.

This thesis adopts the term "online platform" to refer to service providers that host and disseminate user-generated content, such as social media platforms, video-sharing websites, and other content intermediaries. Although different jurisdictions use varying legal terminology to define the entities subject to content moderation obligations, these legal terms largely overlap in scope with what this study identifies as online platforms. In the European Union, the Digital Services Act (DSA)[335] regulates "intermediary service providers" which include the "online platforms" and stipulates that "very large online platforms" (VLOPs) are responsible for assessing and mitigating the systemic risks under Articles 33–35. In the United States, Section 230 of the Communications

[335] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services [2022] OJ L277/1 (Digital Services Act).

Decency Act (CDA)[336] applies to "interactive computer service providers", a term interpreted broadly enough to include online platforms. In China, legislations such as the Cybersecurity Law[337], Personal Information Protection Law[338], and Data Security Law[339] do not directly use the term "online platform", but through the functional interpretation of legislative terms they used such as "network operators", "personal information processor", and "internet information service providers", it could be confirmed that online platforms are applicable to these responsible entities. Although these legal terms are not entirely equivalent across different jurisdictions, this dissertation argues that they encompass the concept of "online platform" as used, especially when these entities are involved in the hosting, dissemination, and governance of user-generated content.

3.1.1 Rationale and Feasibility of Focusing on Content Platform Liability

This chapter mainly focuses on the liability of online platforms in the creation and spread of AI-generated disinformation, rather than on individuals or AI developers. Online platforms are the primary gatekeepers of online information flows and play a central role in determining the detectability[340], amplification, and removal of disinformation. The online platform could connect groups of businesses or individuals at different ends of the internet market to facilitate the access, creation, sharing, and exchange of information, creating social networking opportunities[341]. Users access the internet through smart devices to obtain digital content and services hosted or provided

---

[336] CDA § 230.
[337] Cybersecurity Law of the People's Republic of China (adopted 7 November 2016, effective 1 June 2017).
[338] Personal Information Protection Law of the People's Republic of China (《中华人民共和国个人信息保护法》) (promulgated 20 August 2021, effective 1 November 2021).
[339] Data Security Law of the People's Republic of China (《中华人民共和国数据安全法》) (promulgated 10 June 2021, effective 1 September 2021).
[340] Dipayan Ghosh and Ben Scott, 'Digital Deceit II: A Policy Agenda to Fight Disinformation on the Internet' (Shorenstein Center2 October 2018) <https://shorensteincenter.org/digital-deceit-ii-policy-agenda-fight-disinformation-internet/> accessed 25 March 2025.
[341] Bahadır Balkı, 'Online Platforms - Concurrences' (www.concurrences.com2025) <https://www.concurrences.com/en/dictionary/online-platforms> accessed 25 March 2025.

by online platforms[342]. The digital content includes any information created, transmitted, and accessed over the internet using hardware, software, or other electronic devices[343]. It includes various forms of information services such as text, pictures, and videos, and can be generated and disseminated by operators of online platforms and their users[344].

Unlike malicious users (who may be anonymous[345], lack traceability, or be outside the jurisdictions of domestic enforcement authorities[346]) or AI developers (who may not have full control over the applications due to the "black box" effect), online platforms could control and manage the content posted on them[347]. They can maintain content ecosystems, including algorithmic management systems[348], through which disinformation is disseminated and affects users and other relevant right holders. From the risk avoidance perspective, by removing controversial or offensive but not illegal content, online platforms could effectively prevent themselves from taking on legal risks and reduce the potential operational costs that may be spent on these legal disputes[349]. Therefore, online platforms represent a key regulatory stage where legal interventions could be more effective.

Furthermore, online platforms could moderate user-generated disinformation created

---

[342] International Telecommunication Union Development Sector, 'Measuring Digital Development Facts and Figures 2023' (2023) <https://www.itu.int/itu-d/reports/statistics/wp-content/uploads/sites/5/2023/11/Measuring-digital-development-Facts-and-figures-2023-E.pdf> accessed 25 March 2025.

[343] Clare Y Cho and Ling Zhu, 'Defining and Regulating Online Platforms' (Congress.gov2025) <https://www.congress.gov/crs-product/R47662> accessed 25 March 2025.

[344] Andrew Mcafee and Erik Brynjolfsson, Machine, Platform, Crowd: Harnessing Our Digital Future (W W Norton & Company 2017) <https://books.google.com.hk/books?hl=en&lr=&id=zh1DDQAAQBAJ&oi=fnd&pg=PA1905&dq=Machine> accessed 5 December 2024.

[345] Sam Capper, 'Cyber Criminals: Being Anonymous Online' (Darkinvader.io2024) <https://www.darkinvader.io/blogs/cyber-criminals-being-anonymous-online> accessed 5 December 2024.

[346] Eva Stepney and Clare Lally, 'Disinformation: Sources, Spread and Impact Overview' (2024) <https://researchbriefings.files.parliament.uk/documents/POST-PN-0719/POST-PN-0719.pdf> accessed 2 December 2024.

[347] Cho and Zhu (n 343) para 18.

[348] David Stark and Ivana Pais, 'Algorithmic Management in the Platform Economy' (2020) 14 Sociologica 47 <https://sociologica.unibo.it/article/view/12221>.

[349] Keller (n 331) 3.

by AI models[350], making them liable for ensuring that harmful or false content does not proliferate uncontrollably. Online platforms leverage their technological capabilities and managerial advantages to control data on their platforms, establish their content moderation standards, and even enable some large online platforms to achieve monopoly status[351]. Given these reasons, this thesis argues that placing governance liabilities on online platforms is both necessary and feasible to address the challenges posed by AI-generated disinformation.

A comparative study of platform content moderation liability in the EU, the US, and China is feasible due to the global governance and different legal philosophies of the three jurisdictions[352]. The EU places the protection of citizens' rights at the core of digital regulation, and the DSA's provisions reflect differentiated management of intermediary service providers, such as requiring VLOPs to assume greater liabilities than small or medium-sized platforms[353]. The Cybersecurity Law and the Provisions on the Governance of the Online Information Content Ecosystem[354] emphasize the obligations of online platforms to review and process disinformation, particularly regarding illegal information and disinformation that threatens national security and social stability[355]. The Provisions on Administration of Algorithmic Recommendation in the Internet Information Service[356] require online platforms with huge social

---

[350] Louis Regnier, 'AI Platform Launches in UK to Combat Harmful Content Online | Startups Magazine' (Startups Magazine2025) <https://startupsmagazine.co.uk/index.php/article-ai-platform-launches-uk-combat-harmful-content-online> accessed 13 May 2025.

[351] Yankun Yang, 'Attribution of Liability for Copyright Infringement by Artificial Intelligence Generated Content' (2023) 29 Lecture Notes in Education Psychology and Public Media 115.

[352] Eyal Zilberman, 'Platform Liability Regimes around the World | Heinrich-Böll-Stiftung | Tel Aviv - Israel' (Heinrich-Böll-Stiftung | Tel Aviv - Israel2022) <https://il.boell.org/en/2023/03/30/platform-liability-regimes-around-world> accessed 31 May 2025.

[353] Christoph Busch, 'Platform Responsibility in the European Union' [2025] Cambridge University Press eBooks 20 <https://www.cambridge.org/core/books/defeating-disinformation/platform-responsibility-in-the-european-union/AA3D55C57B0F6A7C18F5CAEF25146557 >.

[354] Provisions on the Governance of the Online Information Content Ecosystem (网络信息内容生态治理规定) (promulgated 15 December 2019, effective 1 March 2020).

[355] Jufang Wang, 'Platform Responsibility with Chinese Characteristics' in Bhaskar Chakravorti and Joel P Trachtman (eds), Defeating Disinformation: Digital Platform Responsibility, Regulation and Content Moderation on the Global Technological Commons (Cambridge University Press 2025) 41.

[356] Provisions on Administration of Algorithmic Recommendation in the Internet Information Service (《互联网信息服务算法推荐管理规定》, 2021, China), art 17.

influence to improve online information review mechanisms, conduct security assessments, and avoid using algorithms to promote the spread of disinformation. While the US's Section 230 of the Communications Decency Act provides broad immunity to online platforms for most user-generated content, except for federal criminal offenses, intellectual property infringement, electronic privacy violations, and other exemptions provided by law, it emphasizes the protection of free speech and of business interests[357]. Comparing the differences in legal regulations across these three jurisdictions demonstrates the scope of regulatory liabilities[358] borne by online platforms, the availability of platform immunity, and the challenges faced in practical implementation. Moreover, this field also benefits from an increasing number of case studies, including real-life cases on AI-generated disinformation, judicial decisions addressing online platform liability, and administrative enforcement actions against platforms that fail to remove or block harmful posted information.

3.1.2 Private Law and Platform Liability

Private law plays a crucial role in governing the liability of online platforms by providing legal remedies and regulatory mechanisms for individuals and businesses when they are deceived or misled by AI-created disinformation. It primarily governs relationships between private entities, including users, online platforms, and third-party content creators. When online platforms facilitate or fail to prevent the generation and dissemination of false content by AI models, they may have liabilities that arise from tort law.

Tort law provides the normative basis of establishing duties of care, including the liability to moderate content, especially where platforms are aware of the dissemination

[357] Aifang MA, 'Digital Legislation: Convergence or Divergence of Models? A Comparative Look at the European Union, China and the United States' (2024) <https://server.www.robert-schuman.eu/storage/en/doc/questions-d-europe/qe-769-en.pdf> accessed 31 May 2025.

[358] Daphne Keller, 'Systemic Duties of Care and Intermediary Liability' (Stanford CIS29 May 2020) <https://cyberlaw.stanford.edu/blog/2020/05/systemic-duties-care-and-intermediary-liability/> accessed 5 December 2024.

of harmful or illegal information[359]. When information published by an online platform constitutes infringement or causes economic losses to users, the users can sue in court based on the rights protection of tort law, requiring the platforms to assume responsibility for infringement and make appropriate compensation.

The regulations of online platform liabilities reveal divergent legal approaches across three selected jurisdictions. In the EU, platform liability under the tort law is predominantly based on fault liability. Besides, the new Product Liability Directive 2024/2853[360] also extends the definition of "product" to include digitally or AI-generated goods or content[361], stating that liability for defectiveness in relation to such products is not limited only to manufacturers, but the online platforms should also be liable in certain circumstances. China's tort liability regime, which is governed by Articles 1194 and 1195 of the Civil Code[362], imposes more proactive obligation on online platforms to take timely measures to remove or block disinformation or illegal content when it is detected, and to assume conditional liability for the portion of the content they review and host that infringes upon the rights of users[363]. These provisions reflect the duty of care requirements for internet service providers and the regulatory framework that places primary responsibility for content moderation[364]. In contrast, US's Section 230 of the CDA clarifies that platforms are not publishers of third-party content, thereby largely exempting platforms from liability arising from third-party content and significantly narrowing the scope of claims based on negligence and

---

[359] Kashish Shamsi, 'Dangerous Social Media Trends: Can Social Media Platforms Be Held Liable?' (Brooklaw.edu2025) <https://sports-entertainment.brooklaw.edu/media/dangerous-social-media-trends-can-social-media-platforms-be-held-liable/?utm_.com> accessed 6 May 2025.

[360] Directive (EU) 2024/2853 of the European Parliament and of the Council of 24 April 2024 on liability for defective products [2024] OJ L202/1.

[361] Christoph von Burgsdorff and Luisa Kramer, 'Increased Liability due to the New EU Product Liability Directive: What Does This Mean for the Medical and Pharmaceutical Industry?' (Ibanet.org2025) <https://www.ibanet.org/increased-liability-eu-product-directive> accessed 9 April 2025.

[362] Civil Code of the People's Republic of China (《中华人民共和国民法典》) (promulgated 28 May 2020, effective 1 January 2021).

[363] Wang (n 355) 71.

[364] Guideline on further pressuring websites and platforms to fulfil the primary responsibilities for content governance (promulgated by the Cyberspace Administration Sept. 15, 2021)<http://www.cac.gov.cn/2021-09/15/c_1633296790051342.htm>accessed 6 May 2025.

defamation[365]. However, online platforms are not exempt from the obligation to review content in all circumstances: if there is a binding agreement between the online platform and users[366] (such as Terms of Service), its content may stipulate that the online platform has the obligation to review content and delete disinformation.

Online platforms are interpreted as entities that could directly prevent harms through timely intervention, rather than as neutral intermediaries that just provide information exchange platforms[367]. Their ongoing governance process is shaped by socioeconomic structure, technical design, and regulatory framework[368], and the constant intervention in users' social activities prevents platforms from being truly eutral intermediaries[369].

The Principle of European Tort Law(PETL)[370] does not explicitly address the liability of internet platforms[371], while its general principle could be used to assess such liabilities, providing a solid framework for understanding liability in the European context. It regulates the situations in which individuals should be liable, such as causing errors intentionally or negligently(Article 4:101), violating the obligation of protecting one specific party(Article 4:103), failing to exercise due diligence with foreseeable damage(Article 5:101), or causing damages by violation of requirements for auxiliaries within the scope of their duties(Article 6:102)[372].

The new EU Product Liability Directive 2024/2853[373] was proposed in October 2024, replacing the previous Product Liability Directive 85/374/EEC[374] from nearly 40 years

---

365 Jeff Kosseff, The Twenty-Six Words That Created the Internet (Cornell University Press 2019).
366 Barnes v Yahoo! Inc (2009) 570 F3d 1096 (9th Cir) 7413, 7435.
367 Erik Valgaeren and Cyril Fischer , 'Online Platforms and Uploading of Protected Works: A Priori No Liability for Operators of Online Platforms' (Stibbe15 July 2021) <https://www.stibbe.com/publications-and-insights/online-platforms-and-uploading-of-protected-works-a-priori-no-liability> accessed 13 May 2025.
368 José Van Dijck and Thomas Poell, 'Understanding Social Media Logic' (2013) 1 Media and Communication 2.
369 Tarleton Gillespie, 'Platforms Intervene' (2015) 1 Social Media + Society 1, 2.
370 European Group on Tort Law, Principles of European Tort Law (Springer 2005).
371 Piotr Machnikowski, 'The Principles of European Tort Law and Product Liability' (2024) 15 Journal of European Tort Law 31.
372 European Group on Tort Law, Principles of European Tort Law: Text and Commentary (Springer 2005) arts 4:101, 4:103, 5:101, 6:102.
373 Directive (EU) 2024/2853 of the European Parliament and of the Council of 13 March 2024 on liability for defective products [2024] OJ L.
374 Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products [1985] OJ

ago. The purpose of this new Directive is to impose more specific requirements on product liability to keep pace with advanced AI technologies and globalized digital platform services[375]. The new Directive significantly expanded the scope of the definition of "product", including digital products, such as the AI system (Article 4(2))[376]. Therefore, independent AI systems and AI-integrated products fall within the scope of application of this Directive, while digital data not classified as software is not considered as a product unless it is in the form of a digital manufacturing file[377]. The directive requires a product to be deemed defective if its characteristics, foreseeable use, potential impact, specific needs of the targeted audience, and safety requirements fail to meet the expectation or legally required safety standard(Article 7(2))[378]. Manufacturers of the defective product could be held liable for unintended "harmful behavior" of such products, and this Directive extends liability to authorized representatives, fulfillment service providers, and distributors(Article 8 (3)-(4))[379]. According to the DSA, the online platform shall be held liable in the same way as a distributor under this regulation if it facilitates the particular transaction in question, by displaying its product or otherwise, in such a way as to lead the average consumer to believe that the product has been supplied by the online platform itself or by a merchant acting under its authorization or control, where the role of the online platform is not limited to that of a neutral intermediary(Recital 38)[380].

In China, Article 1195 of the Civil Code[381] explicitly regulates that if malicious users exploit online platforms to commit an infringement (such as uploading copyright-infringing articles on the platform), the right holder can inform online platforms to take

---

L210/29.

[375] Dr Lena Niehoff and others, 'New Product Liability Directive 2024/2853: New Product Liability Risks for Products in the EU' (Taylorwessing.com6 January 2025) <https://www.taylorwessing.com/en/insights-and-events/insights/2025/01/di-new-product-liability-directive> accessed 7 April 2025.

[376] Directive 2024/2853, art 4(2).

[377] Directive 2024/2853, Recitals 13 and 16.

[378] Directive 2024/2853, art 7(2).

[379] Directive 2024/2853, art 8.

[380] DSA, recital 38.

[381] Civil Code of the PRC, art 1195

necessary measures, such as deletion, blocking, or disconnection, to prevent the dissemination of harmful content. Upon receipt of the notification, network service providers have a liability to forward the notification to relevant users and intervene to deal with false content that has preliminarily constituted infringement. Article 1197[382] states that online platforms that fail to take appropriate measures will be jointly and severally liable with infringing users, emphasizing the obligations on platforms to proactively handle illegal content and moderate information.

In the US, Section 230 of the Communications Decency Act[383] provides broad immunity to online platforms for user-generated disinformation or defamation unless they have contributions to the content's creation or development. For example, in Doe v. MySpace[384], the appeals court pointed out that "MySpace", as an information platform, should not be regarded as the publisher of the content generated by its illegal users. The criminal merely exploited the information exchange function of an online platform to establish contact and communicate with the victims offline, so the platform should not bear the liability for this role[385]. However, user agreements impose reasonable rights and obligations on online platforms, and provide for content review within a certain scope with the consent of both parties to the contract.[386]. Terms of Service(ToS) are legally binding agreements between online platforms and users and are widely used by large online platforms, such as Facebook[387], to outline the rules, obligations, and rights of both parties. Many terms of service would regulate their content policies, including the discretion to remove or moderate the content that violates these policies and block the accounts that post harmful or inappropriate information.

---

382 Civil Code of the PRC, art 1197.
383 CDA § 230.
384 Doe v MySpace Inc, 528 F 3d 413 (5th Cir 2008).
385 StudyBounty, 'Doe v. MySpace: The Case That Changed Social Media' (StudyBounty22 July 2022) <https://studybounty.com/doe-v-myspace-the-case-that-changed-social-media-research-paper> accessed 10 April 2025.
386 Andrew Jankowich, 'EULAw: The Complex Web of Corporate Rule-Making in Virtual Worlds' (2019) 8 Tulane Journal of Technology & Intellectual Property <https://journals.tulane.edu/TIP/article/view/2500> accessed 13 May 2025.
387 Nicolas Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4 Social Media + Society.

3.2 Evolving Legal Foundations of Platform Liability in the Digital Age

3.2.1 The Gatekeeping Theory for Shifting Supervisory Liability for the Online Platforms from Neutral Intermediary to Gatekeeper

As digital technology continues to develop, theories about the liability of online platforms in content review have also evolved. The liability for information supervision of internet platforms has also gone through many stages of transformation[388] from passive management to active censorship.

Gatekeeping theory, first proposed by Kurt Lewin[389], was used to illustrate that gatekeepers can filter out what people deem undesirable according to certain conditions or criteria. For example, housewives could decide the family's habits by controlling what food should be served on the table. David Manning explained gatekeeping as the process of filtering countless messages into limited information delivered to people, introducing the concept of the "gatekeeper" through a case study of a wire editor, "Mr. Gates" [390]. This example analyzes how this editor chose the new stories to be published from the magazine's coverage, revealing the impacts of personal bias, professional norms, and practical constraints on the selection process[391], demonstrating that diverse subjective characteristics of gatekeepers may have a profound effect on the filter criteria of information flow[392].

Content moderation is commonly defined as the process of selecting and evaluating user-generated content posted to websites, social media sites, or other online platforms[393]. As high-level decision-makers who manage the flow of information

---

[388] Joanne van Eennnaam, 'The New Platform Liability: From the E-Commerce Directive to the Digital Services Act Regulation (" DSA ")' (WiseMen Advocaten2023) <https://www.wisemen.nl/en/news/the-new-platform-liability-from-the-e-commerce-directive-to-the-digital-services-act-regulation-dsa-/> accessed 13 May 2025.

[389] Ali Aidroos Albara, 'The Concept of Gatekeeping in Information Science: A Philosophical Reflection' (2018) 8 Global Journal of Information Technology: Emerging Technologies 16.

[390] David Manning White, 'The "Gate Keeper": A Case Study in the Selection of News' (1950) 27 Journalism Quarterly 383.

[391] Gregory Perreault, 'Gatekeeping' [2022] The SAGE Encyclopedia of Journalism.

[392] Gavin Davie, 'Gatekeeping Theory' (Mass Communication Theory2 November 2018) <https://masscommtheory.com/theory-overviews/gatekeeping-theory/> accessed 13 May 2025.

[393] Tarleton Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (Yale University Press 2018), ch 7.

throughout the social system, traditional media usually bear the liability of filtering and excluding the collected information to ensure the content they publish is authentic, reliable, and appealing, in order to maintain the authority and credibility of their own publication. Tushman and Katz[394] expanded the role of the gatekeeper, expanding how the gatekeepers, as intermediaries[395] between clients and products, handle the transfer and communication of information internally and externally. They demonstrate that gatekeepers not only interpret and assimilate the required information from external sources but also facilitate and communicate valuable information to internal members or users[396].

Traditional gatekeeping theory usually considers the following factors when determining whether an entity should be held accountable as a "gatekeeper". Firstly, the gatekeeper needs to have appropriate capacity to deter the publication and dissemination of illegal content, financial ability to detect it with reasonable costs, and sufficient incentives to regulate violations[397]. At the same time, it is also necessary to consider whether the subject that is given the regulatory duty has control over illegal or harmful information[398] and is a more effective approach than imposing practical penalties that may fail to deter serious misconduct[399].

Modernized gatekeeping theory[400] recognizes that online platforms, especially those providing core intermediary services, as powerful gatekeepers that use algorithms to

---

394 R Katz and M Tushman, 'An Investigation into the Managerial Roles and Career Paths of Gatekeepers and Project Supervisors in a Major R & D Facility' (1981) 11 R&D Management 103.
395 Michelle Striepe and Christine Cunningham, 'Gatekeepers, Guides, and Ghosts: Intermediaries Impacting Access to Schools during COVID-19' (2022) 17 Ethnography and Education 1.
396 Thomas Wesley Allen, 'Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within the R&D Organization' (1984) 1 RePEc: Research Papers in Economics.
397 Reinier H Kraakman, 'Gatekeepers: The Anatomy of a Third-Party Enforcement Strategy' (1986) 2 The Journal of Law, Economics, and Organization <https://academic.oup.com/jleo/article/2/1/53/873299> 53, 56.
398 Ibid 55.
399 Jonathan Zittrain, 'A History of Online Gatekeeping' (2006) 19 Harvard Journal of Law & Technology 253, 256.
400 Karine Barzilai-Nahon, 'Toward a Theory of Network Gatekeeping: A Framework for Exploring Information Control' (2008) 59 Journal of the American Society for Information Science and Technology 1493.

amplify, curate, and select the content[401]. In the digital age, even though most online platforms, such as Facebook or X, do not directly produce content, their business approach is to attract worldwide internet users to post, talk, and broadcast to each other, guiding and facilitating users to engage in the production of content actively[402]. As the primary intermediaries in the dissemination of online content, online platforms are often better positioned than governments or law enforcement authorities to manage and regulate the flow of information. Moreover, with the development of information auditing technologies, the online platforms have technical and cost advantages over administrative agencies[403] in content supervision. Therefore, the selected jurisdictions at the legislative level are gradually increasing the online platforms' information moderation regulatory obligation[404]. For example, the EU has recognized the important role that Internet Service Providers (especially Very Large Online Platforms) play in legal enforcement and has directly assigned them the liability of detecting, filtering, and removing specific types of illegal content[405]. The vast amount of personal data collected and processed by online platforms gives them insight into interests, preferences, and needs. Such platforms' knowledge of user preferences links their choices to personalized services and advertisements[406], aggregating and providing this customized content for advertisers who produced commercialized promotional content centered around audiences[407]. Therefore, as the development of online platforms transform data into a primary resource, data mining could foster digital companies to

---

[401] Philip M Napoli, 'Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers' (2014) 39 SSRN Electronic Journal 756.
[402] Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (papers.ssrn.com9 September 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3038939> accessed 4 August 2025.
[403] Nadejda Komendantova and Dmitry Erokhin, 'Artificial Intelligence Tools in Misinformation Management during Natural Disasters' (2025) 25 Public Organization Review.
[404] Jiansheng Xie, Research on Information Supervision Responsibility of Internet Platform (MA thesis, Anhui University of Technology 2024) https://doi.org/10.27790/d.cnki.gahgy.2022.000316 accessed 11 April 2025.
[405] Marco Bassini, 'Fundamental Rights and Private Enforcement in the Digital Age' (2019) 25 European Law Journal 182.
[406] Daniele Archibugi, Andrea Filippetti and Marion Frenz, 'The Impact of the Economic Crisis on Innovation: Evidence from Europe' (2013) 80 Technological Forecasting and Social Change 1247.
[407] James Frederick Hamilton, Robert Bodle and Ezequiel Korin, Explorations in Critical Studies of Advertising (Routledge, Taylor & Francis Group 2019), ch 3.

develop predictive and real-time analytics techniques[408]. This position of control enables them to shape information flows, lead the competition in the marketplace, and master strong economic power[409]. E-commerce platforms, as third-party intermediaries with a natural incentive[410] use a variety of resources to mitigate situations that block transactions, enabling them not only to act as bridges between different user groups but also to play a governance role in interpreting the codes of conduct necessary for the functioning of the online market[411]. While it also leads to malicious users taking advantage of online platforms to spread disinformation more efficiently, thus creating a huge accountability for platform managers. In this context, online information platforms have adopted ground rules for content moderation, being liable for regulating posted information and proposing practical solutions for disinformation detection and warnings. Targeted obligations that have been imposed on them could mitigate the abuse of intermediary power and foster fair competition. For example, one of the regulatory approaches used by Facebook is to flag the controversial posts and provide a way for users to "flag" the information whose authenticity they have suspected[412]. In this case, the control over the communities and a willingness to self-regulate enables online platforms to become gatekeepers who access consumers' data, curate and monitor user-generated information, and provide services[413].

The European Commission has clarified the term "gatekeeper" in the Digital Market Act, referring to technology companies that are able to provide core platform services(CPS) that have an important influence on the EU's internal markets[414]. These

---

[408] José van, Platform Mechanisms (Oxford University Press 2018) 31, 48.
[409] Christian Fuchs, Social Media: A Critical Introduction (SAGE Publications Ltd 2014).
[410] Molly Cohent and Arun Sundararajantt, 'Self-Regulation and Innovation in the Peer-To-Peer Sharing Economy' (2015) <https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1039&context=uclrev_online>.
[411] Chuanman You, 'Law and Policy of Platform Economy in China' (2020) 39 Computer Law & Security Review 105493, 105505.
[412] Mark Wilson, 'Study: Facebook's Fake News Labels Have a Fatal Flaw' (Fast Company4 March 2020) <https://www.fastcompany.com/90471349/study-facebooks-fake-news-labels-have-a-fatal-flaw> accessed 22 April 2025.
[413] Jacques Crémer, Yves-Alexandre De Montjoye and Heike Schweitzer, 'Digital Era a Report by Competition Policy Competition' (2019) <https://euagenda.eu/upload/publications/untitled-257961-ea.pdf> accessed 8 August 2025.
[414] Grant Thornton Ireland, 'Determining Gatekeepers under the Digital Markets Act' (Grant Thornton

digital companies operate platform services that are used by a large number of users or enterprises and present a significant gateway for business users to reach end-users, maintaining or could be expected to have solid market positions[415]. The EU's Digital Market Act[416] has specifically designated certain online platforms as "gatekeepers" and imposed clear obligations on them. Article 3 of DMA states the conditions for a business to be designated as a gatekeeper, including a specified turnover, the number of monthly or annual end-user active users, the business's ability to access and collect personal information or analyze it, and the business's group structure[417]. Besides, Article 5 regulates the directly applicable obligations of gatekeepers, which include protecting users' rights to talk directly to customers outside the platform, the right to uninstall the pre-installed apps, and the right not to have their personal data merged across platforms without their consent[418]. Furthermore, Article 6 sets out further specifications about the requirements of core platform services, such as prohibiting gatekeepers from favoring their own services or goods, or requiring interoperability with third-party software or services[419]. Article 7 emphasize the importance of ensuring gatekeepers to make their messaging services interoperable with those offered by third-party providers[420]. This requirement could reduce user "lock-in" to a single platform and multiple their choices, promoting the platform's innovation while balancing technical feasibility and user's safety[421]. To ensure and demonstrate compliance with the obligations, Article 8 requires the gatekeepers to implement effective measures in

---

Ireland22 October 2024) <https://www.grantthornton.ie/insights/factsheets/determining-gatekeepers-under-the-digital-markets-act/> accessed 4 April 2025.

415 'The Role of the Gatekeepers under the DMA Regulation' (Consent Management Platform (CMP) Usercentrics2023) <https://usercentrics.com/knowledge-hub/role-of-gatekeepers-under-digital-markets-act/> accessed 18 August 2025.

416 Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022] OJ L265/1.

417 DMA, art 3.

418 DMA, art 5.

419 DMA, art 6.

420 DMA, art 7.

421 European Commission, 'The Digital Markets Act: Ensuring Fair and Open Digital Markets' (European Commission2022) <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en>.

achieving the objectives of this regulation and associated obligations[422]. The regulations of DMA control the power structure that allows or prevents the dissemination of certain information, indirectly affecting the gatekeeping role played by online platforms by limiting self-preference and ensuring data access.

DMA's use of the term "gatekeeper" reflects the fact that online platforms control access to the market and audiences in communication and media studies. DMA recognized that online platforms are powerful intermediaries with considerable influence and capacity to manage information available in the public domain. This "gatekeeping" theory strengthens the regulatory liability of online platforms in content supervision: as the main body of network information auditing, online platforms shall play an active role as gatekeepers in the process of content dissemination and immediately remove harmful information that violates legislation. However, in practice, even if the platform's Terms of Service includes explicit rules prohibiting harassment, hate speech, and other forms of harmful content, the inconsistency and opacity of the platform's enforcement mechanisms have led to a serious imbalance between over-enforcement and under-enforcement in some cases[423]. Additionally, to maximize user engagement and advertising revenue[424], online platforms may allow provocative or marginal content to remain accessible, particularly when such information generates significant user interactions[425]. The online platform's business model determines that it needs to attract more advertisers to place advertisements online so it can make a profit, and personalized advertising encourages the platform to provide engaging content[426].

---

[422] DMA, art 8.

[423] Nicolas P Suzor, Lawless: The Secret Rules That Govern Our Digital Lives (Cambridge University Press 2019) <https://www.cambridge.org/core/books/lawless/8504E4EC8A74E539D701A04D3EE8D8DE>, ch 3.

[424] Jack M Balkin, 'Fixing Social Media's Grand Bargain' (papers.ssrn.com15 October 2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3266942> accessed 4 August 2025.

[425] Jeff Gary and Ashkan Soltani, 'First Things First: Online Advertising Practices and Their Effects on Platform Speech' (knightcolumbia.org2019) <https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech> accessed 13 September 2025.

[426] Alice E Marwick, 'Why Do People Share Fake News? A Sociotechnical Model of Media Effects' (Georgetown Law Technology Review21 July 2018) <https://georgetownlawtechreview.org/why-do-people-share-fake-news-a-sociotechnical-model-of-media-effects/GLTR-07-2018/> accessed 13 September 2025.

Once users are attracted to share, discuss and like controversial information, online platforms will predict consumers' potential consumption tendencies and push different advertisements directly to the targeted audience without hiding or deleting such information[427].

Besides, even though the online platform acts as a powerful gatekeeper in moderating content, its enforcement heavily relies on user reports and is subject to detection resource constraints[428]. Online platforms as gatekeepers cannot be limited to self-regulation; they lack the accountability, transparency[429], and procedural fairness[430] that should be provided for in the law, which requires the establishment of a sound supervisory framework at the legislative level.

3.2.2 Early Liability Regulation of Online Platforms (from the 1990s to early 2000s)

In the early years of the internet development(1990s to early 2000s), legislators in various jurisdictions took a hands-off approach to digital regulation[431]. It was widely recognized that the internet is a space for innovation and free expression, and premature regulation could stifle economic and technological development. In the early days of the Internet's development, libertarian social activist John Perry Barlow[432] envisioned the idealized governance of online communities as a space that would not be externally regulated and create its own internal legal governing bodies. While the influence of this declaration is diminishing, in practice, cyberspace remains largely unregulated[433]. Therefore, early online platforms were regarded not as publishers or editorial entities,

---

[427] Louise Matsakis, 'Facebook's Targeted Ads Are More Complex than It Lets On' (Wired25 April 2018) <https://www.wired.com/story/facebooks-targeted-ads-are-more-complex-than-it-lets-on/> accessed 13 September 2025, para 7.
[428] Ibid para 12.
[429] Frank Pasquale, 'The Black Box Society: The Secret Algorithms That Control Money and Information' (2016) 45 Contemporary Sociology: a Journal of Reviews 367.
[430] Natasha Tusikov, Chokepoints : Global Private Regulation on the Internet (Oakland, California University of California Press 2017).
[431] John F Blevins, 'The Use and Abuse of "Light-Touch" Internet Regulation' (2018) 99 SSRN Electronic Journal 177, 226.
[432] John Perry Barlow, 'A Declaration of the Independence of Cyberspace' (Electronic Frontier Foundation8 February 1996) <https://www.eff.org/cyberspace-independence> accessed 13 September 2025.
[433] Suzor (n 423) ch 4.

but as passive channels that merely facilitated content exchange for users. These internet intermediaries, such as forums, searching engines, or social networks, were viewed more like message transfer entities[434], which provide infrastructure but do not curate content. The legal systems at that time generally favored exempting platforms from responsibility for the accuracy or even legality of third-party posted content, fostering digital businesses to develop[435]. Although the underlying objectives were similar across selected jurisdictions, there are significant differences in the legal instruments of the US, the EU, and China.

In the US, Section 230 of CDA, which includes two provisions in subsection(c), has established a legal shield for online platforms. Under Section 230(c)(1), this provision states that "service providers should not be regarded as the publisher or creator of any information posted by another information content provider", shielding platforms from being held liable for third-party harmful information publishers[436]. Section 230(c)(2) provides protection for platforms that choose to delete or restrict access to content that they consider "obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable[437]", even if these speeches are protected under constitutional free speech, as long as they act in good faith in doing so[438]. The Section 512 of Digital Millennium Copyright Act[439](DMCA) has also regulated the safe harbor provision, protecting online service providers that are not engaged in illegal activities from monetary liability for copyright infringement based on the conduct of their users in exchange for cooperating with copyright owners to remove infringing content

---

434 Technology Director, 'From Immunity to Regulation: Turning Point of Internet Intermediary Regulatory Agenda' (The Journal of Law and Technology at Texas8 October 2016) <https://jolttx.com/2016/10/08/immunity-regulation-turning-point-internet-intermediary-regulatory-agenda/> accessed 15 April 2025.

435 Corynne McSherry, 'User Generated Content and the Fediverse: A Legal Primer' (Electronic Frontier Foundation20 December 2022) <https://www.eff.org/deeplinks/2022/12/user-generated-content-and-fediverse-legal-primer> accessed 15 April 2025.

436 Communications Decency Act of 1996, 47 USC § 230(c)(1).

437 Communications Decency Act of 1996, 47 USC § 230(c)(2).

438 Barbara Ortutay, 'What You Should Know about Section 230, the Rule That Shaped Today's Internet' (PBS NewsHour21 February 2023) <https://www.pbs.org/newshour/politics/what-you-should-know-about-section-230-the-rule-that-shaped-todays-internet>.

439 Digital Millennium Copyright Act, 17 USC § 512 (1998)

immediately and satisfy certain conditions. This protection is available only if the online platforms lack actual prior knowledge of infringing activity and promptly adopt and implement actions to remove or disable access to such content after being notified[440]. Moreover, if the platform enables control of the infringing activities, it is not allowed to derive a direct financial benefit from infringement(including infringement appeals to paying customers). Besides, platforms are required to designate an agent to receive takedown notices and implement feasible measures to deal with repeat infringers. This legal framework aims to balance the protection of copyright holders' rights with platforms' business models and technological innovation[441], ensuring that platforms will not be held liable for unlawful content created by third parties, as long as they act responsibly after receiving notifications.

The judgment of Zeran v. America Online Inc.[442] solidified Section 230's liability, ensuring platforms to integrate user-created content without fear of prosecution. The court emphasized that the amount of information dissemination through interactive computer services is astounding, so it is impossible for service providers to filter the potential problems involved in millions of postings. The possibility of being held liable for every message republished by the service platform would lead to a chilling effect because it would naturally incentivize service providers to remove messages that some users find offensive to avoid liability, resulting in a severe restriction on the number and type of messages posted.[443]. Despite ongoing criticism about legal immunity, such as in Malwarebytes, Inc. v. Enigma Software Group USA, LLC[444], Justice Clarence Thomas suggests the liability immunity should be narrowed or eliminated in future cases; section 230 still remains a cornerstone of US internet governance.

---

[440] Kevin J Hickey, 'Digital Millennium Copyright Act (DMCA) Safe Harbor Provisions for Online Service Providers: A Legal Overview' (Congress.gov2025) <https://www.congress.gov/crs-product/IF11478>.

[441] Ibid para1.

[442] Zeran v America Online Inc, 129 F 3d 327 (4th Cir 1997), para 11.

[443] Matt Stroud, 'These Six Lawsuits Shaped the Internet' (The Verge19 August 2014) <https://www.theverge.com/2014/8/19/6044679/the-six-lawsuits-that-shaped-the-internet>.

[444] Malwarebytes Inc v Enigma Software Group USA LLC, 946 F 3d 1040 (9th Cir 2020).

In the EU, the E-Commerce Directive (Directive 2000/31/EC)[445] is the central pillar of the regulatory framework for digital services at the EU level and contains the EU's provisions on conditional immunity for online platforms, which establish the minimum standards of liability for internet intermediaries. The legislation introduce the "safe harbor" regime for three types of intermediaries: mere conduits(Article12), cashing services(Art.13), and hosting providers(Art.14). Under the principle, platform operators are not liable for information stored or hosted by third-parties, subject to two alternative conditions[446]. Firstly, the providers are manifestly unaware of illegal activities or content[447]; secondly, the providers have taken effective actions to delete or disable access to the illegal content or disinformation upon being aware of such circumstances[448]. Unlike Section 230 of the US's CDA, which grants broad immunity to publishers, the E-Commerce Directive exempts online platforms from liability only if they do not aware the illegal content or disinformation[449]. In L'Oréal v. eBay[450] case, CJEU stated that application of the "safe harbor" provision needs to be seen in the context of the role played by the intermediary, and the exemption from liability does not apply to a service provider that plays an active role, which would aware or control over the information they hosted. Besides, Article 15 of this legislation regulates that member states should not impose a general obligation on information service providers to monitor the information they transmitted or stored[451]. However, the prohibition only

---

[445] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L178/1.

[446] Tambiama Madiega, 'Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act | Think Tank | European Parliament' (Europa.eu2020) <https://www.europarl.europa.eu/thinktank/en/document/%20EPRS_IDA(2020)649404> accessed 16 April 2025.

[447] See ECD, supra note 1, art. 14(1)(a).

[448] See ECD, supra note 1, art. 14(1)(b).

[449] Christoph Busch, 'Platform Responsibility in the European Union' [2025] Cambridge University Press eBooks 20 < https://www.cambridge.org/core/books/defeating-disinformation/platform-responsibility-in-the-european-union/AA3D55C57B0F6A7C18F5CAEF25146557?.com >.

[450] L'Oréal SA v eBay International AG (C-324/09) [2011] ECR I-06011, para 116.

[451] Piper UK, 'EU Study on the Legal Analysis of a Single Market for the Information Society - Publications Office of the EU' (Publications Office of the EU2023) <https://op.europa.eu/en/publication-detail/-/publication/a856513e-ddd9-45e2-b3f1-6c9a0ea6c722> accessed 16 April 2025,5.

refer to monitoring of the general nature, does not concern monitoring obligations in alleged illegal activities(such as the specific duties of care imposed by national law[452]) or in the case that authorities request them to provide information that identifies the recipient of the services with whom the storage agreements are concluded. But since this Directive does not specify the scope of responsibility for the duties of care and does not apply to any types of illegal behaviors[453], there is a lack of a clear boundary between the duties of care[454] and general monitoring.

In China, early online platform regulatory liabilities were based on reactive response(such as handling reports after they were received), taking responsibility only for "known" or "should have known" illegal content, and not requiring platforms to actively review all the posted information. In terms of content moderation, platforms only legally cracked down on obscenity, copyright infringement, and other illegal information, while with fewer requirements to regulate disinformation or misleading content. The Regulations on Protection of the Information Transmission Rights on Internet[455], which were proposed in 2006, have introduced the "safe harbor" principle to protect copyright. The platform should delete or remove the infringing content after receiving the notice from the copyright holder, otherwise, it will bear joint and several liability. It emphasizes that network operators are not liable for infringement caused by user-generated content if they have removed harmful information after receiving notification. In Tort Liability Law, which was proposed in 2009[456], Article 36 expanded the circumstances in which platforms are jointly and severally liable, extending from

---

452 Ibid.

453 Aleksandra Kuczerawy, 'To Monitor or Not to Monitor? The Uncertain Future of Article 15 of the E-Commerce Directive' (CITIP Blog10 July 2019) <https://www.law.kuleuven.be/citip/blog/to-monitor-or-not-to-monitor-the-uncertain-future-of-article-15-of-the-e-commerce-directive/> accessed 5 December 2024.

454 Peggy Valcke, Aleksandra Kuczerawy and Pieter-Jan Ombelet, 'Did the Romans Get It Right? What Delfi, Google, eBay, and UPC Tele Kabel Wien Have in Common' (2017) 31 Law, Governance and Technology Series 101.

455 Regulations on the Protection of the Right of Communication through Information Networks (信息网络传播权保护条例) (promulgated by the State Council, 18 May 2006, effective 1 July 2006) (China).

456 Tort Liability Law of the People's Republic of China (Standing Committee of the National People's Congress, adopted 26 December 2009, effective 1 July 2010).

content that infringes copyright to any information that infringes the civil rights of others.

These early legal frameworks show how different political and legal systems responded to the platform liability issue in the digital age. Now all three jurisdictions are reevaluating the platform immunity in light of the advent of AI models[457], widespread of disinformation, and algorithmic amplification, which shift explored in the following research.

3.3 Current Legal Frameworks on Content Platform Liability

This section examines the existing legal frameworks across the EU, the US, and China that establish online platforms' responsibility for the content they host and store, including the disinformation created by AI. It highlights how liability and accountability have evolved from traditional areas of law, such as tort law, and how these developments have been affected by modern challenges of disinformation.

Online platforms have the ability to engage in regulation[458], in particular through direct access to users' data and the deployment of algorithms that instantly remove or flag the disinformation within seconds after detecting it[459].

3.3.1 The European Union: From the E-Commerce Directive to the Digital Services Act

The EU's construction of the social consensus on the governance of disinformation needs to be explained by the EU's series of actions. One of the major reasons why disinformation has attracted worldwide attention from the EU and the world is due to the massive amount of disinformation posted on online platforms during the US

[457] Peter Henderson, Tatsunori Hashimoto and Mark Lemley, 'Where's the Liability in Harmful AI Speech?' (arXiv.org2023) <https://arxiv.org/abs/2308.04635> accessed 17 April 2025.
[458] Christoph Busch, 'Self-Regulation and Regulatory Intermediation in the Platform Economy' (Ssrn.com30 November 2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3309293> accessed 22 April 2025.
[459] Michèle Finck, 'Digital Co-Regulation: Designing a Supranational Legal Framework for the Platform Economy' (papers.ssrn.com20 June 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2990043> accessed 4 August 2025.

presidential election in 2016[460]. How to prevent disinformation from manipulating political activities has become a key issue of concern to governments. In this context, the European Commission has taken practical measures to address the growing threat of online disinformation, combining legal instruments and binding regulatory frameworks. One of the foundational policy documents in this area is the 2018 Communication titled "Tackling Online Disinformation: A European Approach" (COM(2018) 236 final)[461], analyzing the definition, scope, and main cause of online disinformation and highlighting the necessity of effective self-monitoring by platforms. Besides, it seeks to enact practical solutions to reduce the dissemination of disinformation by strengthening self-censorship, improving monitoring techniques, promoting online accountability, and enhancing public media literacy. Building on this strategy, the European Commission promoted the adoption of the "2018 Code of Practice on Disinformation"[462], a self-regulatory code of practice voluntarily endorsed by major online platforms, advertisers, and industry participants. This Code provides a structured framework and sets key benchmarks for platform liability[463], which is an important step in coordinating different stakeholders to develop a uniform standard for combating disinformation. Besides, this Code outlines commitments on multiple fronts, including the censorship of ad placement and transparency around issue-based advertising, strengthening the integrity of safeguards, empowering users and fact-checkers, and undermining the monetization of disinformation. It also led to concrete actions and policy changes by relevant stakeholders to help combat disinformation, serving as the foundation for assessing DSA compliance. While the Code has been very

---

460 Robert Faris and others, 'Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election' (papers.ssrn.com1 August 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3019414> accessed 4 August 2025.

461 European Commission, 'Tackling Online Disinformation: a European Approach' COM (2018) 236 final.

462 European Commission, Code of Practice on Disinformation (2018) <https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation> accessed 10 July 2025.

463 Ramsha Jahangir, 'The EU's Code of Practice on Disinformation Is Now Part of the Digital Services Act. What Does It Mean?' (Tech Policy Press25 February 2025) <https://www.techpolicy.press/the-eus-code-of-practice-on-disinformation-is-now-part-of-the-digital-services-act-what-does-it-mean/> accessed 10 July 2025.

helpful in monitoring and ensuring accountability for the actions of signatories, it has been criticized for lacking a strong enforcement mechanism and inconsistent implementation across platforms[464]. In response, the Commission issued the "2022 Strengthened Code of Practice on Disinformation", which has been included in the framework of the Digital Service Act as a Code of Conduct on Disinformation. This document introduced firmer co-regulatory commitments, adding a number of specific measures and commitments in terms of content review mechanisms, empowering users and fact-checkers, and establishing an external permanent group[465].

E-Commerce Directive(ECD) lays the foundation for the EU's regulatory framework for digital services, stipulating the standards for determining liability immunity for different types of online service providers. The ECD's safe harbor provisions grant platforms significant discretion to establish inconsistent and non-transparent content policies that are applied in ways that sometimes allow for removing lawful content or speech, and often without adequate due process protection[466]. However, this directive only provides limited guidance on the circumstances in which platforms should undertake content censorship, lacks recognition of very large online platforms' influences, and becomes outdated in the face of rapidly evolving technologies and the use of generative AI.

In 2022, the Digital Service Act(DSA)[467] was introduced to modernize and replace the ECD, illustrating the existing rules aimed at regulating the liability of online platforms that provide digital services. The DSA applies to intermediary service providers (including online social networks and online platforms[468]), requiring them to clarify

---

[464] Stephan Mündges and Kirsty Park, 'But Did They Really? Platforms' Compliance with the Code of Practice on Disinformation in Review' (2024) 13 Internet Policy Review <https://policyreview.info/articles/analysis/platforms-compliance-code-of-practice-on-disinformation-review>.

[465] Brooke Tanner, 'EU Code of Practice on Disinformation' (Brookings5 August 2022) <https://www.brookings.edu/articles/eu-code-of-practice-on-disinformation/>.

[466] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (Harvard Law Review10 April 2018) 1598 <https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/> accessed 4 August 2025.

[467] Digital Services Act, Regulation (EU) 2022/2065.

[468] DLA Piper, 'What Is the Digital Services Act? | DLA Piper' (DLA Piper2023) <https://www.dlapiper.com/en/insights/publications/2023/06/what-is-the-digital-services-act> accessed

and disclose the conditions and rules of their content moderation, explain content audit decisions to users, and take proactive and effective actions following notification[469]. In particular, it has set out additional rules for "very large online platforms," forcing them to grant users the right to opt out of recommendation systems and analytics, share key data with researchers and authorities, cooperate with crisis response requests, and undergo external and independent audits[470]. With regard to platform liability, the DSA retains the conditional immunity structure from liability for online platforms, but imposes enhanced regulatory obligations. Compared to ECD, the DSA covered the same categories of intermediaries under the protection of "safe harbor"(Articles 4-6), the enforcement is strengthened through the "notice-and-action" mechanism while maintaining the same exemption conditions for hosting services. Articles 15 and 16 establish a standard process that allows any person or entity to report online platform content that they believe to be illegal, imposing specific requirements for both intermediary service providers and hosting service providers[471]. Firstly, intermediary service providers have liability to publicly post a clear and accessible report in the readable format, of any content censorship that has undertaken during the given period, including its classification and the efficiency with which it has carried out the orders of the Member States' competent authorities, the criteria for classifying content for auditing, and the measures taken to manage it[472]. Secondly, they are legally obliged to act promptly to remove or restrict access to the disinformation after they receive actual knowledge of existing harmful content. About the outcomes of notifications, the intermediary service providers are required to record and disclose in their transparent reports about the number of complaints received through their internal complaint handling systems, as well as the decision, basis and time taken for the eventual handling

---

4 June 2025.

[469] Christoph Busch, 'Regulating the Expanding Content Moderation Universe: A European Perspective on Infrastructure Moderation' (2022) 27 UCLA Journal of Law & Technology.

[470] Emma Roth, 'The EU's Digital Services Act Goes into Effect Today: Here's What That Means' (The Verge25 August 2023) <https://www.theverge.com/23845672/eu-digital-services-act-explained>.

[471] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, arts 15-16.

[472] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 15.

of the complaint, in accordance with their terms and conditions. In the case of hosting service providers who enable to moderation of content, they have been required to have an accessible notice mechanism, providing a user-friendly tool that enables users, social organizations, or public institutions to submit reports[473]. Besides, they should cooperate proactively with notifiers by ensuring and facilitating the submission of relevant information, containing a reasonable explanation of why the content constitutes an offence, an indication of the precise location of illegal information (such as exact URL), the notifier's identity and email address, and a statement confirming the accuracy and completeness of information provided[474].

In the DSA, transparency obligations are not limited to a single aspect of platform operation. Rather, the DSA introduces a multilayer transparency regime that could be broadly categorized into three types, each of which plays a crucial role in improving the regulations of platforms and addressing the challenges posed by AI-generated disinformation. First, Articles 14, 15, and 24 that regulate the content review transparency, focusing on how platforms manage illegal or harmful content[475]. Articles 14 and 15 impose a general obligation on all providers of intermediary services, including hosting service providers, to publish annual transparency reports. These reports should contain clear and detailed information about content moderation activities, such as information about the complaint handling process[476]. Under Article 24[477], platforms are required to publish regular transparency reports detailing their content review practices, such as the amount of content removed, numbers and outcomes of out-of-court dispute settlements, and the reasons and amounts of suspensions. In addition, when a platform restricts or removes user content, it must

---

[473] Theresa Ehlen, 'The Digital Services Act: New Liability Rules?' (Passle10 July 2023) <https://technologyquotient.freshfields.com/post/102iiyf/the-digital-services-act-new-liability-rules> accessed 25 April 2025.

[474] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 16.

[475] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, arts 14, 15 and 24.

[476] European Commission, 'Digital Services Act: Commission Launches Transparency Database | Shaping Europe's Digital Future' (digital-strategy.ec.europa.eu26 September 2023) <https://digital-strategy.ec.europa.eu/en/news/digital-services-act-commission-launches-transparency-database>.

[477] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 24.

provide a clear statement of reasons, including the applicable rules or legal provisions. This transparency ensures that platforms are held liable for their vetting activities, prevents arbitrary decision-making, and enables regulators to monitor how platforms deal with the growing amount of AI-generated disinformation that often evades traditional vetting methods[478].

Second, transparency in the operation of systems targets the internal mechanisms that shape users' online experiences, in particular recommendation systems and advertising practices. Articles 26 and 27 of DSA have strengthened users' rights to transparency and control over the platform's information distribution mechanism by introducing disclosure requirements on the advertising transparency and operation logic of recommendation systems. Article 26[479] has imposed specific transparency obligations on platforms' advertising and recommender systems to address the key role of algorithm-driven content distribution and advertisements in information manipulation and the spread of disinformation. By forcing platforms to disclose algorithmic logic and the basis of pushing advertisements, the DSA hopes to break "black box" control of information flow by online platforms and provide users with greater rights to information and choice, thereby increasing the credibility and accessibility of the entire information environment. Under requirements of Article 26, online platforms should ensure that commercial content has prominent markings to identified as advertisements on the user's interface, informing users about the identity of advertisers, the rules for setting parameters for advertisements to be recommended to specific consumers(e.g., based on users' interests, behaviors, geographic location, etc). Such requirements raise the visibility and accountability of ad pushes and prevent covert manipulation of users, especially in the area of political advertising and public opinion. Secondly, for recommendation systems commonly used on platforms, such as search result ranking,

---

[478] Alessia Zornetta, 'Is the Digital Services Act Truly a Transparency Machine?' (Tech Policy Press11 July 2024) <https://www.techpolicy.press/is-the-digital-services-act-truly-a-transparency-machine/>.
[479] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 26.

Article 27[480] requires platforms to disclose the main functional logic of these systems. They must clearly explain to users how recommendation systems personalize the sorting or pushing of content based on factors such as user behaviors, preferences, and historical searching records. More importantly, platforms need to provide multiple options that allow service recipients to choose and modify the relative order in which information is presented at any time, so they are not limited to the recommended options of profiling and thus avoid being trapped in an information cocoon by the algorithm.[481] Third, transparency requirements for external security establish that VLOPs should grant access to certain data to vetted researchers and authorities, as well as the obligation to undergo an annual independent audit at the external institution at their own expense. Article 37 obliges the VLOPs and VLOSEs to undergo audits at least once a year in order to assess their compliance with their obligations under DSA, in particular with regard efficiency of data review and deletion, transparency reporting, and a compliant dispute resolution process. The audit conducted by qualified and objective auditors should prepare a comprehensive report on the effectiveness of the risk assessment, content review, and provide reasonable mitigation measures, thereby platforms are obliged to address the deficiencies following the issues raised by this paper. Under Article 42[482], these platforms are required to publish a comprehensive transparency report every six months, in at least one of the official languages of member states, detailing their content auditing measures, the use of automated tools, and actions they have taken to address systemic risks such as disinformation. Their reports improved the public supervision of platforms for the fulfillment of their obligations under GDPR, particularly with regard to addressing challenges posed by harmful content.

The DSA also introduced the obligations for Very Large Online Platforms(VLOPs),

---

[480] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 27.
[481] Luca Nannini and others, 'Beyond Phase-In: Assessing Impacts on Disinformation of the EU Digital Services Act' [2024] AI and Ethics.
[482] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 42.

mainly focusing on obligations around systemic risk assessment and mitigation, and how these provisions are important to regulating disinformation, including AI-generated disinformation. By mandating these platforms to proactively identify the potential risks, regulators expect platforms to take preventative measures before problems occur, rather than just reacting to damages after their occurrence. The VLOPs mentioned in DSA apply to online platforms or online search engines that have an average monthly number of active service recipients within the Alliance equal to or greater than 45 million[483](Article 33), such as Facebook, X, and Instagram. Article 34 regulates that platforms(especially very large online platforms) and Very Large Online Search Engines (VLOSEs) must identify, analyze, and assess any systemic risks arising from the functioning or design of their services and their related systems[484]. These risks include the dissemination of illegal information, potential infringements of fundamental rights(such as freedom of expression, data protection, or non-discrimination and the rights of child), as well as risks arising from the intentional manipulation of platform' services, such as the proliferation of dis/misinformation(including disinformation generated by AI), the inauthenticity of service use(like manipulation by phishing account), and spread of harmful or misleading information. In addition, platforms are also required to be concerned about the potential negative impact caused by system risks on the space of public disclosure and the electoral process[485]. The purpose of mandatory risk assessment is to prompt platforms to design and implement appropriate risk mitigation measures, strengthening platforms' responsibility and accountability mechanisms, while at the same time protecting users' fundamental rights and maintaining a safe digital environment. Also, the risk assessment reports could serve as an important basis for regulators to monitor and review the platform's compliance with their liabilities, reducing the probability and severity of risks. DSA's Article 35

[483] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 33.
[484] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 34.
[485] Sally Broughton Micova and Andrea Calef, 'Elements for Effective Systemic Risk Assessment under the DSA' [2023] SSRN Electronic Journal.

elaborates the requirements in Article 34 for VLOPs to take feasible and proportionate measures to mitigate identified risks. Platforms should follow certain standards when developing mitigation measures, such as adjusting news recommendation algorithms to reduce the visibility of harmful content, strengthening content review mechanisms, improving user reporting and appeal processes, and seeking cooperation with independent fact-checkers to limit information abuse[486]. Besides, platforms should consider the principle of risk prevention when designing new features or updating services to reduce the probability of risks at source. At the same time, platforms should regularly adjust and review their mitigation strategies to ensure these measures can be updated according to the development of risks. Article 35 complements the risk assessment obligations of Article 34 by requiring platforms not only to detect risks, but also to take effective actions to control and prohibit them, and collectively to build a more comprehensive regulatory framework for VLOPs under the DSA system. These provisions are particularly important in addressing disinformation posed by generative AI models.

3.3.2 The United States: Section 230 and the Limits of Platform Immunity

The United States currently lacks a unified federal regulatory framework that imposes specific content moderation obligations on online platforms, including with regard to AI-generated disinformation. Instead, the current landscape is shaped by a combination of federal regulations, state-level initiatives, and ongoing debates about the balance between free speech and addressing harmful or illegal information.

At the federal level, the US provides online platforms with a wide range of exemptions from content auditing liability, centered on Section 230[487] of the Communications Decency Act(CDA). Internet service providers' exemptions could always be asserted as a First Amendment defense, but Section 230 significantly complements the First

[486] European Commission, 'The Digital Services Act Package | Shaping Europe's Digital Future' (digital-strategy.ec.europa.eu2022) <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>.

[487] Communications Decency Act, 47 USC § 230 (1996).

Amendment defense. At its core, section 230 provides that interactive computer service providers(including online platforms) should not be treated as the publisher or speaker of the third-party content created by their users. This statute means online platforms are not legally liable for the content posted by internet users unless statutory exceptions apply, such as federal criminal prosecutions or intellectual property claims[488]. Besides, Section 230 also allows platforms to review the posted information and remove harmful or infringing posts without being punished. This allows online platforms to develop and enforce their own information governance rules in a large context, with relatively limited government and mainly relying on platform self-regulation. Section 230 immunity provides additional legal comfort to Internet services as an effective remedial procedure that can help empower courts to dismiss claims at the earliest possible stage of litigation[489], avoiding the need for costly litigation for internet service providers. While exemption clauses would undermine the willingness of online platforms to proactively censor harmful or illegal content, which might cause users to decrease their trust in platforms[490]. Compared to the First Amendment, Section 230 is not limited to regulating government conduct, but extends legal immunity to private companies and grants platforms a moderate content censorship, rather than providing absolute protection from government restrictions on any speech[491]. However, Section 230 of CDA has granted platforms broad legal immunity to user-generated content, which leads to insufficient liability for the spread of harmful content [492] (including

---

[488] Elizabeth Banker, 'A Review of Section 230'S Meaning & Application Based on More than 500 Cases' (Archiveia.org2020) <https://archiveia.org/publications/a-review-of-section-230s-meaning-application-based-on-more-than-500-cases/> accessed 29 April 2025.

[489] Kelly O'Hara, 'What Is Section 230 and Why Should I Care? - Internet Society' (Internet Society24 February 2023) <https://www.internetsociety.org/blog/2023/02/what-is-section-230-and-why-should-i-care-about-it/?gad_source=1&gad_campaignid=138051697&gbraid=0AAAADqyrA_4bQLTagzsVLldz6FPqNC8o&gclid=Cj0KCQjwzrzABhD8ARIsANlSWNPuD-ha2mduJQKarP5lEoWDCMu38Dw0stTXMsSsR_hrdriwEB7YHnsaAui_EALw_wcB> accessed 29 April 2025.

[490] National Academies of Sciences, Engineering, and Medicine, Section 230 Protections: Can Legal Revisions or Novel Technologies Limit Online Misinformation and Abuse? (National Academies Press 2021).

[491] Eric Goldman, 'Why Section 230 Is Better than the First Amendment' (2019) 95 SSRN Electronic Journal.

[492] Michael Brown, 'Rethinking Section 230: Fostering Transparency, Accountability, and User

disinformation, hate speech, or harassment content) and a lack of strong external regulatory mechanisms. While platforms have the right to remove content, section 230 has no request for platforms to provide transparent criteria or redress mechanisms for their censorship practices[493], which triggers platforms to act in a way that their content review actions are neither restricted by governmental constraints nor challenged by users[494].

In the absence of comprehensive federal legislation, several states have enacted laws to regulate disinformation, particularly during election cycles, that could significantly mislead or distress the public. California has enacted several new bills to curb the spread of disinformation and deceptive election content[495]. These bills, AB 730, AB 2013, and AB 2355, seek to strengthen protections against digitally manipulated media in political activities and advertising, as well as enable consumers to be fully informed when using generative AI systems. Assembly Bill 730 (2019)[496] prohibits publishing audio or video media of a political candidate that is materially deceptive within 60 days of the election, such as posts that injure the candidate's reputation or deceive voters, unless a clear disclaimer is provided. It targets manipulative media to address disinformation during elections and sets a time-bound injunction to protect the integrity of campaigning. Similarly, Assembly Bill 2013 [497](2024) requires the developers of generative AI models to disclose specific information about the datasets used to train their models. This bill applies to all generative AI systems and services, or significantly

---

Protection Online – Denver Journal of International Law & Policy' (Djilp.org17 November 2024) <https://djilp.org/rethinking-section-230-fostering-transparency-accountability-and-user-protection-online/?utm_.com> accessed 29 April 2025.

[493] Richard Stengel, 'Revoking the Law That Protects Twitter Could Backfire on Trump' (Vanity FairJune 2020) <https://www.vanityfair.com/news/2020/06/revoking-the-law-that-protects-twitter-could-backfire-on-trump?srsltid=AfmBOoq1vOrVHagyPDsW-qT6mz0pNCK6XWApw6N4cAWEz5hhN5CwKuSS> accessed 28 May 2025.

[494] Anna Wiener, 'Trump, Twitter, Facebook, and the Future of Online Speech' (The New Yorker6 July 2020) <https://www.newyorker.com/news/letter-from-silicon-valley/trump-twitter-facebook-and-the-future-of-online-speech> accessed 28 May 2025.

[495] Stuart D Levi and others, 'California Enacts New Laws to Combat AI-Generated Deceptive Election Content' (Skadden.com27 September 2024) <https://www.skadden.com/insights/publications/2024/09/california-enacts-new-laws>.

[496] California Assembly Bill No 730, Ch 493 (2019).

[497] California Assembly Bill No 2013, Ch 817 (2024).

modified versions of generative AI systems or services, released on or after January 1, 2022, and made available to California residents, with or without a fee. It intends to increase transparency of content generation systems, making the source and composition[498] of the data behind the AI models clearer to users. Assembly Bill No. 2355[499](2024) aims to enhance transparency in political advertising by addressing the use of AI in creating or altering content. The bill requires that any political advertisement containing images, audio, or video generated or substantially modified using AI tools, and the content of which is likely to mislead a reasonable person as to its authenticity, must provide clear and conspicuous disclosures. To ensure the effectiveness of these disclosures, the bill establishes specific formatting requirements based on media of advertisements, including readability of disclosed text, font size, as well as clarity and timing of audio disclosure. Minnesota Statute § 211B.075[500], enacted in 2023, criminalizes the intentional dissemination of disinformation or deepfakes that misrepresent the candidate's speech or conduct, although its enforcement has raised significant First Amendment concerns[501]. Meanwhile, New Mexico's House Bill 182[502] (HB 182) requires that any political advertisement or campaign material utilizing AI or deepfake technologies should include a clear and conspicuous disclaimer to indicate that this content is created or processed by AI tools. This bill is introduced to avoid the potential to distort political discourse by generating disinformation that could confuse or mislead voters[503].

However, some states, such as Texas, have concerns about unfair censorship of

---

[498] Marc S Martin and Sydney Veatch, 'AB 2013: New California AI Law Mandates Disclosure of GenAI Training Data | Perkins Coie' (Perkinscoie.com7 October 2024) <https://perkinscoie.com/insights/update/ab-2013-new-california-ai-law-mandates-disclosure-genai-training-data>.

[499] California Assembly Bill No 2355, Ch 260 (2024).

[500] Minnesota Stat § 211B.075 (2023).

[501] STEVE KARNOWSKI, 'Elon Musk's X Sues to Overturn Minnesota Political Deepfakes Ban' (ABC News25 April 2025) <https://abcnews.go.com/Technology/wireStory/elon-musks-sues-overturn-minnesota-political-deepfakes-ban-121173206>.

[502] New Mexico House Bill 182, 56th Leg, 1st Sess (2024).

[503] Jeremy Werner, 'New Mexico Enacts Law Requiring Disclosure of AI-Generated Content in Political Campaign Ads' (BABL AI27 August 2024) <https://babl.ai/new-mexico-enacts-law-requiring-disclosure-of-ai-generated-content-in-political-campaign-ads/?utm_.com> accessed 30 April 2025.

controversial speech[504]. Texas House Bill 20 (HB 20) prohibits large platforms from censoring users, user-generated content, or users' accounts based on viewpoint and provides users with legal remedies for bringing suits against platforms that violate this bill, restricting large online platforms from moderating AI-generated disinformation that spreads political expressions or parodies legitimate speech[505]. Also, this bill requires platforms to publicly publish an acceptable use policy and semi-annual transparency reports that clearly explain the review rules and actions. Similarly, Florida has also introduced or passed bills aimed at restricting monopolistic bias in content moderation, such as Florida's Senate Bill 7072[506]. This Bill seeks to regulate how large online platforms regulate and manage content, preventing censorship that is deemed arbitrary by the Florida legislature and promoting transparency and fairness in content review practices[507]. It has faced constitutional challenges and has been subject to litigation regarding its compatibility with federal law and free speech protections[508]. The growing interest of state governments in generative AI and intervention in its use highlights the collective concern of people about the potential impact of AI[509]. However, in the US, the lack of federal-level legal regulations to moderate content, and the level regulatory approaches are diverse and sometimes conflicting, lead to fragmented management of AI-generated disinformation[510].

While there is no overarching federal-level law specifically regulating platforms'

---

[504] Mark Joseph Stern, 'The 5th Circuit's Reinstatement of Texas' Internet Censorship Law Could Break Social Media' (Slate Magazine12 May 2022) <https://slate.com/technology/2022/05/texas-internet-censorship-social-media-first-amendment-fifth-circuit.html>.

[505] Mackenzie Cerwick, 'Censoring Social Media: Texas HB 20' (Vanderbilt University2021) <https://www.vanderbilt.edu/jetlaw/2021/10/06/censoring-social-media-texas-hb-20/>.

[506] Florida Senate, Senate Bill 7072, Social Media Platforms, Reg Sess 2021 (enacted 24 May 2021) https://www.flsenate.gov/Session/Bill/2021/7072 accessed 29 April 2025.

[507] Tech Policy Press, 'Florida Social Media Platforms Bill - SB.7072 | TechPolicy.Press' (Tech Policy Press) <https://www.techpolicy.press/tracker/florida-social-media-platforms-bill-sb-7072/>.

[508] Alan Rozenshtein, 'The Real Takeaway from the Enjoining of the Florida Social Media Law' (Lawfare2021) <https://www.lawfaremedia.org/article/real-takeaway-enjoining-florida-social-media-law> accessed 19 October 2024.

[509] Simon Chesterman, 'Lawful but Awful: Evolving Legislative Responses to Address Online Misinformation, Disinformation, and Mal-Information in the Age of Generative AI' (arXiv.org2025) <https://arxiv.org/abs/2505.15067?utm.com> accessed 28 May 2025.

[510] Tammy Whitehouse, 'How AI Governance Can Adapt to a Fragmented Regulatory Landscape' (WSJ2 December 2024) <https://deloitte.wsj.com/cfo/how-ai-governance-can-adapt-to-a-fragmented-regulatory-landscape-23e47f94>.

content moderation liability and addressing AI-generated disinformation, federal agencies have taken enforcement actions to combat specific disinformation. One notable example is the Federal Communications Commission(FCC) taking action against the use of AI-generated fake robot calls to deceive voters[511]. In early 2024, the FCC regulated Lingo Telecom by issuing a $1 million fine and asking this company to stop transmitting suspicious information for its role in making AI-created robocalls impersonating US President Joe Biden, which sent disinformation, and these calls were intended to suppress turnout during the New Hampshire Democratic[512]. These automated calls used AI voice cloning technology, raising serious concerns about election manipulation and public trust[513]. In 2024, the FCC issued the AI-generated voices in robot calls as the "automatic telephone dialing system, or an artificial or prerecorded voice," which is required to be prohibited from being used for malicious or immoral purposes by the Telephone Consumer Protection Act (TCPA)[514]. This declaratory ruling has expanded the FCC's scope of enforcement capabilities, enabling greater regulation of entities that use AI tools in their communications.

Also, the FCC has taken action against autodialed scams, including enforcement against auto warranty fraud autodialed scams[515], and warnings about student loan debt scam robocalls or robot-texts[516], and has achieved significant results (the number of the scams

---

[511] David Shepardson, 'Lingo Telecom Agrees to $1 Million Fine over AI-Generated Biden Robocalls' Reuters (21 August 2024) <https://www.reuters.com/technology/artificial-intelligence/lingo-telecom-agrees-1-million-fine-over-ai-generated-biden-robocalls-2024-08-21/> accessed 13 September 2025.

[512] Federal Communications Commission, 'FCC Settles Spoofed AI-Generated Robocalls Case' (Fcc.gov21 August 2024) <https://www.fcc.gov/document/fcc-settles-spoofed-ai-generated-robocalls-case> accessed 13 September 2025.

[513] Federal Communications Commission , 'FCC SETTLES CASE against PROVIDER THAT TRANSMITTED SPOOFED AI-GENERATED ROBOCALLS for ELECTION INTERFERENCE in NEW HAMPSHIRE Lingo Telecom to Pay $1 Million Civil Penalty and Implement First-of-Their-Kind Compliance Terms Secured by the FCC' (2024) <https://docs.fcc.gov/public/attachments/DOC-404951A1.pdf> accessed 1 May 2025.

[514] Nivedha Soundappan and Pantho Sayed, 'FCC Cracks down on AI-Powered Robocalls' (Harvard Journal of Law & Technology23 February 2024) <https://jolt.law.harvard.edu/digest/fcc-cracks-down-on-ai-powered-robocalls> accessed 13 September 2025.

[515] Brian Fung, 'FCC Issues Historic $300 Million Fine against the Largest Robocall Scam It Has Ever Investigated' (CNN4 August 2023) <https://www.cnn.com/2023/08/04/tech/fcc-robocall-scam-biggest-fine/index.html> accessed 13 September 2025.

[516] Federal Communications Commission, 'FCC & State AGs Warn of Student Loan Debt Scam Robocalls & Robotexts' (2023) <https://www.fcc.gov/document/fcc-state-ags-warn-student-loan-debt-scam-robocalls-robotexts> accessed 1 May 2025.

dropped by 80%[517]). This action demonstrates that federal-level enforcement agency is directly targeting the use of generative AI models in disinformation campaigns as a form of wire fraud and shows that regulators are increasingly focused on curbing AI-driven disruptive activities. In addition, the FCC proposed new transparency requirements mandating the disclosure of the AI-generated content in political ads that run on radio or television[518]. The Notice of Proposed Rulemaking(NPRM)[519] was released to mandate that Broadcasters, Cable Operators, DBS Providers, and SDARS Licensees disclose the use of generative AI to create content, enhancing the election transparency and preventing voter deception.

Imposing liability on online platforms to regulate disinformation, particularly AI-generated disinformation, presents a range of legal and technical challenges. The US's First Amendment strongly protects the freedom of speech, including most forms of false information, unless it falls within narrowly defined exceptions such as defamation or incitement to commit an immediate unlawful act, which may indeed result in the commission of an offence(Brandenburg v. Ohio[520])[521]. This constitutional protection makes it difficult to require platforms to remove or censor false information, even if it is patently false or harmful. Besides, platforms' broad immunity from liability granted by Section 230 of the CDA has significantly restricted their legal responsibility to moderate disinformation unless they actively participate in its creation. The deepfakes and AI-generated disinformation are usually disseminated anonymously or via

---

[517] Karl Bode, 'FCC, State Action Nets an Amazing 80% Reduction in Auto Warranty Scam Robocalls' (Techdirt25 August 2022) <https://www.techdirt.com/2022/08/25/fcc-state-action-nets-an-amazing-80-reduction-in-auto-warranty-scam-robocalls/> accessed 1 May 2025.

[518] Federal Communications Commission, 'Disclosure and Transparency of Artificial Intelligence-Generated Content in Political Advertisements' (Federal Register5 August 2024) <https://www.federalregister.gov/documents/2024/08/05/2024-16977/disclosure-and-transparency-of-artificial-intelligence-generated-content-in-political-advertisements>.

[519] Federal Communications Commission, 'Federal Communications Commission FCC 24-74 NOTICE of PROPOSED RULEMAKING' (2024) <https://docs.fcc.gov/public/attachments/FCC-24-74A1.pdf> accessed 1 May 2025.

[520] Brandenburg v Ohio, 395 US 444 (1969).

[521] Henry Cohen, 'CRS Report for Congress Freedom of Speech and Press: Exceptions to the First Amendment' (2009) <https://resources.saylor.org/wwwresources/archived/site/wp-content/uploads/2014/01/POLSC401-3.1-FreedomofSpeechandPress-PublicDomain.pdf>.

automated accounts and lack traceability and clear intent[522], making it challenging to attribute liability to responsible parties.

3.3.3 China: Multi-layered regulatory framework and special requirements for AI-generated disinformation

China's platform regulatory framework represents a co-regulatory governance model centered on platform interventions, establishing market-oriented industry standards[523]. The legal regulatory framework for content monitoring in China rests on foundational laws, including the Cybersecurity Law(2017), Data Security Law(2021), and Personal Information Protection Law(2021)[524]. These laws establish the obligations of platform operators, containing the duty to monitor and manage user-generated content to prevent the creation and spread of illegal or harmful information. However, these selected laws have different emphases on platform liability: Cybersecurity Law explicitly requires network operators(including online platforms) to proactively manage content generated by users and take monitoring measures such as deletion of unlawful information, as well as passively responding to reports or regulatory requests from users or public; DSL focuses on protection of data security rather than direct content management, so there is no direct provision requiring platforms to monitor information posted by users unless the content involves illegal data processing; PIPL concentrates on reactive responses related to personal information(such as handling complaints and enforcing owner's right), with proactive obligations limited to risk management(such as preventing obviously illegal processing of private data) rather than comprehensive content review. Article 47 of the Cybersecurity Law[525] has imposed information

---

[522] Anqi Shao, 'Beyond Misinformation: A Conceptual Framework for Studying AI Hallucinations in (Science) Communication' (arXiv.org2025) <https://arxiv.org/abs/2504.13777> accessed 2 May 2025.
[523] You (n 411) 12.
[524] Cybersecurity Law of the People's Republic of China (adopted 7 November 2016, effective 1 June 2017) Order No 53 of the President of the PRC; Data Security Law of the People's Republic of China (adopted 10 June 2021, effective 1 September 2021) Order No 84 of the President of the PRC; Personal Information Protection Law of the People's Republic of China (adopted 20 August 2021, effective 1 November 2021) Order No 91 of the President of the PRC.
[525] Cybersecurity Law, art 47.

moderation obligations on network operators, but only with relation to content that is prohibited by laws or administrative regulations from being published or transmitted, requiring operators to stop spreading and remove illegal content from platforms immediately upon discovery of such information. Article 49 mandates that platforms establish a digital information security system and requires them to handle the relevant reports and complaints from users effectively[526]. Besides, Article 48 stipulates that platforms are obliged to provide technical support and information processing assistance to national security authorities in the investigation of criminal activities, including the relevant moderation of posted content[527]. The Data Security Law[528] requires platforms to adopt appropriate technical measures to safeguard data security through conducting data processing activities, so as to ensure that the data is under effective protection and lawful use(Article 27). Under this requirement of data safety protection, if unlawful or disinformation involves data misuse, such as forging data or illegal scraping of information, the online platforms are obliged to delete or block the flows of relevant data[529]. Articles 29 and 30 regulate that the processors of important data are required to carry out risk assessments of their data analysis activities regularly and to take immediate remedial measures if deficiencies, loopholes, and other risks to data are identified, highlighting the platform's duty to supervise such data[530]. The Personal Information Protection Law[531] stipulates that platforms that enable access to and processing of personal information are obliged to correct, supplement, or remove the content posted on public platforms at the request of their owners(Articles 44-47)[532], and to handle complaints of private data effectively(Article 50)[533]. Besides, Article 57

---

[526] Cybersecurity Law, art 49.
[527] Cybersecurity Law, art 48.
[528] Data Security Law, art 27.
[529] Barbara Li, 'IAPP' (Iapp.org2024) <https://iapp.org/news/a/china-issues-the-regulations-on-network-data-security-management-what-s-important-to-know>.
[530] Data Security Law, arts 29-30.
[531] Personal Information Protection Law of the People's Republic of China (adopted 20 August 2021, effective 1 November 2021) Order No 91 of the President of the PRC.
[532] Personal Information Protection Law, arts 44-47.
[533] Personal Information Protection Law, art 50.

states that platforms need to remedy incidents of personal information leakage, tampering, and loss (such as deleting leaked information)[534].

At the same time, several Chinese departmental regulations and normative documents impose general obligations on content auditing as well as specific provisions on false information. Article 8 of the "Provisions on the Governance of the Online Information Content Ecosystem" explicitly states that online content platforms should fulfil their main responsibility for information management, and Article 10 clarifies that disinformation belongs to the category of inaccurate information, requiring platforms to prevent and resist the dissemination of such information[535]. The "Interim Measures for the Management of Generative Artificial Intelligence Services"[536] introduced special provisions for disinformation created by generative AI: Article 4 imposes an obligation on generative AI service providers to be responsible for created content, requiring them not to take advantage of algorithms and data to generate disinformation by infringing on the legitimate rights and interests of others, and to provide security audits of the services they provide following the requirements of articles 17 and 18[537]. Immediately stop generating and deleting content when disinformation is found, and report it to the regulatory authorities. In addition, Articles 8 and 12 of the Measures require platforms to clearly and accurately label images, videos, and other content generated using AI technology, and timely dispose of illegal content[538].

This comparative study demonstrates that broad legal and political traditions deeply shape the platform regulations of AI-generated disinformation. The EU's approach offers a balanced and comprehensive framework, blending user protection, platform accountability, and requirements of algorithmic transparency. Risk assessment obligations, transparency duties, and independent auditing requirements create a robust

---

534 Personal Information Protection Law, art 57.

535 Provisions on the Governance of the Online Information Content Ecosystem, arts 8 and 10.

536 Interim Measures for the Management of Generative Artificial Intelligence Services (生成式人工智能服务管理暂行办法) (promulgated 10 July 2023, effective 15 August 2023).

537 Interim Measures for the Management of Generative Artificial Intelligence Services, arts 4,17, and 18.

538 Interim Measures for the Management of Generative Artificial Intelligence Services, arts 8 and 12.

framework for mitigating systemic risks posed by generative AI. The US, grounded in the First Amendment protections and Section 230 of the CDA, takes a market-driven and expression-protective approach. While this approach fosters the technical innovation and protection of free speech, it limits governmental authority to impose mandatory content moderation, making it difficult to require platforms to proactively manage AI-generated disinformation. China follows a model centered on national governance, with internet service providers as bridges. The existing legal regulations require online platforms to monitor and remove harmful or prohibited content, including disinformation generated by AI tools. However, the criteria of harmful information are vague and overly broad, and platforms may over-censor content to comply with strict laws and regulations and avoid being penalized, creating a chilling effect on freedom of expression[539].

3.3.4 Changing Patterns of Legal provisions on the Online Platform Regulatory Liability

This section examines the evolution of legal provisions on platform liability in the EU, the US, and China, identifying their respective patterns of legal change. It also assesses the impact of changes in the regulatory framework on platforms and explains the underlying reasons for these patterns of change.

The EU's regulatory evolution shows a clear pattern of progressive codification, moving from the principle of limited intermediary liability under the E-Commerce Directive to a comprehensive liability framework under the DSA[540].

The E-commerce Directive provides a safe harbor for online intermediaries[541], shielding them from liability for illegal information transmitted through their services, as long as they promptly remove the information or disable access upon learning of the

[539] Larry Diamond and Orville Schell, China's Influence and American Interests (Hoover Press 2019).
[540] Martin Husovec, 'Introduction to Liability Framework', Principles of the Digital Services Act (Oxford University Press 2024) <https://academic.oup.com/book/58088> accessed 29 November 2024.
[541] Maria Lillà Montagnani and Alina Trapova, 'New Obligations for Internet Intermediaries in the Digital Single Market—Safe Harbors in Turmoil?' (2019) 22 Joural of Internet Law.

illegal activity[542]. This “notice and takedown” system institutionalizes the passive liability of online platforms to regulate disinformation[543], imposing mandatory regulations to ensure the management of disinformation. As platforms shift their role to become gatekeepers to the flow of online information, this Directive's provisions on platform liability have shown their limitations[544]. For example, Article 15[545] prohibits imposing general monitoring obligations that restrict platforms' rights to proactively review content, giving them broad discretion to address harmful but lawful disinformation[546]. The Code of Practice on Disinformation and the Strengthened Code of Practice on Disinformation encourage platforms to sign and voluntarily adhere to these codes, thereby reducing the spread of disinformation[547]. These codes complement the E-commerce Directive's requirements for platform liability, but as soft law, they can only be implemented voluntarily by online platforms and are not legally binding[548]. The DSA is based on the requirements of the E-Commerce Directive, retaining its platform liability immunity while introducing due diligence requirements proportionate to the platform's size and social influence, primarily targeting VLOPs[549]. Furthermore, the DSA formally incorporates the voluntary Code of Conduct on Disinformation into its regulatory framework, requiring VLOPs to anticipate and mitigate disinformation that poses systemic risks through mandatory laws[550].

---

[542] Directive 2000/31/EC, art 14.
[543] Gergely Gosztonyi, ‘The Contribution of the Court of Justice of the European Union to a Better Understanding the Liability and Monitoring Issues Regarding Intermediary Service Providers’ (2020) 59 Annales Universitatis Scientiarum Budapestinensis De Rolando Eötvös Nominatae. Sectio Iuridica 133.
[544] Toygar Hasan Oruç, ‘The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in Light of Recent Developments: Is It Still Necessary to Maintain It?’ (2022) 13 JIPITEC–Journal of Intellectual Property, Information Technology and E-Commerce Law 176 <https://www.jipitec.eu/jipitec/article/view/354> accessed 19 October 2025.
[545] Directive 2000/31/EC, art 15.
[546] Kuczerawy (n 453).
[547] Emmanouil Papadogiannakis and others, ‘Before & After: The Effect of EU’s 2022 Code of Practice on Disinformation’ (ACM Digital Library22 April 2025) 1577 <https://dl.acm.org/doi/10.1145/3696410.3714898>.
[548] Gergely Gosztonyi, ‘How the European Union Had Tried to Tackle Fake News and Disinformation with Soft Law and What Changed with the Digital Services Act?’ (2024) 3 Frontiers in Law 102.
[549] DSA, arts 34–35.
[550] DSA, recital 104.

Since the E-Commerce Directive, the EU has consistently developed digital governance through a gradual legislative process, with subsequent legal instruments or voluntary regulatory guidelines systematically extending the previous regulatory framework[551]. This evolutionary model ensures that stakeholders such as online platforms, regulators, and users can predict the direction of future reforms[552], even if specific obligations might be changed.

In conclusion, the evolution of EU legislation governing platform liability has not only imposed stricter requirements regarding the subjects and scope of content moderation obligations but has also introduced greater procedural predictability.

The evolution of online platform regulatory liability in the US has been primarily shaped by judicial interpretation of statutory provisions and case-based adjudication[553], rather than by comprehensive legislative reform. This approach relies heavily on the principle of precedent, whereby judicial rulings in individual cases accumulate to establish binding legal principles[554], thereby progressively defining the scope of liability for online platforms[555].

In the US, the key provision governing online platform liability remains Section 230 of the CDA, which effectively grants online platforms broad immunity for user-generated content. However, courts have interpreted Section 230 inconsistently across cases, resulting in a fragmented and unpredictable judicial landscape[556]. For example, in Force

---

[551] Giancarlo Frosio, 'From the E-Commerce Directive to the Digital Services Act' (SSRN2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4914816> accessed 4 August 2025.

[552] Charis Papaevangelou and Votta Fabio, 'Trading Nuance for Scale? Platform Observability and Content Governance under the DSA' (2025) 14 Internet Policy Review <https://policyreview.info/articles/analysis/platform-observability-and-content-governance>.

[553] Eduardo Manuel de Almeida Leite and Maria André Ramos Leite, 'Platform Liability, Free Speech, and Market Fairness: Transatlantic Legal Responses to Commercial Defamation and Digital Competition' [2025] International Review of Law, Computers & Technology 1.

[554] Andrea O'Sullivan, 'Section 230 Isn't an Aberration, It's a Distillation of Common Law Trends' (Mercatus Center16 July 2019) <https://www.mercatus.org/economic-insights/expert-commentary/section-230-isnt-aberration-its-distillation-common-law-trends> accessed 19 October 2025.

[555] Leilani Jimenez, 'Tech Regulation Digest: Sunsetting Section 230—the Future of Content Moderation, Ads, and AI | Milken Institute' (Milken Institute3 March 2025) <https://milkeninstitute.org/content-hub/collections/articles/tech-regulation-digest-sunsetting-section-230-future-content-moderation-ads-and-ai>.

[556] Alan Rozenshtein, 'Interpreting the Ambiguities of Section 230' (Yale Journal on Regulation17 April 2024) 60 <https://www.yalejreg.com/bulletin/interpreting-the-ambiguities-of-section-

v. Facebook Inc.[557], the Second Circuit ruled that Facebook's recommendation algorithm was protected by Section 230 immunity; while in Anderson v. TikTok[558], the Third Circuit held that Section 230 does not protect TikTok from liability for its own recommendations when a platform's algorithmic recommendations constitute "expressive activity".

Driven by judicial discretion, different courts have interpreted the scope of Section 230's exemptions differently on a case-by-case basis[559]. Therefore, despite the flexibility of US legal regulations on platform liability[560], the consequences are more uncertain than those of the EU and China[561].

The evolution of China's regulatory liabilities for online platforms reflects a developmental model that combines legal regulations with national administrative oversight[562]. An analysis of the trajectory of its laws and regulations reveals a trend toward stricter platform regulation[563] and expanded platform obligations, particularly reflected in the Cybersecurity Law (2017), Data Security Law (2021), and Personal Information Protection Law (2021). Subsequent documents issued by the Cyberspace Administration of China (CAC) and other departments, such as the "Provisions on Administration of Algorithmic Recommendation in the Internet Information Service" (2021) and the "Administrative Provisions on Deep Synthesis in Internet-based Information Services" (2022), force online platforms to preemptively review and

---

230/>accessed 19 October 2024.

557 Force v Facebook Inc, 934 F 3d 53 (2d Cir 2019).

558 Anderson v TikTok Inc, No 22-3061 (3rd Cir 27 Aug 2024).

559 Tom Wheeler, 'The Supreme Court Takes up Section 230' (Brookings31 January 2023) <https://www.brookings.edu/articles/the-supreme-court-takes-up-section-230/>.

560 Will Duffield, 'Circumventing Section 230: Product Liability Lawsuits Threaten Internet Speech' (Cato Institute26 January 2021) <https://www.cato.org/policy-analysis/circumventing-section-230-product-liability-lawsuits-threaten-internet-speech?> accessed 19 October 2025.

561 Dickinson (n 88).

562 Anu Bradford, 'The Chinese State-Driven Regulatory Model', Digital Empires: the Global Battle to Regulate Technology (Oxford University Press 2023) <https://academic.oup.com/book/46736/chapter-abstract/418514383?redirectedFrom=fulltext&login=false>.

563 Angela Huyue Zhang, 'Agility over Stability: China's Great Reversal in Regulating the Platform Economy' (papers.ssrn.com28 July 2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3892642> accessed 4 August 2025.

remove harmful or false content[564] and label deeply synthesized information[565].

This pattern reflects not only the increasing intensity of regulation but also the predictability of the laws imposing platform liability. China's internet regulatory enforcement efforts are primarily coordinated by the CAC, thereby establishing a regulatory framework dominated by administrative oversight, supplemented by judicial involvement[566]. As regulatory actions within this governance framework align with overarching national policy priorities, it provides procedural predictability[567].

Overall, selected jurisdictions are trending to strengthening platforms' responsibility for content regulation. By clarifying moderation obligations and encouraging proactive review, these regulations are becoming stricter and more predictable, particularly in the EU and China. In the US, however, platform liability remains difficult to predict due to variations in state laws and evolving case law.

3.4 Comparative Analysis of Platform Liability Attribution Rules

3.4.1 Principles of Liability Attribution for Online Platforms in the Regulation of Disinformation

The central issue in regulating disinformation on online platforms lies in determining the scope and triggering conditions of their legal liability. From a legal perspective, determining the principle of liability attribution is a prerequisite for analyzing the triggering conditions of the duty of care[568]. The principle of liability attribution defines the circumstances under which platforms should assume liability for moderating and

---

564 dministrative Provisions on Deep Synthesis in Internet-based Information Services (互联网信息服务深度合成管理规定) (promulgated 11 November 2022, effective 10 January 2023).

565 Administrative Provisions on Deep Synthesis in Internet-based Information Services, arts 12 and 16.

566 Jamie P Horsley, 'Behind the Facade of China's Cyber Super-Regulator' (DigiChina8 August 2022) <https://digichina.stanford.edu/work/behind-the-facade-of-chinas-cyber-super-regulator/>.

567 Ming Zeng and Yongshin Kim, 'Institutional Reforms and Regulatory Shifts in China's Digital Platform Sector: How Domain-Specific Centralization Shaped the 2020–2022 Transition' [2025] Business and Politics 1.

568 Edward J Janger and Aaron D Twerski, 'Functional Tort Principles for Internet Platforms: Duty, Relationship, and Control
| Yale Journal of Law & Technology' (2025) 26 Yale Journal of Law & Technology <https://yjolt.org/functional-tort-principles-internet-platforms-duty-relationship-and-control>.

regulating disinformation when it appears on online platforms.

First, the fundamental basis for pursuing liability should be determined: whether the platform's liability arises only when fault is established (fault liability), when fault is presumed but can be disclaimed by proving reasonable care (presumed fault liability), or irrespective of fault (strict liability). Different liability rules in selected jurisdictions determine the circumstances under which online platforms are held liable for AI-generated disinformation. The distinction between fault liability and strict liability reflects the differences in how different jurisdictions allocate risks between online platforms, users, and national regulators. In exploring these two attribution principles, this thesis adopts the unified terms "injured party" and "liable party". Here, the "injured party" refers to the entity whose rights have been infringed or whose legitimate interests have been harmed[569], and can be an individual, a legal person, or an organization[570]. The "liable party" refers to the entity required to bear primary or contributory liability, including direct infringers and intermediaries, such as online platforms.

In tort law[571], fault liability requires fault on the part of the rights holder in a subjective sense (including intent and negligence), as well as a causal relationship between such fault and the harm, to establish that the fault was sufficient to cause the harm or constituted a direct cause of the harm[572]. Applying fault liability to online platforms, they are not liable unless they breach their duty of care in content moderation. Presumed fault liability still belongs to the fault-based liability system, but it alleviates the evidentiary asymmetry between plaintiff and defendant by reversing the burden of proof[573]. Presumed fault liability shifts the burden of proof to the liable party, thereby

---

569 Cornell Law School, 'Legal Information Institute (Cornell LII)' (LII / Legal Information Institute2024) <https://www.law.cornell.edu/wex/tort>.

570 Adam Scott Wandt, 'Tort: Property' in Lauren R Shapiro and Marie-Helen Maras (eds), Encyclopedia of Security and Emergency Management (Springer International Publishing 2020).

571 Michael Moore, 'Causation in the Law' (Stanford.edu3 October 2019) <https://plato.stanford.edu/archIves/fall2024/entries/causation-law/#LawsExplDefiCaus> accessed 13 August 2025.

572 Page Keeton and William L Prosser, Prosser and Keeton on Torts (West Pub Co 1984).

573 Christian von Bar, The Common European Law of Torts: Volume One, vol. 2 (Oxford University Press 1998) <https://academic.oup.com/book/37023> accessed 7 October 2025.

reducing the injured party's burden of proof. Once the injured party proves the existence of damage and the causal relationship, platforms are presumed at fault unless they prove that they have exercised reasonable care[574]. In the context of online platform regulation, the presumed fault liability is typically implemented through a "notice and action"[575]. If an online platform fails to take appropriate measures promptly after formally receiving notification of illegal or harmful content, it would be presumed to be at fault and thus liable. By introducing a presumption of negligence, the burden of proof shifts to online platforms, thereby strengthening their incentive to exercise due care, particularly in circumstances where evidentiary limitations make it difficult to establish an infringement[576]. This attribution method effectively promotes the liable party's supervision of disinformation, alleviates the difficulty of proof for the injured party due to insufficient evidence, thereby reducing the generation of negligent behavior, and ensures that the liability party fulfills its regulatory obligation[577].

Strict liability is a liability regime independent of negligence and intent, based on allocating risks to those who can control such activities, rather than to the injured party[578]. This method is outcome-oriented: once a causal link between the infringer's action and the damage is established, the liable party is objectively taking liability, regardless of whether subjective fault exists, with only very limited defenses, such as force majeure or the wrongful acts of a third party[579]. Such a way of assigning liability allocates risks to the party with greater control, thereby prompting them to moderate

---

[574] Basil S Markesinis, 'Tort - Tort Law and Alternative Methods of Compensation', Encyclopædia Britannica (2019) <https://www.britannica.com/topic/tort/Tort-law-and-alternative-methods-of-compensation>.

[575] Theresa Ehlen, 'The Digital Services Act: New Liability Rules?' (Passle10 July 2023) <https://technologyquotient.freshfields.com/post/102iiyf/the-digital-services-act-new-liability-rules>.

[576] Bruce L Hay and Kathryn E Spier, 'Burdens of Proof in Civil Litigation: An Economic Perspective' (1997) 26 The Journal of Legal Studies 413.

[577] Alice Guerra, Barbara Luppi and Francesco Parisi, 'Do Presumptions of Negligence Incentivize Optimal Precautions?' (2022) 54 European Journal of Law and Economics 349.

[578] John Goldberg, Harvard School and Benjamin Zipursky, 'Fordham Law Review Fordham Law Review the Strict Liability in Fault and the Fault in Strict Liability the Strict Liability in Fault and the Fault in Strict Liability Recommended Citation Recommended Citation' (2016) <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5250&context=flr> accessed 4 August 2025.

[579] Jules L Coleman, 'Fault and Strict Liability', Risks and Wrongs (Oxford University Press 2002).

and manage posted information more carefully and invest more resources to prevent the generation of disinformation[580]. While strict liability is not directly applicable to online platforms in any selected jurisdiction, it remains an important theoretical framework in liability attribution. The exclusion of strict liability from the EU, the US, and China reflects that even if online platforms act as gatekeepers for user-generated content and bear the risks of the generation and dissemination of disinformation[581], they are only liable to the extent of their negligence.

The principle of liability attribution determines the scope and circumstances under which online platforms are held liable for regulating and managing disinformation. The legal rules governing the liability of online platforms for user-generated disinformation differ significantly between the EU, the US, and China. In the EU, the conditional fault liability model, which combines knowledge-triggered threshold with safe harbor protection, reflects the balance in the EU legal framework between ensuring freedom of expression and protecting data subjects from harmful online disinformation. In contrast, Section 230 of the CDA grants online platforms broad immunity for user-generated content, reflecting both the First Amendment's commitment to free speech and a policy preference for fostering innovation[582]. While for the statutory exceptions listed under Section 230, liability attribution principles for online platforms typically follow a fault-based liability model. China's legal system defines the liability of online platforms to monitor disinformation as fault-based and reinforces the proactive management obligations of platforms in terms of administrative liability requirements (e.g., the Cybersecurity Law), so the threshold for meeting the duty of care depends not only on notification, but also on the platform's technological ability to foresee risks.

---

[580] Gregory C Keating, 'The Idea of Fairness in the Law of Enterprise Liability' (1997) 95 Michigan Law Review 1266.

[581] Martin Husovec, 'Amicus Curiae Delfi as v Estonia' (Scribd2025) <https://www.scribd.com/document/232759055/Amicus-Curiae-Delfi-AS-v-Estonia> accessed 16 August 2025.

[582] Ashley Johnson and Daniel Castro, 'Overview of Section 230: What It Is, Why It Was Created, and What It Has Achieved' (itif.org22 February 2021) <https://itif.org/publications/2021/02/22/overview-section-230-what-it-why-it-was-created-and-what-it-has-achieved/> accessed 10 July 2025.

While online platforms face high compliance requirements and regulatory pressure under laws and regulations, and the boundaries of liability are interpreted more strictly in practice, China's legal attribution still falls within the framework of fault liability. The online platforms retain the right to defend themselves by proving that they have fulfilled their moderation obligations or duty of reasonable care.

The E-Commerce Directive (2000/31/EC) provides the EU's foundational liability framework. Article 14[583] establishes a safe harbor for hosting service providers, which exempts online platforms from liability for information stored at user request if they lack actual knowledge of the illegal activity or act promptly to remove or disable access to the content upon becoming aware of it. The Courts have interpreted this provision that the role of the platform may shift from passive hosting service provider to active participant if the platform's involvement gives it "knowledge or control" over specific unlawful information; in such cases, safe harbor protections may fail, and liability may be imposed on grounds of fault[584]. The EU standard aims to prevent platforms from willful blindness to avoid liability, while avoiding the imposition of a general obligation to monitor content[585]. This reflects the conditional fault-based liability, under which attribution of liability depends on the platform's actual or constructive knowledge of illegal content[586]. Online platforms that lack awareness of unlawful activity, or act expeditiously upon obtaining such awareness, are exempt from liability under the safe harbor provisions. While the Directive does not explicitly classify disinformation as "illegal content", it does impose liability when it overlaps with illegal expressions, such as hate speech, defamation, or election interference[587]. The DSA maintains the fault-

---

583 E-Commerce Directive, art 14.

584 case C-324/09 L'Oréal SA v eBay International AG [2011] ECR I-6011, ECLI:EU:C:2011:474.

585 Christina Angelopoulos, 'On Online Platforms and the Commission's New Proposal for a Directive on Copyright in the Digital Single Market' (papers.ssrn.com1 January 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2947800>accessed 3 February 2025.

586 Aleksandra Kuczerawy, 'Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative' (Ssrn.com2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2560257> accessed 25 February 2019.

587 Sarah Ziedler, 'Navigating Platform Power: From European Elections to the Regulatory Future' (HIIG18 July 2024) <https://www.hiig.de/en/dsa-navigating-platform-power/> accessed 18 August 2025.

based liability framework established by the E-Commerce Directive. Articles 4-6[588] of the DSA explicitly reaffirm the safe harbor regime, preserving the principle that online platforms are not directly liable for illegal content of which they are unaware and must promptly remove it upon notification[589]. However, the DSA requires VLOPs to identify, analyze, and mitigate systemic risks, adopting a proactive approach that goes beyond the passive "notice and action" regime of the E-Commerce Directive[590]. While these obligations do not transform the fault liability regime into a strict liability regime, they raise the threshold for a duty of care and effectively narrow the scope of the safe harbor protections[591]. Failure to fulfill the DSA's systemic risk management obligations does not automatically give rise to civil liability but may serve as evidence of negligence under the fault-based liability framework[592]. In this sense, the DSA encourages specific platforms (such as VLOPs) to play a proactive governance role in managing online harms, including disinformation. In conclusion, in the EU, online platforms basically bear fault liability for the generation and spread of disinformation, and embed prevention liabilities, risk assessments, and mitigation measures[593] into the platform operation through a clear institutional chain[594].

Unlike the EU and China, the US has taken a distinct approach to the liability of online

---

[588] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1, arts 14–16.

[589] Gerhard Wagner, 'Liability Rules for the Digital Age' (2022) 13 Journal of European Tort Law 191.

[590] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1, arts 34—35.

[591] Giancarlo Frosio and Christophe Geiger, 'Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime' (2023) 29 European Law Journal 31.

[592] Marc Tiernan and Goran Sluiter, 'The European Union's Digital Services Act and Secondary Criminal Liability for Online Platform Providers: A Missed Opportunity for Fair Criminal Accountability?' (SSRN Electronic Journal2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4731220> accessed 5 December 2024.

[593] Giulia Giannasi, 'Risk in the Digital Services Act and AI Act: Implications for Media Freedom, Pluralism, and Disinformation - Centre for Media Pluralism and Media Freedom' (Centre for Media Pluralism and Media Freedom27 May 2025) <https://cmpf.eui.eu/risk-in-the-digital-services-act-and-ai-act-implications-for-media-freedom-pluralism-and-disinformation/> accessed 18 August 2025.

[594] Christoph Busch, 'Platform Responsibility in the European Union', Defeating Disinformation (Cambridge University Press 2025) <https://www.cambridge.org/core/books/defeating-disinformation/platform-responsibility-in-the-european-union/AA3D55C57B0F6A7C18F5CAEF25146557>.

platforms for disinformation. The US has chosen a liability immunity system that almost protects online platforms from being held liable for user-generated content, but there are still exceptions to this immunity system. This approach is embodied in section 230 of the CDA, which regulates that interactive computer service providers or users are only liable for the information they create[595]. Consequently, online platforms generally cannot be held liable for third-party content, including disinformation, whether based on fault or not. Courts have consistently interpreted this provision expansively, effectively shielding online platforms from defamation, negligence, or similar tort claims[596]. For example, in Zeran v. America Online, the Fourth Circuit ruled that immunity applies even if online platforms ignore notice of defamatory materials and warned that imposing liability could have a limiting effect on content moderation[597]. Likewise, in Blumenthal v. Drudge, the court dismissed the claim against AOL for defamatory content authored by a user, attributing immunity under Section 230[598]. In Gonzalez v Google, the Supreme Court refrained from narrowing this immunity, thereby reaffirming that practices of attributing user-generated disinformation to online platforms remain highly limited[599]. Dickinson pointed out that nearly all interpretations of the immunity clause are broad and vaguely worded, thus exempting internet intermediaries from tort liability if they are not the direct authors of disinformation[600]. But this immunity does not apply to all categories of disinformation. Online platforms may be held liable on a fault-based principle under statutory exceptions (federal criminal violations, intellectual property infringement, and information fraud)[601]. In cases involving federal criminal crimes, platforms can be held liable if they knowingly aided or encouraged the commission of the crime, reflecting the fact that they have

---

595 Communications Decency Act 1996, 47 USC § 230(c)(1).
596 Louis Shaheen, 'Section 230'S Immunity for Generative Artificial Intelligence' (SSRN.com15 December 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4886463> accessed 12 October 2025.
597 Zeran v. America Online Inc, 129 F.3d 327 (4th Cir 1997).
598 Blumenthal v. Drudge, 992 F Supp. 44 (D D C 1998).
599 Gonzalez v Google LLC 143 S Ct 1191 (2023).
600 Dickinson (n 88).
601 CDA § 230(e)(1).

subjective fault[602]. Section 230(e)(2) of the CDA expressly excludes intellectual property infringement from the scope of liability immunity afforded to online platforms[603]. If a platform has actual knowledge of an infringement, or if it exercises sufficient control over the infringing activity to be presumed to have such knowledge[604], and makes a material contribution to it[605], the platform may be held jointly liable with the infringer. In conclusion, the US has adopted a relatively broad immunity model, exempting platforms from liability for disinformation generated by third parties, but under statutory exceptions, platforms may be liable based on their fault.

China's liability attribution system combines fault-based liability with presumed fault-based liability. In terms of civil liability, Chinese law follows the principle of fault liability. The core of this principle is that online platforms are not primarily liable for all the disinformation that is posted on their platforms, but rather for their inaction or negligence. According to the "notice-and-delete" rule established in Articles 1194-1197 of the Civil Code[606], if an online platform fails to take necessary measures after being aware or should have known of disinformation that infringes on the civil rights of others, it shall bear joint and several liability with the infringing users. At the administrative liability level, strict supervision based on the principle of presumption of fault is implemented. On the one hand, Article 47 of the Cybersecurity Law[607] explicitly requires network service providers to take immediate remedial measures once they know or should have known that their users have disseminated information prohibited by law. Articles 3-9 of the Provisions on the Governance of the Online Information

---

602 Charles Matula, 'Any Safe Harbor in a Storm: SESTA-FOSTA and the Future of § 230 of the Communications Decency Act' (2019) 18 Duke Law & Technology Review 353 <https://scholarship.law.duke.edu/dltr/vol18/iss1/24/> accessed 14 October 2025.

603 CDA § 230(e)(2).

604 A&M Records, Inc v Napster, Inc 239 F 3d 1004 (9th Cir 2001).

605 Larry Wayte, 'Contributory and Vicarious Liability for Peer-To-Peer File Sharing Services: The Napster and Grokster Cases', Pay for Play: How the Music Industry Works, Where the Money Goes, and Why (Pressbooks 2023) <https://opentext.uoregon.edu/payforplay/chapter/chapter-38-contributory-and-vicarious-liability-for-peer-to-peer-file-sharing-services-the-napster-and-grokster-cases/> accessed 14 October 2025.

606 Civil Code of the People's Republic of China (promulgated by the National People's Congress, 28 May 2020) arts 1194–1197.

607 Cybersecurity Law of the People's Republic of China (adopted 7 November 2016, effective 1 June 2017) art 47.

Content Ecosystem[608] legalize the main responsibilities of online platforms, requiring them to act as proactive managers. By establishing and improving internal systems such as information moderation and real-time inspections, online platforms should proactively identify, and address disinformation prohibited by legal regulations (including violent, terrorist, and pornographic content), rather than simply passively responding to reports[609]. Similarly, Articles 7-10 of the Provisions on the Administration of Algorithmic Recommendations for Internet Information Services[610] require service providers to adjust or disable recommendation features that amplify harmful content, prominently label algorithmically generated synthetic information, and establish a signature database for identifying illegal or harmful information. On the other hand, the triggering of the corresponding penalties (such as Article 68 of the Cybersecurity Law[611]) does not require that disinformation endanger cybersecurity. As long as the regulatory authorities determine that the online platforms have "failed to fulfill their management obligations", administrative penalties can be initiated[612]. Such penalties target the platform's inaction, rather than subjective fault or actual damage caused. In this case, the burden of proof effectively shifts to the platform: if the platform tries to avoid administrative penalties, it must bear the burden of proving that it has fulfilled its management obligations.

3.4.2 The Threshold for Triggering the Online Platform's Duty of Care

After examining the attribution principles used to determine online platforms' content moderation liabilities in selected jurisdictions, this thesis will explain the conditions under which a platform's duty of care is triggered[613]. The principle of liability

---

[608] Provisions on the Governance of the Online Information Content Ecosystem (网络信息内容生态治理规定) (promulgated 15 December 2019, effective 1 March 2020) arts 3-9.
[609] Provisions on the Governance of the Online Information Content Ecosystem (网络信息内容生态治理规定) (promulgated 15 December 2019, effective 1 March 2020) art 10.
[610] Provisions on the Administration of Algorithmic Recommendations for Internet Information Services (adopted 31 December 2021, effective 1 March 2022) arts 7–10.
[611] Cybersecurity Law of the PRC, art 68.
[612] Ibid.
[613] Janger and Twerski (n 568) 39-42.

attribution provides the normative foundation for determining platform liability[614], while the threshold of duty of care assumes regulatory liability by determining when the actor must take measures[615]. The threshold for an online platform's duty of care defines the conditions that trigger its intervention, such as receiving a notice from the infringed party requesting the removal of disinformation[616].

The concept of the duty of care is important for determining liability, defining the necessary measures that online platforms should adopt in preventing foreseeable harm[617]. The modern concept of the duty of care originates from the common law tradition[618], particularly the judgments in Donoghue v. Stevenson. As articulated in Donoghue v Stevenson[619], parties should exercise reasonable care to avoid actions or omissions that are foreseeably likely to harm their "neighbors" (those closely and directly affected by their actions), extending protection to individuals who have no direct legal or economic relationship with the defendant (such as consumers harmed by a defective product)[620]. In Robinson v Chief Constable of West Yorkshire Police[621], the court emphasized that the existence of the duty of care should be determined by established categories of duty, as well as by foreseeability and proximity to the plaintiff[622]. Since online platforms are not merely participants in the dissemination of

---

[614] Uta Kohl, 'Toxic Recommender Algorithms: Immunities, Liabilities and the Regulated Self-Regulation of the Digital Services Act and the Online Safety Act' (2024) 16 Journal of Media Law 1.

[615] Karel Roynette, 'Drawing the Line of the Scope of the Duty of Care in American Negligence and French Fault-Based Tort Liability' (2015) 8 Journal of Civil Law Studies <https://digitalcommons.law.lsu.edu/jcls/vol8/iss1/4/>.

[616] Daphne Keller, 'Systemic Duties of Care and Intermediary Liability' (Stanford CIS29 May 2020) <https://cyberlaw.stanford.edu/blog/2020/05/systemic-duties-care-and-intermediary-liability/> accessed 5 December 2024.

[617] Richard Castle, 'Lord Atkin and the Neighbour Test: Origins of the Principles of Negligence in Donoghue v Stevenson' (2003) 7 Ecclesiastical Law Journal 210 <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/CBCF36E5E5998EB037E232CAAE3317ED/S0956618X00005214a.pdf/lord_atkin_and_the_neighbour_test_origins_of_the_principles_of_negligence_in_donoghue_v_stevenson.pdf>.

[618] Peter Cane and James Goudkamp, Atiyah's Accidents, Compensation and the Law (Cambridge University Press 2018) 85 85–90.

[619] Donoghue v Stevenson [1932] AC 562 (HL).

[620] Law Teacher, 'Donoghue v Stevenson [1932] Doctrine of Negligence' (Lawteacher.net7 March 2018) <https://www.lawteacher.net/cases/donoghue-v-stevenson.php>.

[621] Robinson v Chief Constable of West Yorkshire Police [2018] UKSC 4, [2018] AC 736.

[622] KM Stanton, 'Professional Negligence: Duty of Care Methodology in the Twenty First Century' (2006) 22 Tottel's Journal of Professional Negligence 134 <https://research-information.bris.ac.uk/en/publications/professional-negligence-duty-of-care-methodology-in-the-

information but also intermediaries of information exchange, their role enables them to review, manage, and influence content to varying degrees[623]. Therefore, the duty of care not only requires online platforms to protect the personal interests of information owners, but also requires them to design review and reporting systems to mitigate foreseeable damages in platform operations[624].

A key issue in this regulatory framework is the threshold for exercising the duty of care[625]. The threshold of online platform liability refers to the standard that determines when a platform shifts from an intermediary position that does not bear content liability to one that has an obligation to manage and review disinformation on its services[626]. This threshold is closely tied to the principle of liability attribution adopted by selected jurisdictions. If the attribution of liability is based on fault, the duty of care is often triggered only when the online platform knows (or should have known) of the illegal conduct or disinformation[627]. If liability is based on strict or presumed fault liability, the duty of care arises more proactively, requiring online platforms to adopt preventive and regulatory measures to mitigate potential harm, even in the absence of notifications of disinformation[628]. In this sense, the establishment of the liability principle is the logical foundation for analyzing the framework of the duty of care: this principle not only determines the subject of liability but also the triggering threshold of the duty of

---

twenty-fi> accessed 21 August 2025.

[623] Tarleton Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (Yale University Press 2018).

[624] Lorna Woods and Will Perrin, 'Obliging Platforms to Accept a Duty of Care', Regulating Big Tech: Policy Responses to Digital Dominance (Oxford University Press 2021) <https://academic.oup.com/book/39213/chapter/338717347>.

[625] James C Plunkett, 'The Historical Foundations of the Duty of Care' (2015) 41 bridges.monash.edu <https://bridges.monash.edu/articles/journal_contribution/The_Historical_Foundations_of_the_Duty_of_Care/10065659>.

[626] Paddy Leerssen, 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems' (papers.ssrn.com24 February 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3544009> accessed 4 August 2025.

[627] Trevor Cook, 'Online Intermediary Liability in the European Union' (2012) 17 Journal of Intellectual Property Rights 157 <https://www.researchgate.net/publication/295162458_Online_intermediary_liability_in_the_European_Union>.

[628] Carsten Ullrich, 'Standards for Duty of Care? Debating Intermediary Liability from a Sectoral Perspective' (2017) 8 Journal of Intellectual Property, Information Technology and E-Commerce Law 111 <https://www.jipitec.eu/jipitec/article/view/197> accessed 18 September 2025.

care[629].

Comparatively, common triggering conditions can be categorized into three categories. First, most jurisdictions (in this study, the EU and China) consider actual knowledge as the baseline trigger for a platform's duty of care, typically arising upon notification by the right holders or the injured party[630]. A platform is deemed to have actual knowledge when it directly learns of illegal or harmful content or is subjectively aware of the illegal nature of that content[631]. This standard ensures effective prevention of the spread of illegal content or disinformation while avoiding imposing an excessive burden on monitoring that could undermine freedom of expression[632] or lead to the excessive removal of legal content[633]. From a legal perspective, actual knowledge provides a clear threshold for liability by attributing fault to the platform's inaction once it has received notice or the platform is otherwise aware of it[634]. This explains why notice-action mechanisms premised on actual knowledge have become a common standard in regimes such as the EU's E-Commerce Directive (Article 14) and the DSA (Article 16). Also, it is reflected in China's Civil Code (Articles 1195 and 1196), whereby once the internet service provider (online platform) is aware of the existence of unlawful

---

[629] Miriam C Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation' (2021) 12 Journal of Intellectual Property, Information Technology and E-Commerce Law 361 <https://www.jipitec.eu/jipitec/article/view/331>.

[630] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), art 16; Cybersecurity Law of the People's Republic of China, art 47; Measures for the Administration of Internet Information Services, art 14.

[631] Nigel Miller, 'Navigating the Web 2.0 Safe Harbour - Internet for Lawyers Newsletter' (*Internet for Lawyers Newsletter*September 2008) <https://www.infolaw.co.uk/newsletter/2008/09/navigating-the-web-20-safe-harbour/> accessed 24 August 2025.

[632] Giovanni Sartor, 'DIRECTORATE GENERAL for INTERNAL POLICIES POLICY DEPARTMENT A: ECONOMIC and SCIENTIFIC POLICY Providers Liability: From the ECommerce Directive to the Future IN-DEPTH ANALYSIS' (2017) <https://www.europarl.europa.eu/RegData/etudes/IDAN/2017/614179/IPOL_IDA(2017)614179_EN.pdf>.

[633] Christoph Schmon and Haley Pedersen, 'Platform Liability Trends around the Globe: Taxonomy and Tools of Intermediary Liability' (Electronic Frontier Foundation25 May 2022) <https://www.eff.org/deeplinks/2022/05/platform-liability-trends-around-globe-taxonomy-and-tools-intermediary-liability>.

[634] João Quintais, 'The New Copyright in the Digital Single Market Directive: A Critical Look' (*papers.ssrn.com*14 October 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3424770> accessed 4 August 2025.

information, it is required to assume liability for monitoring and removing it[635]. In systems that emphasize regulatory intervention, such as the EU and China, actual knowledge is often used as a liability threshold, although its impact on platform liability varies across selected jurisdictions.

Second, both the EU and China have expanded their baseline beyond actual knowledge to encompass constructive knowledge, whereby platforms may be held liable if they should have known of unlawful or harmful content under the circumstances[636].

In discussing the liability of online platforms for the dissemination of disinformation, this thesis adopts the term "constructive knowledge[637]". This concept refers to situations where the platform does not have actual knowledge of illegal or disinformation but is deemed to have "ought reasonably to have known" of such information in the exercise of its duty of care. In the EU, Article 14 of the E-Commerce Directive[638] provides online platforms with immunity from liability when they had no actual knowledge of illegal content or activities, or when such illegal circumstances were not so obvious[639] as to allow for constructive knowledge[640]. In the US, the Digital Millennium Copyright Act (DMCA) distinguishes between "actual knowledge" and situations where a service provider is "aware of facts or circumstances from which infringing activity is apparent"[641]. The latter provision, codified in 17 U.S.C. §512(c)(1)(A)(ii), indicates that the online service provider is deemed to have constructive knowledge when infringement would be obvious to a reasonable

---

635 GRUR International, 'Liability of E-Commerce Platform Operators for Taking Necessary Measures' (2023) 72 GRUR International 566.

636 Graeme B Dinwoodie, 'A Comparative Analysis of the Secondary Liability of Online Service Providers' (Ssrn.com17 May 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2997891> accessed 24 August 2025.

637 Donoghue v Stevenson [1932] AC 562 (HL); Restatement (Second) of Torts § 12 (American Law Institute 1965); Jane Stapleton, 'Duty of Care Factors: A Selection from the Judicial Menus' (2000) 20 OJLS 101; Carsten Ullrich, 'Standards for Duty of Care? Debating Intermediary Liability from a Sectoral Perspective' (2017) 8 JIPITEC 207.

638 E-Commerce Directive, art 14(1)(a).

639 Aleksandra Kuczerawy, 'The Good Samaritan That Wasn't: Voluntary Monitoring under the (Draft) Digital Services Act' (Verfassungsblog12 January 2021) <https://verfassungsblog.de/good-samaritan-dsa/> accessed 5 December 2024.

640 Joined Cases C-236/08 to C-238/08 Google France SARL and Google Inc v Louis Vuitton Malletier SA and Others [2010] ECR I-2417, para 114.

641 17 USC § 512(c)(1)(A)(ii).

operator[642]. In China, the standard is commonly expressed as "knows or ought to know", which is generally considered equivalent to constructive knowledge. For these reasons, the term 'constructive knowledge' is used throughout this thesis to denote ideas that are presumed to be known.

The following two reasons explain the expansion of constructive knowledge to include platforms' knowledge of disinformation. On the one hand, if liability were based solely on actual knowledge, it would be easy for platforms to evade liability through willful blindness[643], intentionally ignoring obvious signs of illegal or harmful content to avoid liability[644]. Therefore, in the context of platform regulation, limiting liability to those with actual knowledge would allow online platforms to evade liability by ignoring obvious disinformation before receiving formal notice[645]. To address this loophole, both EU and Chinese laws and regulations stipulate that if an online platform has received notification from the relevant rights holder or the disinformation is obviously wrong (for example, debunking of such disinformation has been widely disseminated), it is deemed to be aware of illegal or harmful content. On the other hand, unlike traditional publishers or passive intermediaries, online platforms possess the technological means of algorithmic analysis, targeted advertising, and recommendation systems[646], and therefore can amplify, rank, or organize user-generated content. These mechanisms allow platforms to gain extensive insights into user behaviors and content distribution patterns, making it difficult to remain a neutral position[647]. In L'Oréal v. eBay[648], the Court of Justice of the European Union (CJEU) clarified that platforms could lose safe

---

642 Angelopoulos (n 585) 28.

643 Giancarlo F Frosio, 'Reforming Intermediary Liability in the Platform Economy: A European Digital Single Market Strategy' (Northwestern Pritzker School of Law Scholarly Commons2017) <https://scholarlycommons.law.northwestern.edu/nulr_online/251/> accessed 18 September 2025.

644 Global-Tech Appliances, Inc. v SEB SA 563 US 754 (2011).

645 Jack M Balkin, 'Information Fiduciaries and the First Amendment' (papers.ssrn.com3 February 2016) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2675270> accessed 14 April 2025.

646 Miriam Fernández, Alejandro Bellogín and Iván Cantador, 'Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation' [2021] arXiv:2103.14748 [cs] <https://arxiv.org/abs/2103.14748> accessed 14 April 2025.

647 Zeran v America Online Inc 129 F3d 327 (4th Cir 1997).

648 Case C-324/09 L'Oréal SA v eBay International AG [2011] ECR I-6011, ECLI:EU:C:2011:474, para 113.

harbor protection if they played an "active role" in knowing about or controlling the content they hosted, thereby introducing a presumption of knowledge standard[649]. This is reinforced by the DSA, which mandates proactive risk assessment and systemic risk mitigation, particularly for VLOPs and VLOSEs, suggesting that platforms may incur liability if they fail to take reasonable measures to address foreseeable harms from disinformation(Article 34). Similarly, in China, the Civil Code (Article 1197) and subsequent regulations, such as the 2020 Provisions on the Governance of the Online Information Content Ecosystem[650], explicitly stipulate that platforms are liable if they "know or should have known" their internet users exploiting its services to infringe upon the civil rights and interests of others, requiring them to implement preventative content review systems, flag disinformation, and report serious incidents to superior departments[651]. One of the main sources of risks in the spread of disinformation is the online platforms' behaviors, so the degree of control that the online platform exercises over the dissemination of information will affect the presumption of the possibility that it "should have known"[652]. Therefore, the constructive knowledge expands the platform's duty of care. If the technological capabilities of an online platform enable it to identify illegal content or disinformation, it could no longer claim ignorance of such disinformation[653]. Ensure that regulatory liability applies not only upon notification, but also potentially holds platforms accountable for disinformation when the platform's technical capabilities[654] or the obviousness of the harm make ignoring it unreasonable.

---

[649] Enrico Bonadio, 'Trade Marks in Online Marketplaces: The CJEU's Stance in L'Oreal v. EBay' (*Ssrn.com*7 March 2012) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2017741> accessed 4 August 2025.

[650] Provisions on the Governance of the Online Information Content Ecosystem (网络信息内容生态治理规定) (promulgated 15 December 2019, effective 1 March 2020).

[651] Xiao Ma, 'Establishing an Indirect Liability System for Digital Copyright Infringement in China: Experience from the United States' Approach - NYU Journal of Intellectual Property & Entertainment Law' (NYU Journal of Intellectual Property & Entertainment Law4 May 2015) <https://jipel.law.nyu.edu/vol-4-no-2-3-ma/>.

[652] Bin zhang, 'Limitation of Safety-Guard Responsibility of Platform Operator' (2019) 22 Economic Law Review.

[653] Qian Tao, 'Legal Framework of Online Intermediaries' Liability in China' (2012) 14 Info 59.

[654] Xiping Zhou, 'E-Commerce Platforms' Security Obligations and Legal Responsibilities' (Southcn.com2022) <https://theory.southcn.com/node_203ed94b00/5fbbe625d3.shtml> accessed 27 August 2025.

This development reflects the dynamic interaction between the threshold of the duty of care and the principle of liability attribution, that is, actual knowledge requires proof of awareness, while constructive knowledge imports a presumed fault where awareness should reasonably have been obtained.

Third, in China's regulatory approach[655], administrative regulations require online platforms to conduct proactive review. Therefore, the presence of content that violates laws and administrative regulations triggers the duty of care[656]. Article 9 of the "Provisions on the Governance of the Online Information Content Ecosystem" requires online platforms to fulfill their primary governance responsibility for content management[657]. Article 10 emphasizes that platforms should establish systems for both manual and machine review, institutionalizing this obligation of proactive review[658]. These two provisions examine whether an online platform has established a supervisory system (including review teams, technical filtering models, and inspection systems) appropriate to its scale, core business, and risks, to determine whether it has assumed proactive management responsibilities[659]. If the platform fails to establish such a system, it is presumed that it should have known that there was a large amount of disinformation on the platform but failed to fulfill its proactive liability and was therefore subject to administrative penalties[660]. Even the supervisory system has been established, but if it fails to operate effectively, it will also be held liable because it constitutes "should have

---

655 Tao Qian, 'The Knowledge Standard for the Internet Intermediary Liability in China' (2011) 20 International Journal of Law and Information Technology 1.

656 Regulation on Internet Information Service (State Council Decree No. 292) (中国国务院令第 292 号) (2000); Cybersecurity Law of the People's Republic of China (adopted 7 November 2016, effective 1 June 2017); Provisions on the Governance of the Online Information Content Ecosystem (CAC, 2020). See generally Qian Tao, 'Legal Framework of Online Intermediaries' Liability in China' (2012) 14 Info 59; Tao Qian, 'The Knowledge Standard for the Internet Intermediary Liability in China' (2011) 20 International Journal of Law and Information Technology 1.

657 Provisions on the Governance of the Online Information Content Ecosystem (promulgated by the Cyberspace Administration of China, effective 1 March 2020) art 9.

658 Provisions on the Governance of the Online Information Content Ecosystem (promulgated by the Cyberspace Administration of China, effective 1 March 2020) art 10.

659 Jiayi Chen and Chang Shi, 'Proactive Governance by Official Administrators on Chinese Social Media Platforms: Boundary Discourse and Governance Legitimacy' [2025] Media Culture & Society.

660 Sisi Zhang, 'Research on the Security Guarantee Obligation of E-Commerce Platform Operators' (2025) 13 E-commerce Reviews <https://www.hanspub.org/journal/paperinformation?paperid=100698&> accessed 27 August 2025.

known" without handling it[661]. For example, China's Cybersecurity Law (Article 47)[662] claims that when online service providers discover illegal information on their networks, they should immediately stop transmitting and report to the relevant authorities. This provision not only includes the liability for content moderation after passively receiving reports but also requires platforms to identify such information through their own review mechanisms proactively. This means that platforms have an active obligation to moderate content and cannot evade responsibility by claiming "ignorance". Therefore, combined with a comparative study of the principles of liability attribution across selected jurisdictions, it is shown that the threshold for triggering the duty of care depends on the underlying principle of attribution. If the platform assumes liability based on fault, actual knowledge of the disinformation is the triggering threshold for the duty of care; if the platform assumes presumed fault liability, the constructive knowledge triggers the duty of care.

The specific conditions triggering a platform's duty of care vary across jurisdictions. In the EU, a platform's safe harbor depends on its ignorance; therefore, once a platform acquires actual knowledge or becomes aware of facts or circumstances that are clearly unlawful, it must take action to remove or block such information[663]. In L'Oréal v eBay[664] (Case C-324/09), the European Court of Justice held that under Article 14 of the E-commerce Directive[665], a hosting service provider is exempt from liability provided that it does not know illegal activities; once it becomes aware of such activities, it must act promptly to remove or disable access to disinformation. The DSA maintains that the duty of care of ordinary intermediaries only arises upon their having actual

---

[661] Feng Xiao, 'Improvement of E-Commerce Platform Responsibility Legislation for Consumer Protection from the Perspective of Informational Interests' (2022) 24 Journal of Shanghai University of Finance and Economics.

[662] Cybersecurity Law 2016, art 47.

[663] Julián López Richart, 'A New Legal Framework for Online Platforms in the European Union (and Beyond)' (2024) 59 Review of European and Comparative Law.

[664] L'Oréal SA v eBay International AG (C-324/09) EU:C: 2011:474, [2011] ECR I-6011.

[665] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce) [2000] OJ L178/1, art 14.

knowledge of illegal content[666] but raises the threshold for duty of care obligations to certain platforms. The intermediary service providers are required to operate 'notice and action mechanisms' and provide their reasons for removing user-generated content, thus establishing a benchmark for actual knowledge (Articles 16-17). Articles 34-35 regulated that VLOPs and VLOSEs are additionally subject to systematic risk assessment, risk mitigation, and independent audits. These requirements of obligations reflect that VLOPs' algorithmic technology and recommendation systems may lead to the amplification of disinformation[667] that can generate systemic risks[668], and that these platforms can foresee the damages caused by such information. Therefore, VLOPs should be subject to a higher duty of care by law[669]. The requirements of these obligations reflect VLOPs and VLOSE's foreseeability of the harm that their algorithmic techniques and recommendation systems may amplify disinformation to a systemic risk, and due to the damage is foreseeable, the law can legitimately impose a higher duty of care. Therefore, the duty of care of VLOPs is triggered upon the identification or potential existence of systemic risks, without the need for specific notification[670].

In the US, due to the broad immunity granted to online platforms by Section 230, user-generated disinformation usually does not trigger a duty of care[671], and courts have been reluctant to impose a proactive monitoring obligation on platforms to protect free speech and avoid a chilling effect[672]. However, online platforms are not exempt from a duty of care regarding all false information. The Digital Millennium Copyright Act (DMCA) emphasizes a conditional immunity for online platforms: under Section

---

666 DSA, arts 6 and 16.
667 Paddy Leerssen, 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems' (papers.ssrn.com24 February 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3544009> accessed 4 August 2025.
668 DSA, art 34.
669 Donoghue v Stevenson [1932] AC 562 (HL).
670 Niklas Eder, 'Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation' (2024) 25 German Law Journal 1.
671 Rozenshtein (n 556) 75.
672 Michael Rustad and Thomas Koenig, 'The Case for a CDA Section 230 Notice-And-Takedown Duty' (2023) 23 Nevada Law Journal <https://scholars.law.unlv.edu/nlj/vol23/iss2/7/> accessed 4 August 2025.

512(c)(1)(A)[673], a platform loses its immunity from copyright infringement if it has actual knowledge of copyright infringement or is aware of facts or circumstances from which infringement is apparent but fails to promptly remove or disable access to the relevant material. This means that a platform's duty of care is triggered when it receives a valid notice of infringement or is presumed to have red flag knowledge[674] of the infringement due to its obviousness. For disinformation that violates federal criminal law, an online platform's duty of care with respect to criminal content is triggered only if it has actual knowledge of the specific illegal conduct and knowingly participates in or facilitates that conduct[675]. Based on Section 230(e)(1) of the CDA[676], in conjunction with 18 U.S.C. §2 and §371[677], platforms can be held criminally liable only if they knowingly and willfully participated in assisting or materially contributing to a crime. Therefore, for disinformation that violates federal criminal law, the online platform would not be held liable[678] if it merely had actual knowledge of the potentially illegal user-generated content or simply failed to remove such content.

The triggering conditions for platforms' duty of care in China are primarily based on their knowledge standard regarding disinformation and the type of content involved. Article 47 of the Cybersecurity Law[679] requires network operators to immediately halt the transmission of disinformation and implement regulatory measures upon discovering it violates laws and administrative regulations. This legal provision reflects that online platforms can assume a duty of care based on actual knowledge after being notified, or they can review and supervise the content posted on the platform through their own proactive management[680]. Secondly, online platforms are required to prevent

---

673 Digital Millennium Copyright Act, 17 USC § 512(c)(1)(A).
674 Jane C Ginsburg, 'Separating the Sony Sheep from the Grokster Goats: Reckoning the Future Business Plans of Copyright-Dependent Technology Entrepreneurs' (2008) 50 Arizona Law Review <https://journals.librarypublishing.arizona.edu/arizlrev/article/id/7378/> accessed 15 October 2025.
675 Gonzalez v Google LLC 598 US ___ (2023), available at https://www.supremecourt.gov/opinions/22pdf/21-1333_19m2.pdf accessed 15 October 2025
676 CDA § 230(e)(1).
677 18 USC §§ 2 and 371.
678 Twitter, Inc v Taamneh 598 US ___ (2023).
679 Cybersecurity Law, art 47.
680 Wang (n 355).

and delete disinformation related to high-risk areas such as political security, social stability, public safety, and financial risks[681], and ensure that such disinformation does not appear in the proactive recommendation areas of their webpages[682]. To facilitate the fulfillment of this obligation, online platforms are required to have information review teams commensurate with the scale of their services and establish convenient reporting channels as a powerful way to eliminate disinformation[683].

3.4.3 Factors Shaping the Attribution of Platform Liability

In this comparative analysis, I first clarify the attribution principles adopted by selected jurisdictions to hold platforms liable for AI-generated disinformation. Having established this attribution principle, this thesis will next examine the factors that influence the attribution of liability. In different jurisdictions, the boundaries of online platforms' liability in regulating disinformation are influenced by various factors, such as the type of services provided by the platform, the type of information processed, the platform's scale, and its influence[684]. Therefore, it can be argued that the attribution principle establishes the basic framework for attributing liability, while the attribution factors determine how liability is shaped and enforced in practice. State what the factors are before venturing into details.

Under the DSA, service type determines the level of liability: "mere conduit" (Article 4) and caching services (Article 5)[685] enjoy minimal obligations, enjoying broad exemptions from liability as long as they voluntarily investigate, detect, identify, and remove or disable access to illegal content[686]. Whereas hosting services and,

---

[681] Provisions on the Ecological Governance of Network Information Content, art 6.
[682] Provisions on the Ecological Governance of Network Information Content, art 11.
[683] Provisions on the Ecological Governance of Network Information Content, art 9.
[684] Chris Marsden, Trisha Meyer and Ian Brown, 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?' (2019) 36 Computer Law & Security Review.
[685] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1, arts 4–5.
[686] Andrej Savin, 'The EU Digital Services Act: Towards a More Responsible Internet' (papers.ssrn.com16 February 2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3786792> accessed 4 August 2025, 6.

particularly VLOPs and VLOSEs, are subject to procedural obligations[687], including establishing notice and action systems, transparent content review procedures, internal complaint mechanisms, and systematic risk assessments and independent audits under the DSA. For example, message pop-ups and advertising systems[688] trigger further risk mitigation requirements, including systematic risk assessments and transparency obligations[689].

The type of information is particularly decisive when assessing platform liability in relation to AI-generated disinformation. The DSA has a different approach to illegal content than to harmful but legal content[690]. Where AI-generated disinformation relates to illegal content, for example, deeply synthesized information created to defame individuals or manipulate the electoral process in breach of national law, the platform is liable if it has actual knowledge of the content but fails to act swiftly to remove it or disable access to it[691]. By contrast, the DSA does not impose strict removal obligations when AI-generated disinformation is lawful but harmful[692], such as fake news reports that mislead the public but do not violate specific legal requirements[693]. While it introduces systemic risk governance obligations, particularly for VLOPs, which are required to assess and implement proportionate mitigating measures for disinformation

---

687 Ibid 10.

688 Muhammad Ali and others, 'Discrimination through Optimization' (2019) 3 Proceedings of the ACM on Human-Computer Interaction 1 <https://dl.acm.org/doi/10.1145/3359301>.

689 Pieter Wolters and Frederik Zuiderveen Borgesius, 'The EU Digital Services Act: What Does It Mean for Online Advertising and Adtech?' (2025) 33 International Journal of Law and Information Technology.

690 Giancarlo Frosio, 'Platform Responsibility in the Digital Services Act: Constitutionalising, Regulating and Governing Private Ordering' (Queen's University BelfastOctober 2023) <https://pure.qub.ac.uk/en/publications/platform-responsibility-in-the-digital-services-act-constitutiona> accessed 30 August 2025.

691 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1, recs 28–29.

692 Aleksander Stawowy, 'User Content Moderation under the Digital Services Act – 10 Key Takeaways – Law Firm Traple Konarski Podrecki and Partners' (Law Firm Traple Konarski Podrecki and Partners24 October 2023) <https://www.traple.pl/en/user-content-moderation-under-the-digital-services-act-10-key-takeaways/> accessed 18 September 2025.

693 Peter Church and Ceyhun Necati Pehlivan, 'The Digital Services Act (DSA): A New Era for Online Harms and Intermediary Liability' (2023) 4 Global Privacy Law Review <https://kluwerlawonline.com/journalarticle/Global+Privacy+Law+Review/4.1/GPLR2023005> accessed 3 August 2025.

that may pose a systemic risk (Article 34) and provide corresponding mitigating measures (Article 35), as well as being subject to independent audits (Article 37). Thus, fault-based liability applies to knowingly unlawful disinformation, while harmful but lawful disinformation primarily involves risk management liability[694], rather than direct content management liability.

In the US, the primary statute that shields online platforms from liability for user-generated content (including AI-generated disinformation) is Section 230 of the CDA. This provision provides platforms with broad immunity from liability, meaning that they are not held liable for the content created by users, even if that content constitutes disinformation[695]. However, there are important exceptions to this immunity where platforms can be held liable for specific types of content. These exceptions primarily address situations where the content violates federal law, intellectual property rights, or is linked to certain criminal activities. The most notable exception is when the content infringes intellectual property rights, such as copyright infringement under the DMCA[696]. Following the enactment of the FOSTA-SESTA amendments[697] in 2018, the Section 230 immunity for platforms no longer covers content related to sex trafficking. These laws were a response to growing criticism that platforms were abusing Section 230 to evade liability for facilitating online sex trafficking, particularly the online platforms like Backpage[698] had been used to facilitate sex trafficking, where sites were accused of knowingly posting advertisements related to sex trafficking. By amending Section 230, Congress explicitly created an exception for sex trafficking, allowing federal and state authorities, as well as victims, to take legal action against platforms that knowingly aid, support, or facilitate such illegal activity. Unlike

---

694 Martin Husovec, 'The Digital Services Act's Red Line: What the Commission Can and Cannot Do about Disinformation' (2024) 16 Journal of Media Law 1.

695 Tomas A Lipinski, Elizabeth A Buchanan and Johannes J Britz, 'Sticks and Stones and Words That Harm: Liability vs. Responsibility, Section 230 and Defamatory Speech in Cyberspace' (2002) 4 Ethics and Information Technology 143.

696 17 USC § 512 (Digital Millennium Copyright Act of 1998).

697 Allow States and Victims to Fight Online Sex Trafficking Act of 2018, Pub L No 115–164, 132 Stat 1253 (2018) ('FOSTA-SESTA').

698 Jane Doe No 1 v Backpagecom LLC 817 F 3d 12 (1st Cir 2016).

disinformation, sex trafficking constitutes a serious criminal offense and a clear violation of fundamental human rights, thus justifying the removal of immunity in this area[699]. The Communications Decency Act itself provides that "nothing in this section shall be construed to impair the enforcement of… any other Federal criminal statute"[700]. This means that online platforms are not immune if they themselves engage in conduct that constitutes a federal crime, including fraud. The most relevant statute here is the Wire Fraud Act (18 USC §1343)[701], which criminalizes the use of interstate telecommunications communications (including Internet transmissions) to facilitate any fraudulent scheme. Courts have consistently held that an online platform will not be liable only because a user exploits its services to engage in fraudulent activity unless the platform materially contributes to the fraudulent scheme or knowingly participates in it[702]. Under the US's laws and regulations, exceptions to platform liability exemptions depend on the content of the information and the activities related to such disinformation. Unless it falls into the specific categories mentioned above, the platform is generally not liable for false information, even if it is harmful.

In China, the allocation of legal liabilities among online platforms depends largely on the type of services they provide and the type of information they process and disseminate.

First, the type of platform service determines the applicable compliance obligations. Platforms with different service models play different roles in social life, and the nature of the risks they are able to foresee and mitigate varies significantly, as reflected in the extent of the legal obligations imposed on them. For platforms that provide basic technical services such as network access and cloud computing, the Cybersecurity Law

---

[699] Elizabeth Carney, 'Protecting Internet Freedom at the Expense of Facilitating Online Child Sex Trafficking? An Explanation as to Why CDA's Section 230 Has No Place in a New NAFTA' (2019) 68 Catholic University Law Review 353 <https://scholarship.law.edu/lawreview/vol68/iss2/8/>.

[700] Communications Decency Act, 47 USC § 230(e)(1).

[701] 18 USC § 1343.

[702] Danielle Citron and Benjamin Wittes, 'The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity' (2017) 86 Fordham Law Review 401 <https://ir.lawnet.fordham.edu/flr/vol86/iss2/3/> accessed 4 August 2025.

only requires them to fulfill general cybersecurity protection obligations (Article 21) and the obligation to assist in law enforcement (Article 28), and assume basic liabilities[703]. Services that use algorithmic recommendation technology[704](such as using personalized push or search filtering algorithm technology to provide information to users) must comply with the Provisions on Administration of Algorithmic Recommendation in the Internet Information Service(2021)[705]. These regulations require platforms to fulfill algorithm transparency and explainability obligations (Article 12), provide a prominent mark for algorithmically synthesized information (Article 9), protect special groups such as minors (Article 18), and prohibit the use of recommendation algorithms to disseminate harmful or illegal content[706]. In addition, the Administrative Provisions on Deep Synthesis in Internet-based Information Services[707] apply to platforms that provide AI-generated content (including deepfakes), and consider them to no longer be neutral technical channels, but important participants in shaping and amplifying such risks[708]. They are required to assume governance responsibilities through rumor-busting mechanisms and security assessments. Most importantly, they must refrain from generating or disseminating false or illegal information. For e-commerce platforms (Articles 27 and 29 of the Electronic Commerce Law[709]) and internet app stores (Articles 6 and 7 of the Administrative Provisions on Information Services of Mobile Internet Application Programs[710]), their core responsibilities focus on vetting the qualifications of operators or application

[703] Cybersecurity Law of the People's Republic of China (2016), arts 21 and 28.
[704] Fei Yang and Yu Yao, 'A New Regulatory Framework for Algorithm-Powered Recommendation Services in China' (2022) 4 Nature Machine Intelligence 802 <https://www.nature.com/articles/s42256-022-00546-9>.
[705] Provisions on Administration of Algorithmic Recommendation in Internet Information Services (互联网信息服务算法推荐管理规定) (promulgated 31 December 2021, effective 1 March 2022).
[706] Provisions on Algorithmic Recommendation, arts 9, 12 and 18.
[707] Administrative Provisions on Deep Synthesis in Internet-based Information Services (互联网信息服务深度合成管理规定) (promulgated 11 November 2022, effective 10 January 2023).
[708] Xuanting Liu, 'Normative Construction of Platform Criminal Liability in the Governance of Deepfake Technology' (2025) 16 Advances in Social Behavior Research 40.
[709] Electronic Commerce Law of the People's Republic of China (2018), arts 27 and 29.
[710] Administrative Provisions on Information Services of Mobile Internet Application Programs (2016), arts 6–7.

providers within the platforms and, by requiring them to review and authenticate the real identity information provided by users, reducing the generation of disinformation at the source.

Secondly, for platforms providing the same service, China dynamically adjusts the intensity of legal obligations according to the type of information disseminated and processed, and implements a risk-based tiered management system. For illegal and harmful information that directly endangers national security and the public interest (such as terrorist and pornographic content, as per Article 21 of the Data Security Law[711]), platforms have the highest duty of care and must proactively monitor and remove it through technical means. For information that infringes on the civil rights of third parties, such as copyright, the platform mainly applies the provisions of Article 1195 of the Civil Code[712] and promptly deletes the infringing information after receiving notification from the right holder. Finally, for platforms that access and process personal privacy information, the Personal Information Protection Law(PIPL) imposes specific obligations at every stage of information dissemination, including notification and consent (Articles 13, 14, and 17), storage security protection (Article 9), and use and processing restrictions (Article 6)[713], as well as stricter protection obligations for important data.

In summary, China's online platform liability system establishes a multi-layered legal obligation framework through tiered governance based on service and information types. This aims to achieve precise and effective regulation, recognizing the functional and technological differences between platforms, thereby preventing small and medium-sized platforms from being unable to comply with regulatory obligations due to indiscriminate obligations. It also allows limited regulatory resources to be focused on areas of highest risk (such as national security and public interests).

---

711 Data Security Law of the People's Republic of China (2021), art 21.
712 Civil Code (2020), art 1195.
713 Personal Information Protection Law of the People's Republic of China (2021), arts 6, 9, 13, 14, 17.

3.5 Conclusion

This chapter focuses on the legal and regulatory provisions governing online platforms' liabilities for reviewing and regulating disinformation in three jurisdictions. First, I have described the shift in the role of online platforms, from neutral intermediary service providers to gatekeepers capable of proactively reviewing and controlling information, thereby emphasizing the necessity for online platforms to assume liability for content moderation. Second, I have reviewed the legal foundations of platform liability across three jurisdictions, examining the scope of legal liability borne by early internet platforms for content published on their services, and the purpose for which such legal provisions were established. Finally, I have analyzed how current laws and regulations define online platforms' regulatory liabilities for disinformation in these three representative jurisdictions, and their specific regulatory scope, review requirements, and detailed provisions. Furthermore, I have compared and analyzed the patterns and trends of legal regulatory changes in these three jurisdictions from the early days to the present and examines the primary factors behind these diverging patterns. Both the EU and China have enacted stricter and more specific laws or departmental regulations for the regulation of disinformation, and changing patterns in both jurisdictions is predictable, but the pattern of change in the US is flexible but unpredictable. Finally, I have focused on a comparative analysis of the attribution principles of platform liability. By examining the laws and regulations governing platform liability in three jurisdictions, I have demonstrated the attribution principles adopted, the factors influencing the choice of attribution principles, and the thresholds for triggering the duty of care.

4. Challenges in Enforcing Content Regulation by Platforms from a Cross-Jurisdictional Perspective

Overview

This chapter will explore the enforcement challenges faced by online platforms in selected jurisdictions in their efforts to regulate disinformation, including AI-generated disinformation, through content moderation. By examining the causes of enforcement difficulties, I will demonstrate how the limitations of laws and regulations in selected jurisdictions create difficulties in governing online platforms. Also, I will analyze and compares the liability attribution rules adopted by three jurisdictions for online platforms, demonstrating how different liability approaches impact the threshold for triggering the duty of care.

4.1 Common Issues in Implementation Across the EU, the US, and China

Despite significant differences in the legislative approaches and regulatory models for internet content governance in the EU, the US, and China, the three regions present some similar issues in the implementation of platform obligations.

While the EU's AI Act[714] and Digital Services Act introduce mandatory requirements for providing reasons for content removal, platforms still lack standardization in the method, content categories, and frequency of such disclosure. Trujillo and others' research reveals significant differences among platforms in content review methods, response frequency, and reasoning categorization, with platforms retaining significant discretion in both structure and content, resulting in a lack of consistency in disclosure practices[715].

---

[714] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L,1689/1.

[715] Amaury Trujillo, Tiziano Fagni and Stefano Cresci, 'The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media' (arXiv.org2023)

In China, regulations such as the "Administrative Provisions on Deep Synthesis in Internet-based Information Services[716]" mandate the distinctive labeling of deep synthetic information or services (Article 17) and provide users with a convenient channel for filing complaints (Article 12). However, different platforms have their own interpretations of "distinctive", and without a unified review standard for labeling, different platforms' labeling of deep synthesis information cannot completely avoid users' confusion. In addition, some online platforms lack clear explanations of the complaint process and feedback mechanism, resulting in the lack of practicality of their complaint channels[717].

Under the broad immunity provided by Section 230, US online platforms have the right to remove user-generated content, but this is not based on statutory obligations or subject to specific standards. Therefore, although major platforms (such as Google and X) voluntarily publish transparency reports, the disclosed indicators, format, content, and frequency are all determined by the platforms themselves, which leads to huge differences in the disclosed content between platforms[718].

4.2 EU: Unclear Implementation Standards Lead to Difficulties in Content Moderation

4.2.1 The Regulatory Scope of the EU Legal Framework on Disinformation

In the EU, the governance of disinformation could be achieved not only through the legal regulations imposed on platforms, but also by relying on internal moderation policies developed and implemented by online platforms themselves. The EU has established a comprehensive regulatory framework that combines legal regulations and

<https://arxiv.org/abs/2312.10269> accessed 14 April 2025.

[716] Administrative Provisions on Deep Synthesis in Internet-based Information Services (互联网信息服务深度合成管理规定) (promulgated 11 November 2022, effective 10 January 2023).

[717] Jun Liu, 'Internet Censorship in China: Looking through the Lens of Categorisation' (2024) 0 Journal of Current Chinese Affairs 1,2.

[718] Aleksandra Urman and Mykola Makhortykh, 'How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms' (2023) 47 Telecommunications Policy 102477.

self-regulation measures on platforms, providing institutional safeguards and practical paths for combating disinformation.

The EU's framework for governing disinformation is primarily based on mandatory obligations under the DSA, with the Code of Practice on Disinformation[719]serving as a soft-law complement to enhance and support its enforcement.

The DSA does not use the term "disinformation" in any legal provision that imposes an obligation on online platforms to detect or remove such content. In the main Articles, the DSA also does not specifically require online platforms to act against disinformation, but this does not mean the DSA ignores the negative impacts that disinformation has caused on society and democracy. For example, Recital 70[720] recognized that the generation and dissemination of disinformation would amplify societal harms, such as undermining the protection of public health or interfering with electoral processes. Recital 5[721] emphasizes that exponential growth in the use of intermediary services may also exacerbate their role in disseminating illegal or otherwise harmful content. This recital shows that DSA's management of information is not limited to illegal content but also pays attention to all harmful content.

Although DSA does not directly stipulate the governance of disinformation in its binding provisions, its regulations of illegal content and systemic risks can be applied to disinformation. Article 3(h) of DSA defines "illegal content" as "any information that, in itself or to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which complies with Union law, irrespective of the precise subject matter or nature of that law"[722]. It includes not only information that is illegal, such as hate speech or information related to terrorism, but also information associated with illegal activities, such as unauthorized

---

[719] European Commission, Code of Practice on Disinformation (2018) https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation accessed 25 July 2025.

[720] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1, recital 70.

[721] Digital Services Act, recital 5.

[722] Digital Services Act, art 3(h).

distribution of copyrighted works or the sale of counterfeit or substandard products. Regarding to the regulatory approach of illegal content, the DSA imposes mandatory obligations on platforms to take prompt and appropriate actions upon notification (via the notice-and-action mechanisms, Article 16), to provide users with "Statements of Reasons" when removing illegal information (Article 17), to prioritize reports from trusted flaggers(Article 22), and to publish transparency reports on removal and content moderation actions(Articles 15, 24, and 42).

In addition, DSA's requirements for assessing and mitigating systemic risks could be seen as provisions applicable to disinformation. Articles 33-35 stipulate the obligation of VLOPs and VLOSEs to assess and mitigate systemic risks and require them to impose mandatory and binding penalties on those who fail to take necessary measures or fail to comply with the regulations. According to the definition of systemic risk in Article 34, it can be inferred that if disinformation contains "illegal information, information that infringes on the basic rights of citizens, or content that may affect citizen discourse, election processes, and public safety", it can be included in the scope of regulation. In Recital 104[723], the DSA highlighted that systemic risks may include "disinformation" and other forms of manipulative or abusive activities and emphasized that when such information manipulation is used to obtain economic benefits, it is particularly harmful to vulnerable service recipients. In response to Article 35 requiring VLOPs and VLOSEs to formulate effective mitigation measures, Recital 88[724] believes that they should consider strengthening their internal procedures to supervise any activities and take corrective measures or other self-regulatory measures to reduce the risk of the disinformation campaign. Therefore, while the DSA does not specifically provide a clear definition or provisions for disinformation, the above-mentioned recitals and articles have stated that their applicability and governance objects include disinformation.

---

[723] Digital Services Act (n 2) recital 104.
[724] Digital Services Act (n 2) recital 88.

Online platforms could address AI-generated disinformation not only through legal obligations under the DSA but also by adapting voluntary frameworks[725], such as the Code of Practice on Disinformation[726] (hereinafter "2018 Code") and the 2022 Strengthened Code of Practice on Disinformation (hereinafter "2022 Code")[727]. These codes are voluntary instruments, whereby online platforms choose to become signatories and commit to a series of obligations aimed at mitigating the dissemination of disinformation. Although compliance with these codes is not mandatory under EU law, the 2022 code has evolved into a common regulatory took that complements the DSA. Based on the respect for fundamental rights such as freedom of expression, the EU has successively issued these two Codes as tools for self-regulation[728]. Essentially, these codes are industry guidelines developed by the EU Commission, online platforms, and other stakeholders, relying on platforms' voluntary compliance rather than legal enforcement[729]. The 2022 Code addresses the shortcomings of the 2018 Code in terms of enforceability and transparency by introducing specific commitments, supervision mechanisms, and coordination with the DSA. It defines disinformation as "verifiably false or misleading information that is created or disseminated for economic gain and may cause public harm", focusing on its intentionality, falsity, and potential harm to develop economic and operational strategies to reduce its spread[730]. Besides, the 2022 Code encourages signatories to implement specific measures, including detecting, labeling, and demoting synthetic or distorted content, collaborating with fact-checkers to ensure the transparency of algorithmic systems that may amplify false or misleading

---

[725] Kirsty Park and Eileen Culloty, 'BEYOND PERFORMATIVE TRANSPARENCY: LESSONS LEARNED from the EU CODE of PRACTICE on DISINFORMATION' [2023] Selected Papers of Internet Research.

[727] European Commission, 2022 Strengthened Code of Practice on Disinformation (16 June 2022) https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation accessed 21 July 2025.

[728] Paula Gori, 'The Strengthened Code of Practice on Disinformation – Many Stakeholders, One Goal - MediaLaws' (MediaLaws9 January 2023) <https://www.medialaws.eu/the-strengthened-code-of-practice-on-disinformation-many-stakeholders-one-goal/> accessed 10 June 2025.

[729] Mündges and Park (n 464) 3.

[730] Ibid 14.

AI-generated content[731].

4.2.2 Specific Legal Provisions and Issues They Seek to Address

The EU has not established mandatory measures for all online platforms to detect, report, and delete disinformation, including AI-generated disinformation[732]. However, it has established a series of detailed obligations for VLOPs to prevent, mitigate, and manage systemic risks. Furthermore, in addition to mandatory legal provisions, online platforms can also acquire the right to moderate content through their Terms of Services agreements with users to protect their users' interests and safeguard their own reputations.

DSA's requirements for content moderation by online platforms mainly focus on systemic risks, and continuously improving the platform's future practice guidelines through the process of defining systemic risks (Art.34(1)(2)), imposing content moderation and risk mitigation obligations on platforms (Art.34(3) and Art.35(1)), and assessing platforms' measures by independent auditors (Art.37)[733]. Although there are no uniform obligations for all platforms at the EU level, the DSA sets out targeted compliance requirements for specific platforms. Article 34 clearly imposes obligations on VLOPs and VLOSEs to conduct annual risk assessments and implement effective measures to address systemic risk assessment[734]. The platforms identified as VLOPs should identify, analyze, and assess the systemic risks that may be caused by their own system design, algorithmic mechanisms, or functions within the EU. Article 34(1) summarizes the systemic risks that should be detected and regulated, requiring platforms to pay attention to identifying their sources when conducting risk evaluation.

---

[731] Peter H Chase, Senior Fellow and The German Marshall Fund of the United States, 'The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem †' (2019) <https://www.ivir.nl/publicaties/download/EU_Code_Practice_Disinformation_Aug_2019.pdf>.

[732] Koen Vranckaert, 'Disinformation as a Cyber Threat under EU Law: Which Approach to Take in the Age of AI?' (Faculteit Rechtsgeleerdheid En Criminologische Wetenschappen2024) <https://www.law.kuleuven.be/ai-summer-school/blogpost/Blogposts/disinformation-as-a-cyber-threat-under-eu-law-which-approach-to-take-in-the-age-of-ai> accessed 21 July 2025.

[733] Rebecca Tushnet, 'A Hobgoblin Comes for Internet Regulation' (VerfBlog (short for Verfassungsblog)19 February 2024) <https://verfassungsblog.de/a-hobgoblin-comes-for-internet-regulation/> accessed 28 June 2025.

[734] Digital Services Act, art 34(1).

The results of this assessment would form an important basis for compliance audits (Article 37), regulatory moderation, and potential enforcement actions. Disinformation generated by AI, although not explicitly included in systemic risks, due to its scalability, synthetic nature, and deceptiveness, has the potential to cause risks falling under Article 34(1)(b) and (c) and should also be included in the scope of systemic risk mitigation obligations required by Article 34.

At the same time, DSA provides that VLOPs and VLOSEs keep supporting documentations of their risk assessments[735] and publish a comprehensive report annually, focusing on summarizing the most prominent and recurring systemic risks in the EU member states[736]. Additionally, Article 45(2) stipulates that in cases of significant systemic risks involving multiple VLOPs and/or VLOSE, the EU Commission encourages the involvement of stakeholders at the union level in developing a code with specific risk mitigation measures and a regular reporting framework. According to Articles 35(2)(b) and (3) of the DSA, the European Commission may, in cooperation with Digital Service Coordinators, develop feasible guidelines on risk mitigation based on the risk assessment reports provided by online platforms and, if necessary, require the VLOPs or VLOSEs to provide alternative measures. If the platform fails to fulfill its corresponding obligations, the Commission may also impose an administrative fine of up to 6% of its global annual turnover on it under Article 74. To assess whether the platforms have effectively identified, evaluated, and mitigated systemic risks (such as disinformation or algorithmic hazards), DSA requires VLOPs and VLOSEs to proactively undergo independent audits and improve their practices based on objective evaluations. Article 37 requires VLOPs and VLOSEs to undergo a compliance audit at least once a year by a qualified and independent auditor with no conflict of interest[737]. The audit agencies not only conduct a formal review of the risk assessment report that platforms submit, but also comprehensively

[735] Digital Services Act, art 34(3).
[736] Digital Services Act, art 35(2).
[737] Digital Services Act, art 37.

assess whether VLOPs and VLOSEs have fulfilled their systemic risk management obligations. The audit report should assess the platform's compliance with all due diligence obligations mentioned in Chapter 3 and provide recommendations for improvement on the specific measures conducive to achieving compliance. Besides, the VLOPs and VLOSEs under review should, upon request of the institution and committee, grant access to the required data within a reasonable period to conduct research that helps discover, identify, and understand systemic risks and evaluate risk mitigation measures[738].

Procedural fairness is a crucial aspect of the governance design of online platforms[739], which is reflected in the public's perception of legitimacy and transparency of the platform's exercise of its right during the review process[740]. In the context of content moderation, procedural fairness ensures that online platforms make and implement their review and management of online information in a way that is transparent, consistent, unbiased, and respectful of users' rights[741]. The regulatory obligations imposed on online platforms by laws or regulations are an important means to ensure procedural fairness, aiming at ensuring fairness and transparency of the content review process. These requirements include notifying users when content is removed or blocked, providing a statement of reasons, offering an opportunity to appeal the decision, and implementing moderation decisions within a reasonable time[742].

DSA emphasizes that online platforms or search engines must report their content moderation decisions, as well as follow-up notifications, appeals, and other activities

---

[738] Digital Services Act, art 40(1).

[739] Nicolas P Suzor and others, 'What Do We Mean When We Talk about Transparency? Toward Meaningful Transparency in Commercial Content Moderation' (2019) 13 International Journal of Communication 1526 <https://ijoc.org/index.php/ijoc/article/view/9736>, 1538.

[740] Yunhee Shim and Shagun Jhaver, 'Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms' (Arxiv.org2025) <https://arxiv.org/html/2409.08498v1#bib.bib98> accessed 6 June 2025.

[741] Renkai Ma and Yubo Kou, '"I'm Not Sure What Difference Is between Their Content and Mine, Other than the Person Itself"' (2022) 6 Proceedings of the ACM on Human-Computer Interaction 1.

[742] Dawn Carla Nunziato, 'The Digital Services Act and the Brussels Effect on Platform Content Moderation | Chicago Journal of International Law' (cjil.uchicago.edu2024) <https://cjil.uchicago.edu/print-archive/digital-services-act-and-brussels-effect-platform-content-moderation> accessed 6 June 2025.

in response to these results. Article 17 focuses on individual-level transparency, ensuring users receive the "statements of reasons" that are clear and specific explanations when posted information is restricted, removed, or otherwise moderated. Article 24 emphasizes the system's transparency, requiring online platforms and search engines to regularly publish reports that increase the transparency in their content moderation practices and dispute resolution processes. While considering the data privacy and interest protection of VLOPs and VLOSEs, the content involving personal data or commercial secrets would not be disclosed to maintain the security of their services[743].

In addition to legal requirements, online platforms could also obtain the right to detect, identify, and review user-generated content through their Terms of Service (ToS), which users agree to abide by. Article 14 of DSA empowers intermediary service providers to establish their terms and conditions, imposing any restrictions on the information provided by the recipients of the service, including restrictions on any policies, measures, or tools used for content moderation[744]. As the contractual basis for the relationship between online platforms and their users, the ToS, once agreed to by users, constitute a binding agreement between both parties. Consequently, online platforms have the right to make decisions regarding content removal or account management based on the ToS.

While online platforms generally have the freedom to set their terms of use, the DSA sets out some basic rules regarding the content and enforcement of these terms to protect users' rights, increase the transparency of enforcement, and prevent unfairness resulting from unilateral or disproportionate platform practices[745]. For example, Article 14(5) emphasizes that online platforms should explain any restrictive information in their ToSs to users in a clear, user-friendly, and unambiguous language. When drafting

---

[743] Digital Services Act, art 40(2).
[744] Digital Services Act, art 14(1).
[745] João Pedro Quintais, Naomi Appelman and Ronan Ó Fathaigh, 'Using Terms and Conditions to Apply Fundamental Rights to Content Moderation' (2023) 24 German Law Journal 1.

and applying their terms and conditions, platforms should do so with due respect for users' fundamental rights, such as safeguarding their freedom of expression and the right to effective remedies. Besides, while DSA does not directly interfere with the platform's right to write its terms and conditions, its requirement for VLOPs to assess systemic risks limits the online platform's discretion in customizing its terms of service.

The 2022 Code introduces a co-regulatory framework that requires online platforms to proactively detect, identify, and mitigate disinformation. Commitments 1 to 5 on demonetizing disinformation stipulate that parties involved in advertising sales must not subsidize the spread of disinformation[746]. Especially, the online platforms acting as advertising carriers must deny advertising revenue to actors who repeatedly disseminate known disinformation or misleading content. This measure targets the economic incentives behind the disinformation activities, aiming to reduce the profitability of the creation and spread of disinformation online.

Under Commitments 14 to 16 (Integrity of Services), platforms are required to adopt clear policies for identifying and restricting manipulative behaviors and practices commonly associated with the spread of disinformation, such as the creation and use of fake accounts, malicious deepfakes, and coordinated inauthentic behavior[747]. These obligations are directly linked to disinformation regulation, as such tactics are frequently used to amplify false and misleading content or impersonate legitimate sources[748].

Additionally, Commitments 17 to 25 empower users to identify and report disinformation or misleading content, recognizing the importance of user engagement and provenance technology as tools for understanding and accessing disinformation[749]. For example, Measure 22.1 requires platforms to display credibility labels on content

---

[746] Strengthened Code of Practice on Disinformation (2022), Commitments 1–5.
[747] Strengthened Code of Practice on Disinformation (2022), Commitments 14–16.
[748] Richard Wingfield, 'A Human Rights-Based Approach to Disinformation | Global Partners Digital' (Global Partners Digital15 October 2019) <https://www.gp-digital.org/a-human-rights-based-approach-to-disinformation/>.
[749] Strengthened Code of Practice on Disinformation (2022), Commitments 17–25.

verified by independent fact-checkers to help users make informed choices[750]. Measure 22.7, on the other hand, encourages such labels to appear in more prominent forms, such as banner ads and pop-ups[751]. These provisions ensure that disinformation is not only detected but also contextually marked to limit its impact and empower user discernment.

Overall, these commitments represent a shift for online platforms from passive content removal to proactively taking responsibility for content moderation, addressing disinformation through detection requirements, increased transparency, and reducing economic incentives.

4.2.3 An Evaluation of the Effect of the Legal Regulations

The DSA's ambiguous definition of "systemic risk" leads to a lack of unified standards for platforms' content review and management. While VLOPs and VLOSEs require independent audits of their implementation, the auditor's employment relationship with the platform and the platform's protection of private data may undermine the audit's independence. Meanwhile, online platforms manage user-generated content based on their terms of service, but this approach is influenced by factors such as the platform's core business and external oversight policies, resulting in varying regulatory priorities. DSA emphasizes procedural obligations rather than substantive ones, meaning it focuses on how platforms manage risks rather than defining exactly what content is harmful or illegal in each instance. It does not provide concrete and substantive criteria for what constitutes a “systemic risk”. Article 34 lists four different but broad types of risks, the list that read more like a broad enumeration of common concerns for platforms than a clear regulatory framework. The lack of precise definitions and measurable standards makes online platforms difficult to identify systemic risks in practice, especially in the context of emerging threats, such as AI-generated disinformation, which may be harmful but not necessarily illegal. According to the requirements of

[750] Strengthened Code of Practice on Disinformation (2022), Measure 22.1.
[751] Strengthened Code of Practice on Disinformation (2022), Measure 22.7.

DSA, the existence of systemic risk is a preliminary condition for VLOPs and VLOSEs to conduct regulatory measures[752]. The systemic risks are defined in the DSA as significant risks arising from the design, operation, or use of VLOPs and VLOSEs. Systemic risks mainly include the following four categories: the dissemination of illegal content, actions that have a potential or actual negative impact on the exercise of fundamental rights(such as data protection, freedom of expression, and consumers' protection), and the deliberate manipulation of online platforms to undermine the democratic processes, public safety, and protection of mental and physical health of the public[753]. Although DSA has stipulated these four types of risks as "systemic risks" to be assessed and prevented, there is ambiguity around the core concept of this term, which leads to different views among VLOPs in their practice of risk assessment and management[754].

First, there are two perspectives on the criteria for determining whether a risk is "systemic". The first view holds that the understanding of systemic risk depends on the scope of its impact, which would pose a threat to the wider social structure[755]. The requirement of risk coverage is reflected in the following aspects: the expansion of potential harm caused by the cross-platform dissemination of disinformation; the potential for causing significant impacts at the social level, such as interference with electoral processes or emergencies; and the potential for impacts on multiple interrelated forms of fundamental rights. This outcome-oriented interpretation[756] aligns with Article 34(1) of the DSA, which defines systemic risk as foreseeable

---

[752] Claire Stravato Emes, 'Exploring New Frontiers in Digital Governance: Addressing the Ambiguities of Risk-Based Regulation Approach for Platforms' <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5242418> accessed 4 August 2025.

[753] Beatriz Botero Arcila , 'Systemic Risks in the DSA and Its Enforcement' (DSA Decoded2024) <https://www.dsadecoded.com/systemic-risks-in-the-dsa-and-its-enforcement> accessed 4 August 2025.

[754] Luca Nannini and others, 'Beyond Phase-In: Assessing Impacts on Disinformation of the EU Digital Services Act' (2024) 5 AI and Ethics.

[755] David Sullivan and Jason Pielemeier, 'Unpacking "Systemic Risk" under the EU's Digital Service Act' (Tech Policy Press19 July 2023) <https://www.techpolicy.press/unpacking-systemic-risk-under-the-eus-digital-service-act/> accessed 26 June 2025.

[756] Magdalena Jóźwiak, 'The DSA's Systemic Risk Framework: Taking Stock and Looking Ahead' (Dsa-observatory.eu2025) <https://dsa-observatory.eu/2025/05/27/the-dsas-systemic-risk-framework-taking-stock-and-looking-ahead/> accessed 15 July 2025.

negative effects on the public interest. The second view emphasizes the causes of risk, arguing that systemic risk not only arises from the content itself, but also from the design and operation mechanism of online platforms[757]. For example, harmful content automatically pops up through the algorithmic recommendation mechanism[758], which is not the content actively retrieved by the users; the existence of a regulatory vacuum when users share information across platforms makes it difficult for regulators to track the spread of false information. This perspective shows that the systemic risks may stem from the structured problems within the online platform, rather than merely from users' behavior or individual content. This cause-oriented opinion is supported by Recital 84 and Article 34(2), which require platforms to focus on the likelihood that their recommendation and advertising systems would spread deceptive information, and whether their algorithmic systems would amplify the systemic risks, when conducting risk assessment[759]. While both interpretations are supported in the text of the DSA, their coexistence leads to uncertainty in regulatory compliance and enforcement.

Secondly, regarding the ambiguity in the definition of "systemic risk", neither the European Commission nor online platforms is in a position to bear sole responsibility for providing additional clarification on the concept. Although DSA's Article 34 imposes relevant identification and governance obligations on VLOPs and VLOSEs, the self-assessment of platforms under the profit-oriented business logic is prone to conflicts of interest, and it is difficult to ensure the objectivity and credibility of their assessment results[760]. The transparency and risk assessment reports provided by the VLOPs may lack credibility to fill the compliance gap caused by the vague definition

---

[757] Amélie P Heldt, 'EU Digital Services Act: The White Hope of Intermediary Regulation' [2022] Palgrave Macmillan 69.

[758] Svea Windwehr, 'Systemic Risk Reporting: A System in Crisis?' (Electronic Frontier Foundation16 January 2025) <https://www.eff.org/deeplinks/2025/01/systemic-risk-reporting-system-crisis> accessed 15 July 2025.

[759] Tarleton Gillespie, Pablo J Boczkowski and Kirsten A Foot, 'The Relevance of Algorithms' in Tarleton Gillespie and others(eds), Media Technologies: Essays on Communication, Materiality, and Society (MIT Press 2013) 167.

[760] David Sullivan, 'Systemic Risk Assessments Hold Clues for EU Platform Enforcement' (Lawfare2025) <https://www.lawfaremedia.org/article/systemic-risk-assessments-hold-clues-for-eu-platform-enforcement> accessed 15 July 2025.

of systemic risk. It would also be inappropriate for the EU's public bodies, such as the Digital Commission or a national institution undertaking the role of digital service coordinator, to take the role of regulating systemic risks. While public authorities can intervene to limit the discretion of online platforms, such interventions have the potential for the State to indirectly shape or control the public discourse[761]. The establishment of content moderation standards involves the protection of free expression and democratic participation[762]. Therefore, giving such institutions significant power over how online platforms review and curate content is bound to raise several justified concerns, as it may blur the boundary between platform regulation and speech regulation, potentially manipulating the public discussion.

Unlike providing a precise definition directly, DSA tends to create a "virtuous loop mechanism" for evaluating and recalibrating a platform's risk assessment and mitigation efforts when a new potential emerges[763]. The DSA adopts a co-regulatory approach, which relies on ongoing collaboration between online platforms, regulators, civil society organizations, and researchers to continually develop and improve the understanding and detection criteria of systemic risks[764]. While this framework provides regulatory flexibility and adaptability to address emerging hazards such as AI-generated disinformation, it also creates significant uncertainty regarding platform compliance. In the absence of uniform standards, different platforms may interpret systemic risks differently, leading to inconsistent enforcement. This definitional ambiguity poses practical challenges to measuring platform compliance, assessing the effectiveness of risk mitigation measures, and ensuring regulatory accountability[765]. In

---

[761] Eder (n 670) 8.

[762] Laura Fichtner, 'Content Moderation and the Quest for Democratic Legitimacy' (2024) 4 Weizenbaum Journal of the Digital Society <https://ojs.weizenbaum-institut.de/index.php/wjds/article/view/2_2_2>.

[763] Niklas Eder, 'Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation' (2024) 25 German Law Journal 1.

[764] Martin Husovec, 'The DSA as a Co-Regulatory System' [2024] Oxford University Press eBooks 443 <https://academic.oup.com/oxford-law-pro/book/58088/chapter/478884189#515183574> accessed 28 June 2025.

[765] Rachel Griffin, 'Governing Platforms through Corporate Risk Management: The Politics of Systemic Risk in the Digital Services Act' [2025] European Law Open 1 <https://www.cambridge.org/core/journals/european-law-open/article/governing-platforms-through-

particular, for new forms of harm such as AI-generated disinformation, which generally do not fall into the clear category of illegal content, the ambiguity of systemic risk assessments further exacerbates the complexity of content governance under the DSA framework.

Although DSA attempts to enhance the transparency and ensure accountability of platform regulation by introducing a mandatory independent audit mechanism, the mechanism faces many structural problems in practice. First, the DSA has not established sufficiently clear technical and methodological standards for auditing VLOPs and VLOSEs, which makes it easy for the audit process to become a formality. Independent auditing agencies under the requirements of DSA usually focus on confirming whether VLOPs and VLOSEs' treatment of systemic risks that appeared on these platforms complies with company policies, industry standards, and legal regulations[766]. Meßmer and Degeling point out that the DSA does not provide sufficiently specific guidelines for platform audit, and the lack of uniform assessment metrics to guarantee the implementation of the effective audit process allows platforms to potentially use audits as a means of legitimizing their operations and avoiding substantial corrective action[767]. Secondly, platform auditing is a new field that requires auditors to have highly sophisticated skills, but these required skills and resources need to be built up through formal training and years of work in a particular industry, and are therefore difficult to acquire quickly[768]. Besides, these audits require auditors to have expertise in multiple fields, such as content moderation, digital rights protection, and recommendation system governance[769]. This professional requirement significantly

---

corporate-risk-management-the-politics-of-systemic-risk-in-the-digital-services-act/287159FD68134232851133FEFF451D42> accessed 15 July 2025.

[766] Inioluwa Deborah Raji and others, 'Closing the AI Accountability Gap: Defining an End-To-End Framework for Internal Algorithmic Auditing' [2020] Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 33 <https://dl.acm.org/doi/abs/10.1145/3351095.3372873>.

[767] Anna-Katharina Meßmer and Martin Degeling, 'Auditing Recommender Systems -- Putting the DSA into Practice with a Risk-Scenario-Based Approach' (arXiv.org2023) <https://arxiv.org/abs/2302.04556?.com> accessed 30 June 2025.

[768] Claire Pershan, 'Cutting through the Jargon - Independent Audits in the Digital Services Act' (Mozilla Foundation30 January 2023) <https://www.mozillafoundation.org/en/blog/cutting-through-the-jargon-independent-audits-in-the-digital-services-act/> accessed 30 June 2025.

[769] Johann Laux, Sandra Wachter and Brent Mittelstadt, 'Taming the Few: Platform Regulation,

increases the complexity of audit work[770]. Third, the audit agency's dependence on the platform may affect the independence of the audit. DSA requires VLOPs and VLOSEs to commission independent auditors to conduct systemic risk audits every year, with the cost of the audit borne by the platform[771]. This audit system is at risk of being "audit captured" by the platform, that is, the platform uses its own market influence to exert indirect influence on the audit institutions that rely on its business[772]. This mechanism means that these platforms naturally become the main demand side of the audit service market, affecting the auditors' motivation to audit platform activities[773]. Due to the demand for industry-specific information, scarcity of specialization resources, and economic incentives, regulators may tend to serve the interests of the industry they regulate[774]. Although DSA requires the audited platform to provide “access to all relevant data and premises”, information asymmetries between online platforms and third-party auditors continue to play a role[775]. The platform may refuse to provide all the required information to the audit institution on the grounds of maintaining information confidentiality, such as trade secrets or core technology patents[776]. Audit institutions need to establish a mutually beneficial cooperative relationship with the regulated companies and expect the platform to achieve cooperation by disclosing information[777]. This demand and supply relationship may weaken the independence of auditors and affect the quality of audits, because audit institutions have the motivation

---

Independent Audits, and the Risks of Capture Created by the DMA and DSA’ (2021) 43 Computer Law & Security Review 105613.

[770] Daniel Holznagel, ‘Shortcomings of the First DSA Audits — and How to Do Better - DSA Observatory’ (DSA Observatory - a Hub of Expertise on the DSA package.11 June 2025) <https://dsa-observatory.eu/2025/06/11/shortcomings-of-the-first-dsa-audits-and-how-to-do-better/> accessed 30 June 2025.

[771] Digital Services Act, art 37(1).

[772] Laux, Wachter and Mittelstadt (n 769) 8.

[773] Ibid.

[774] Luigi Zingales, ‘Preventing Economists’ Capture’ [2013] Preventing Regulatory Capture 124.

[775] Susan J Smith and Marja Elsinga, International Encyclopedia of Housing and Home (Elsevier 2012).

[776] Marie-Therese Sekwenz and others, ‘Doing Audits Right? The Role of Sampling and Legal Content Analysis in Systemic Risk Assessments and Independent Audits in the Digital Services Act’ (arXiv.org2025) <https://arxiv.org/abs/2505.03601> accessed 1 July 2025, 16.

[777] Siddhant Chatterjee, ‘Rules for Independent Audits under the EU’s Digital Services Act (DSA)’ (Holisticai.com2023) <https://www.holisticai.com/blog/rules-for-independent-audits-digital-services-act> accessed 1 July 2025.

to cater to the platform's preferences to ensure future cooperation and revenue sources. Similarly, auditors and researchers under the DSA may form similar dependencies to gain access to platform-specific data or to maintain collaborative relationships, reducing their objectivity and criticality in risk identification and audit reporting[778]. In the absence of professional capabilities and independence guarantees, the current audit mechanism is unlikely to assume the function of ensuring that platforms fulfill their obligations to assess and mitigate systemic risks. Therefore, unless the audit process is institutionally revised, the vision of building a platform accountability system based on independent audits may be difficult to achieve.

Online platforms' terms and conditions specify not only the types of content that are prohibited from being updated and distributed on platforms, but also the measures that will be taken if such a violation is detected[779]. Different platforms will have distinctive rules and priorities based on their targeted audience, main function, legal obligations in specific jurisdictions, company policies, or the code they have signed up to. Arora et al. conducted a comparative analysis of the publicly available content guidelines of 42 representative online platforms, examining the scope of harmful content categories that are designated for removal[780]. While this article's scope extends beyond disinformation, its findings provide important insights into the inconsistencies in platform content moderation standards and diversity of regulatory scope, all of which are directly relevant to the regulatory treatment of disinformation. Based on this comparative study, it is evident that while different platforms share the same approaches to addressing widely recognized categories of harmful content[781], they also demonstrate significant differences in their focus on specific types of online harms.

---

[778] Sekwenz and others (n 776) 6.

[779] Mohit Singhal and others, 'SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice' [2022] arXiv:2206.14855 [cs] <https://arxiv.org/abs/2206.14855> accessed 14 April 2025.

[780] Arnav Arora and others, 'Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go' (2023) 56 ACM Computing Surveys 1.

[781] Semiu Salawu, Yulan He and Joanna Lumsden, 'Approaches to Automated Detection of Cyberbullying: A Survey' (2017) 11 IEEE Transactions on Affective Computing 1.

Firstly, platforms with different functions present personalized terms and conditions. For the platforms that focus on E-commerce and App distribution (such as Amazon and Apple), they emphasize their ToS on rules regarding product quality, copyright infringement, payment fraud, as well as licensing and authorization of apps. While for social media platforms (such as Facebook, Twitter, Google), they provide extensive and detailed ToS that prioritize the regulation of harmful content related to users' speech and sharing of personal imagery[782], including hate speech, harassment, disinformation, graphic content, self-harming and violent content, sexual exploitation, and spam[783]. This is especially evident in the platform categories centered on interpersonal communication, such as dating apps, particularly where such content poses risks to individual safety or dignity. For example, "Bumble" typically prohibits sexual solicitation, banning the posting of explicit sexual content, harassment, or unsolicited sexual advances, empowering users to report violating content for removal[784]. Besides, as these social media platforms primarily facilitate the user's expression and online interactions, their terms of service particularly focus on regulating risks associated with personal expressions, sexual content, and non-consensual imagery, while striking a balance between freedom of speech and combating harmful content[785]. For online platforms that provide specific services, their ToS are usually formulated for single or limited services and products. In addition to clearly defining its function, access permission, and scope of application, it also formulates disclaims for specific risks related to these services and products[786]. For forums that provide specific services or products, they adapt their rules to the risks associated with their domains, public opinion,

---

[782] Kim Barker, Olga Jurasz and Stirling Law School, 'Online Harms White Paper Consultation Response' (2019) <https://oro.open.ac.uk/69840/1/Barker%20&%20Jurasz%20-%20Online%20Harms%20White%20Paper%20Consultation%20Response%20(2019)%20.pdf> accessed 18 June 2025.

[783] Fiona Dennehy, 'Almost 90% of Young People Exposed to Harmful Content on Social Media' (The Alan Turing Institute2023) <https://www.turing.ac.uk/news/almost-90-young-people-exposed-harmful-content-social-media> accessed 18 August 2025.

[784] Bumble, 'Bumble's Community Guidelines | Bumble' (Bumble2025) <https://bumble.com/en/guidelines?.com> accessed 21 June 2025.

[785] Arora and others (n 780) 2.

[786] F Lagioia and others, 'AI in Search of Unfairness in Consumer Contracts: The Terms of Service Landscape' (2022) 45 Journal of Consumer Policy 481.

and regulatory pressure. Arora and collogues compared the types of content that are prohibited from being uploaded in the ToS of different platforms, trying to explore whether the platform's main business is related to the strictness of its content review policy. According to the number of policy topics covered by each platform per category[787], despite the narrow topic range, the terms and conditions coverage of gaming forums is more comprehensive than that of general-purpose forums. The main reason for this difference is the high attention from the media and regulatory authorities, as well as the forums' previous failure to take sufficient measures to moderate harmful content, which forced these platforms to adopt a stricter management framework[788]. In contrast, finance forums exhibit the least extensive coverage, likely due to limited attention from the public and regulators[789]. These illustrate that the scope and intensity of content moderation are determined not only by the core functionality or main business of online platforms but are also significantly influenced by external factors such as public discourses or regulatory policies.

The platform-specific variation in implementation is particularly evident when assessing how platforms meet their transparency obligations under the DSA. For example, the provision of SoR reveals a counterintuitive trend that the number of SoR provided by online platforms to users whose content has been removed or restricted is not proportional to their daily activity. The research conducted by Kaushal and others shows that, while all the VLOPs engage in content moderation practices, the number of Statements of Reasons (SoRs) is not proportional to the number of monthly active users on these platforms in Europe[790]. For example, given the volume of data submitted by VLOPs and the share of each platform's data across the datasets, Google Shopping accounts for more than half (52.2%) of the SoRs, while those online platforms with

---

[787] Arora and others (n 780) 5.

[788] BBC News, 'TikTok and Twitch Face Fines under New Ofcom Rules' BBC News (5 October 2021) <https://www.bbc.co.uk/news/technology-58809169> accessed 18 August 2025.

[789] Arora and others (n 780) 4.

[790] Rishabh Kaushal and others, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database' (arXiv.org2024) <https://arxiv.org/abs/2404.02894> accessed 11 June 2025.

more users, such as TikTok, Amazon, and Facebook, submit less[791]. This suggests that there are differences in the implementation standards of DSA's content moderation obligations among different platforms, resulting in platforms with high activity volumes providing fewer SoRs.

4.3 US: Operational challenges of disinformation moderation in a deregulated environment

4.3.1 US legal framework for content governance on online platforms

In the US, the governance of disinformation by online platforms is shaped primarily by two legal instruments: the First Amendment of the US Constitution and Section 230 of the CDA. The First Amendment only restricts state actors from interfering with free speech and does not apply to private entities such as online platforms. This distinction is significant because it allows the private platforms to remove or correct user-generated content without violating the constitutional protection of free speech. Complementing this constitutional structure, section 230 of the CDA provides online platforms with immunity from liability and broad discretion to moderate content. Online platforms are not held liable for illegal or harmful user-generated content and can voluntarily take "good faith" action to limit access to information they deem objectionable. The disinformation generated by users falls within the exemptions from platform liability. This statutory immunity encourages online platforms to actively engage in content moderation while shielding them from potential legal consequences resulting from the removal of information. Together, the First Amendment and Section 230 create a governance framework that empowers private platforms to serve as actual regulators of online speech.

As Jack Balkin has pointed out, this regulatory framework reflects a "new school" of

[791] European Commission, 'Supervision of the Designated Very Large Online Platforms and Search Engines under DSA | Shaping Europe's Digital Future' (digital-strategy.ec.europa.eu2025) <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses> accessed 18 August 2025.

speech regulation, where private online platforms, rather than governments, determine the boundaries of permissible expression[792]. The "new school" is an approach to regulation in which governments no longer directly constrain the subject of expressions, but rather target online platforms by threatening liability[793], offering incentives, or cooperating with them to fulfill content moderation functions[794]. When online platforms conduct content moderation, it often takes the form of digital prior restraint[795], and even in the absence of a formal injunction, platforms would delete the user's posts. These platforms proactively develop terms and conditions of services in internet communities, filtering and selecting user-generated content through algorithmic detection and reactive content moderation systems[796]. Under this model, users need to obtain the permissions from platforms to post their content, as opposed to the "old school" regulation that users need to bear risks when expressing their opinions[797]. However, this regulatory framework raises significant concerns about the accountability of disinformation, as online platforms are neither constitutionally obliged to uphold free speech nor legally required to provide clear justifications for their content moderation decisions.

The First Amendment of the US explicitly provides the protection of free speech, requiring the "Congress shall make no law… abridging the freedom of speech, or of the press…". In Hudgens v. NLRB[798], the court has determined that only state actions could create an affirmative obligation under the First Amendment, which means the Constitution only limits government actions. In Marsh v. Alabama, the Court

---

[792] Jack M Balkin, 'Old School/New School Speech Regulation' (Ssrn.com2014) 2296 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2377526> accessed 4 August 2025.
[793] Christina Mulligan, 'Technological Intermediaries and Freedom of the Press' (2013) 66 SSRN Electronic Journal.
[794] Balkin (n 792) 2299.
[795] Thomas I Emerson, 'The Doctrine of Prior Restraint' (1955) 20 Law and Contemporary Problems 648 <https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=2658&context=lcp> accessed 10 July 2020.
[796] Emma J Llansó, 'No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem' (2020) 7 Big Data & Society 1.
[797] Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7 Big Data & Society 1 <https://doi.org/10.1177/2053951720943234>.
[798] Hudgens v. NLRB, 424 U.S. 507 (1976) 513.

distinguished between private and state actors and held that it is difficult to determine when the actions of private parties constitute state action and assume public functions[799]. In Cyber Promotions v. American Online[800], by comparing the activities of AOL and Marsh, the court held that private platforms, particularly those that provide hosting or search engine services, do not assume any traditional municipal powers or indispensable public service functions[801]. The court clarified that the Internet is a global network that is not placed under the exclusive control of the government, and its actions are not governmental, so the First Amendment does not apply to private companies such as Facebook, Twitter, or YouTube[802]. While users may believe that content removal or account banning violates their free speech, private platforms are legally entitled to enforce their terms of service, and such moderation actions do not constitute a violation of constitutional free speech, which applies only to state actors[803]. Very Large online platforms with digital infrastructures have the technical capacity to control, filter, or delete false content, to monitor access to their device, and to manage user-generated content[804]. While private online platforms in the US have no constitutional obligation to comply with the First Amendment, many have adopted content policies aligned with their values, such as encouragement of diverse expressions and providing compliant mechanisms[805]. These practices are not legal requirements, but are intended to maintain their reputations and user trust. But online platforms also remove harmful content, such as hate speech or disinformation, for business or ethical reasons[806].

---

[799] Sheila Kennedy, 'Holding "Governance" Accountable' (Sheila Kennedy3 May 2005) <https://sheilakennedy.net/2005/05/holding-governance-accountable/> accessed 3 July 2025.
[800] Cyber Promotions, Inc. v. American Online, Inc., 948 F. Supp. 436 (E.D. Pa. 1996).
[801] Kyle C Bailey, 'Regulating ISPs in the Age of Technology Exceptionalism | Texas Law Review' (Texas Law Review4 May 2020) <https://texaslawreview.org/regulating-isps-in-the-age-of-technology-exceptionalism/> accessed 3 July 2025.
[802] Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (papers.ssrn.com9 September 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3038939> accessed 4 August 2025.
[803] Jack M Balkin, 'Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society' (2004) 79 SSRN Electronic Journal.
[804] Klonick (n 466) 1672.
[805] Daphne Keller, 'Who Do You Sue?' (2019) <https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf> accessed 4 August 2025.
[806] Evelyn Douek, 'Governing Online Speech: From "Posts-As-Trumps" to Proportionality and

In the US, Section 230 of the Communications Decency Act plays a fundamental role in shaping the regulatory environment for online platforms. This provision exempts platforms from liability for third-party content by providing that providers or users of 'interactive computer services' shall not be deemed to be publishers or speakers of content provided by other users[807]. The legislative intent is to strike a balance between encouraging the development of internet services and promoting the voluntary moderation of harmful or offensive content[808]. This provision does not specifically list the types of content which platforms are exempt from liability, instead granting them broad protection through the term "any information". In judicial practice, disinformation is typically treated as third-party content, meaning that online platforms are generally immune from liability for hosting or distributing such content. While Section 230(c)(2) protects the right of online platforms to remove objectionable information "in good faith", it does not compel the platforms to modify and manage the posted content in this way, nor does it impose procedural fairness obligations on how content management is conducted.

This broad immunity shields online platforms from liability even if they algorithmically amplify the disinformation, including AI-generated deepfakes or misleading narratives, as long as they are not considered as content creators[809]. As a result, online platforms retain significant discretion to moderate or ignore harmful content, which often results in inconsistent or opaque enforcement practices. A growing number of scholars are questioning whether such broad immunity could inhibit proactive review and create a regulatory blind spot in the face of technologically advanced disinformation campaigns[810].

Probability' (2020) 121 SSRN Electronic Journal 759.

807 CDA § 230.

808 Zeran v America Online Inc, 129 F3d 327 (4th Cir 1997).

809 Danielle Citron and Benjamin Wittes, 'Fordham Law Review the Internet Will Not Break: Denying Bad Samaritans § 230 Immunity' (2017) <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5435&context=flr> accessed 4 August 2025.

810 Klonick (n 466) 1613.

In the US, although the First Amendment offers strong protection for free speech, certain categories of illegal content trigger affirmative obligations for online platforms, particularly under the frameworks addressing child sexual abuse material (CSAM) and copyright infringement. Under 18 U.S.C. § 2258A, electronic communication service providers and remote computing service providers are required to report any apparent violations involving CSAM to the National Center for Missing and Exploited Children (NCMEC). The provision explicitly states that platforms "must report… any facts or circumstances from which there is an apparent violation of section 2251… involving child pornography"[811]. Once such content is detected, the provider must file a report via the CyberTipline and preserve the content and the user's information for 90 days after submission, thereby facilitating investigation and prosecution. This imposes not merely a reactive duty but a legal obligation to monitor and report if the platform becomes aware of such material, even if not directly notified.

Also, the Digital Millennium Copyright Act (DMCA), codified at 17 U.S.C. § 512[812], provides platforms with a conditional "safe harbor" from liability for user-generated content Section 512 of the DMCA establishes a "notice-removal mechanism" that requires platforms to promptly remove or prohibit access to content when they learn of the existence of copyright infringement or content, and not to obtain direct economic benefits from the infringing material. This provision effectively provides internet service providers with a safe harbor from liability for copyright infringements by internet users, as well as helping copyright holders to quickly remove allegedly infringing material from the internet[813]. Unlike Section 230, which protects the platform's "inaction" on user content, this provision requires the platform to be exempted from infringement liability only through active action. If they fail to comply, they risk secondary liability. Although AI-generated disinformation does not necessarily involve copyright infringement, the original materials used in its generation

[811] 18 USC § 2258A (2023).
[812] 17 USC § 512 (1998).
[813] 17 USC § 512(d) (1998).

process may infringe the copyright of the original works, such as the unauthorized use of other people's images, videos, or text[814]. In these cases, the DMCA provides this remedy to prompt platforms to remove such AI-created content[815]. This illustrates a crucial point: while U.S. law generally promotes a hands-off approach to content moderation, it does require active intervention from platforms in specific legal contexts. The US's regulatory structure for platform content moderation is mainly based on this system: on the one hand, there are public law restrictions from the First Amendment, and on the other hand, Section 230 of CDA provides online platforms with immunity from liability. The First Amendment's protection of free speech enables American corporations to extend their rights and protection standards beyond their own territory when using online platforms abroad[816]. Section 230 creates a legal immunity shield for platforms by excluding them from legal responsibility for third-party content, leaving them free to decide whether or not to delete content without worrying about being held accountable. The purpose of such a regulatory framework is to provide information intermediaries such as Internet platforms with "legal exemptions" and "safe harbors" to encourage the free flow of information[817]. While this fosters a permissive regulatory environment that supports platform discretion, it imposes minimal enforceable procedural standards for content decisions. Platforms may adopt self-regulatory procedures[818], but users often lack statutory guarantees for fair treatment or appeal.

---

[814] Jennifer M Urban and Laura Quilter, 'Efficient Process or "Chilling Effects"? Takedown Notices under Section 512 of the Digital Millennium Copyright Act' (Ssrn.com23 May 2006) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2210935> accessed 4 August 2025.

[815] Jerome H Reichman, Graeme B Dinwoodie and Pamela Samuelson, 'A Reverse Notice and Takedown Regime to Enable Public Interest Uses of Technically Protected Copyrighted Works' (Ssrn.com2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1007817> accessed 7 August 2025.

[816] Giovanni De Gregorio and Roxana Radu, 'Digital Constitutionalism in the New Era of Internet Governance' (2022) 30 International Journal of Law and Information Technology.

[817] Anupam Chander, 'How Law Made Silicon Valley' (Ssrn.com15 August 2013) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2340197> accessed 7 July 2025.

[818] Nicolas Suzor and Rosalie Gillett, 'Self-Regulation and Discretion' [2022] Palgrave Global Media Policy and Business 259.

4.3.2 From Posts-as-Trumps to Proportionality: A Shift in Platform Governance Models

While the US's framework has historically given platforms wide freedom in content moderation, the rise of generative AI technologies is exposing institutional loopholes in this structure.

First, the lack of unified standards for content review leads to inconsistent implementation on different platforms. The online platform need to strike the balance between determining what content they carry and protecting user's rights to free speech[819]. The traditional "posts-as-trumps" model has been deeply influenced by the First Amendment, which places great emphasis on the protection of users' freedom of speech and therefore takes a cautious approach to user content to avoid excessive intervention[820]. In the early days of the Internet, online platforms generally adopted limited, clearly classified exceptions to manage content. For example, Facebook followed the principles of John Stuart Mill to establish its early community speech guidelines, believing that speech was only worth restricting when it could cause physical harm to others[821]. This model is increasingly unsuitable for managing large amounts of user-generated content, especially as generative AI becomes increasingly widespread on the internet, making the generation of disinformation more accessible. [822]. The public is increasingly aware that the surprising personalization capabilities of online platforms have increased the scope and corrosiveness of disinformation, leading to a deepening impact of information on public discourse and society[823]. In this context, the public's perception of the responsibilities that online platforms should bear has

---

[819] Douek (n 806) 764.
[820] Tarleton Gillespie, Custodians of the Internet (Yale University Press 2018) <https://yalebooks.yale.edu/book/9780300261431/custodians-of-the-internet/>.
[821] Alexis Papazoglou, 'What Would John Stuart Mill Do—to Fix Facebook?' (The New Republic28 January 2019) <https://newrepublic.com/article/152939/john-stuart-mill-doto-fix-facebook> accessed 7 July 202.
[822] Ava Kofman, Francis Tseng and Moira Weigel , 'The Hate Store: Amazon's Self-Publishing Arm Is a Haven for White Supremacists' ProPublica (16 April 2020) <https://www.propublica.org/article/the-hate-store-amazons-self-publishing-arm-is-a-haven-for-white-supremacists> accessed 7 July 202.
[823] John Bowers and Jonathan Zittrain, 'Answering Impossible Questions: Content Governance in an Age of Disinformation' (2020) 1 Harvard Kennedy School Misinformation Review.

changed, including whether platforms should be held liable for disinformation generated by AI and social harm caused by algorithmic manipulation[824]. In contrast, the proportionality principle framework, which can balance various conflicts of interest, has been widely adopted by various platforms.

Proportionality no longer focuses solely on the speech interests[825] of individual posts but also needs to consider other societal interests, such as public health, electoral integrity, or safety, rather than treating free speech as an absolute protection[826]. The existence of these interests can justify the proportionality of platform restrictions on content. An important example of the online platforms shifting their content policies from the "posts-as-trumps" model to proportionality is the emergence of conspiracy theorist Alex Jones and his Infowars website has prompted several popular platforms to change their speech policies[827]. The hate speech represented and disseminated by Alex Jones can and has resulted in revenge and violence against individuals or groups, and Facebook, Spotify, YouTube, and Apple, as private corporations using their content review rights to actively intervene with and filter such information[828], should not be subject to free speech protections[829].

The principle of proportionality recognizes that the platform has value judgment standards and can make such judgments clearly, rather than denying the existence of the platform's will and only dividing it through content classification, so that these judgment standards cannot be applied to more complex contexts. [830]. There is competition between different interests, and such competition requires platforms to

---

[824] Jonathan L Zittrain, 'Three Eras of Digital Governance' (papers.ssrn.com23 September 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458435> accessed 4 August 2025.
[825] Jamal Greene, 'Rights as Trumps?' (2018) 132 Harv. L. Rev. 28 <https://scholarship.law.columbia.edu/faculty_scholarship/2920/>.
[826] Douek (n 806) 759.
[827] Jane Coaston, 'Alex Jones Banned from YouTube, Facebook, and Apple, Explained' (Vox6 August 2018) <https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-youtube-conspiracy-theories> accessed 7 July 202.
[828] Meta Newsroom, 'Standing against Hate' (Meta Newsroom27 March 2019) <https://about.fb.com/news/2019/03/standing-against-hate/> accessed 8 July 2025.
[829] Marissa Lang, 'Blocked and Banned by Social Media: When Is It Censorship?' (SFChronicle.com31 August 2016) <https://www.sfchronicle.com/business/article/Blocked-and-banned-by-social-media-When-is-it-9193998.php> accessed 8 July 2025.
[830] Niels Petersen, Proportionality and Judicial Activism (Cambridge University Press 2017).

evaluate and weigh these conflicts in the specific contexts where disputes arise, to formulate their own content review rules[831]. The core advantage of this framework is that it does not rely on an abstract and fixed value hierarchy system, but allows for flexible adjustment of judgments in specific contexts, thereby achieving a dynamic balance between rights. For content review, if platforms adopt the principle of proportionality as the basis for their decision-making, they will be more inclined to make case-by-case judgments based on specific contexts, user impacts, and potential risks, rather than relying on a unified set of rules[832]. This approach helps to improve the rationality and legality of the platform's judgment on disinformation. However, due to different understandings and applications of the principle of proportionality, different platforms have formulated different review policies, further exacerbating the lack of unified content governance standards among platforms and the difficulty for users to predict whether the information they post will be deleted[833].

4.3.3 Practical Challenges in Content Moderation of Disinformation by Online Platforms

As the platform content governance model shifts to "proportional measurement", more platforms have begun to formulate differentiated content review policies based on their characteristics. Although this approach has improved the flexibility and relevance of content governance, it has led to governance fragmentation in implementation. Different standards across various platforms have created regulatory barriers, making the spread of false information more concealed and fluid, thereby weakening cross-platform governance of the regulators and causing users to lose trust in the online platforms due to inconsistent treatments.

---

831 Greene (n 825) 62.

832 Mike Ananny, 'Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance' (knightcolumbia.org2019) <https://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance> accessed 8 July 2025.

833 Grégoire CN Webber, 'Proportionality, Balancing, and the Cult of Constitutional Rights Scholarship' (2010) 23 Canadian Journal of Law & Jurisprudence 179.

First, for users, the lack of uniform standards for regulating disinformation does not provide equal protection for users. For example, during the COVID-19 pandemic, the dissemination of COVID-19 disinformation is often not confined to a single platform but flows and spreads across multiple social media and forums[834]. Malicious disseminators avoid censorship by constantly adjusting the languages and images they use, pushing it from marginal communities to mainstream platforms and amplifying the reach of disinformation[835]. As the influence of false information grows exponentially, disinformation presented on different platforms gradually shows consistency in discourse and focus on themes, making it more difficult to block its dissemination path[836]. The drawback of this lack of a unified content review policy is also reflected in the handling of disinformation generated by AI tools. Some platforms try to delete AI-generated content that is identified as false or misleading, while others choose to reduce its dissemination impact through labeling or downranking[837]. For example, Meta forces political advertisers to label when using AI or digital manipulation in ads on Facebook and Instagram[838], while TikTok has no requirements to label or remove disinformation. Besides, even after receiving complaints, the platform's ad review mechanism remains inconsistent[839]. These differences reflect that the management strategies adopted by multiple platforms in dealing with AI-generated disinformation are highly susceptible to profit pressures or public relations considerations, thus revealing obvious

---

834 N Velásquez and others, 'Online Hate Network Spreads Malicious COVID-19 Content Outside the Control of Individual Social Media Platforms' (2021) 11 Scientific Reports 11549 <https://www.nature.com/articles/s41598-021-89467-y>.

835 Oriol Artime and others, 'Effectiveness of Dismantling Strategies on Moderated vs. Unmoderated Online Social Platforms' (2020) 10 Scientific Reports.

836 Michael Röder, Andreas Both and Alexander Hinneburg, 'Exploring the Space of Topic Coherence Measures' [2015] Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15 399 <http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf>.

837 Trisha Meyer and Claire Pershan, 'Room for Improvement. Analysing Redress Policy on Facebook, Instagram, YouTube and Twitter - EU DisinfoLab' (EU DisinfoLab2025) <https://www.disinfo.eu/publications/room-for-improvement-analysing-redress-policy-on-facebook-instagram-youtube-and-twitter> accessed 9 July 2025.

838 BBC NEWS, 'Meta Requires Political Advertisers to Mark When Deepfakes Used' BBC News (9 November 2023) <https://www.bbc.co.uk/news/technology-67366311> accessed 18 August 2025.

839 ITV News, 'TikTok Failed to Detect Disinformation on Adverts to Voting ahead of Irish General Election' (ITV News29 November 2024) <https://www.itv.com/news/2024-11-29/tiktok-failed-to-detect-disinformation-relating-to-irish-general-election> accessed 8 July 2025.

uncertainty[840]. This situation may not only encourage the implicit bias against a certain position or economic interest group in the platform's internal logic but also make users unable to determine the reliability of the information, thereby weakening the public's trust in these online platforms[841].

Second, for regulators, inconsistent cross-platform content moderation policies and the cross-platform dissemination of harmful information are serious impediments to the enforcement of content regulation on a large scale. Governance consistency is challenged by the fact that different platforms have considerable discretion in setting and implementing their moderation standards.

On the one hand, disinformation often appears quickly on another platform after being deleted from one platform and is re-spread by taking advantage of the differences and loopholes in content review policies between platforms. Malicious disseminators circumvent the strict content management of a single platform, making it impossible for the single platform to contain the spread of information in the entire digital environment. For example, during the COVID-19 period, a large amount of disinformation migrated from strictly managed platforms to fringe platforms (such as Gab and Telegram), expanding the scope and life cycle of harmful information and circumventing strict review systems[842]. In addition, Mekacher and others tracked malicious users banned by Twitter and found that these users quickly migrated to Gettr after being removed and continued to post similar disinformation, revealing that if regulatory measures are only concentrated on a single platform, they will not be able to effectively cut off the dissemination chain of disinformation[843]. Cinelli and others'

---

[840] Otávio Vinhas and Marco Bastos, 'The WEIRD Governance of Fact-Checking and the Politics of Content Moderation' (2023) 27 New Media & Society.

[841] Wilberforce Murikah, Jeff Kimanga Nthenge and Faith Mueni Musyoka, 'Bias and Ethics of AI Systems Applied in Auditing - a Systematic Review' (2024) 25 Scientific African <https://www.sciencedirect.com/science/article/pii/S2468227624002266>.

[842] Heidi Schulze and others, 'Far-Right Conspiracy Groups on Fringe Platforms: A Longitudinal Analysis of Radicalization Dynamics on Telegram' (2022) 28 Convergence: the International Journal of Research into New Media Technologies 1103, 1106.

[843] Amin Mekacher, Max Falkenberg and Andrea Baronchelli, 'The Systemic Impact of Deplatforming on Social Media' (2023) 2 PNAS Nexus.

research also pointed out that although the design logic of the platforms (such as push mechanisms and user structures) is different, the same information presents similar propagation curves on different platforms, such as propagation speed, coverage, and user response methods[844]. This means that once information appears, it often arouses similar reactions on multiple platforms simultaneously[845], rather than being effectively suppressed by the governance measures of a certain platform, indicating that the role of governance boundaries between platforms is extremely limited.

On the other hand, the platforms' regulatory policies against disinformation are not comprehensive enough, lacking clearly stated terms and well-established remedies for users. As Schaffner and others show in their comparative study of 43 user-generated content (UGC) moderation policies, each platform takes a very different approach to regulating copyright infringement, hate or harmful speech, and disinformation[846]. Most platforms illustrate possible violations with examples rather than clearly defining review standards in their terms, and their actual review practices may differ from stated policies[847]. Besides, in contrast to copyright infringement, which has a well-developed system of remedies, users lack a clear recourse for disinformation that has been removed because it is false, such as a clear legal basis or a dedicated policy page[848]. It was found that although most mainstream platforms (such as YouTube and Facebook) generally claim to have a complaint mechanism, there are huge differences in response speed, review transparency, and interpretation of review logic[849]. Some platforms do not even provide clear complaint paths or remedies. DisinfoLab has also criticized that

---

[844] Matteo Cinelli and others, 'The COVID-19 Social Media Infodemic' (2020) 10 Scientific Reports <https://www.nature.com/articles/s41598-020-73510-5>.

[845] Daniel M Romero, Brendan Meeder and Jon Kleinberg, 'Differences in the Mechanics of Information Diffusion across Topics' [2011] Proceedings of the 20th International Conference on World Wide Web - WWW '11.

[846] Brennan Schaffner and others, '"Community Guidelines Make This the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies' [2024] ArXiv (Cornell University).

[847] J Nathan Matias, Austin Hounsel and Nick Feamster, 'Software-Supported Audits of Decision-Making Systems: Testing Google and Facebook's Political Advertising Policies' (2022) 6 Proceedings of the ACM on Human-Computer Interaction 1.

[848] Schaffner and others (n 846) 2.

[849] Meyer and Pershan (n 837).

there is currently no uniform and effective mechanism to ensure that users have access to internal redress for wrongful take-downs or content blocking, and that users can only resort to public pressure or legal action, this 'systemic failure' that further exacerbates procedural unfairness [850]. Besides, the fragmentation of moderation standards exacerbates cross-border enforcement challenges [851], as platforms based in other countries or subject to different national laws can create regulatory disputes, resulting in regulations enforceable in one jurisdiction being unenforceable in another[852]. This regulatory inconsistency allows content to exploit jurisdictional loopholes and migrate to platforms with more lax standards, undermining effective enforcement.

Another major challenge for online platforms in practical implementation is that, due to their multiple roles in the generative AI ecosystem with concentrated power, their legal responsibilities are difficult to track, resulting in a systemic imbalance in the governance structure. The platforms may be both the provider and developer of AI tools and the distributor and regulator of information, but the boundaries of their responsibilities in the entire process are not clearly defined[853]. The platform's multiple identities not only blur the attribution of its accountabilities but also trigger widespread controversy over whether the platform bears legal responsibility for disinformation generated by generative AI[854].

The functional positioning of the online platform in AI-generated information is highly ambiguous, which has caused disputes over the attribution of responsibility. On the technical level, the platform may only serve as a custodian of the model and assume a neutral role[855]; but in actual operation, the platform often has decisive control over the

---

[850] EU DisinfoLab, 'When Platforms Make Mistakes, Users Need Redress' (EU DisinfoLab2022) <https://www.disinfo.eu/publications/when-platforms-make-mistakes-users-need-redress/> accessed 10 July 2025.

[851] Minako Morita-Jaeger and others, 'Interoperability of Data Governance Regimes: Challenges for Digital Trade Policy | CITP' (Citp.ac.uk2024) <https://citp.ac.uk/publications/interoperability-of-data-governance-regimes-challenges-for-digital-trade-policy> accessed 10 July 2025.

[852] Giovanni De Gregorio and Simona Demková, 'The Enforcement Dilemmas in Europe's Digital Rulebook' (Tech Policy Press19 May 2025) <https://www.techpolicy.press/the-enforcement-dilemmas-in-europes-digital-rulebook/> accessed 10 July 2025.

[853] Gregorio and Radu (n 816).

[854] Balkin (n 792)1174.

[855] Christoph B Graber, 'Bottom-up Constitutionalism: The Case of Net Neutrality' (2017) 7 SSRN

generation logic, algorithm configuration, and content recommendation path[856]. In the applications of generative AI, platforms are not only intermediaries of content but may also be "co-generators" of content. For example, Meta' Emu is a generative model capable of text-to-image and image-to-image, designed to help develop high-quality and controllable image generation tools, but this AI model may contain disinformation or misleading content, causing it to spread on the platform[857]. According to research by CDT (Center for Democracy & Technology), the public's distrust of disinformation generated by the platform is generally over 50% across different languages[858], but platforms usually refuse to take responsibility and continue to claim that they are "neutral channels."

The internal decision-making mechanism for content review is opaque, concealing the platform's preferences and making it impossible for the public to understand and monitor the platform's decision-making logic effectively[859]. The 2023 report from NYU's Stern Center illustrated that most of the major technology platforms for generative AI do not provide adequate information about how they use AI in their recommendation, review, and ranking systems[860]. It highlights a lack of transparency in the disclosure of training data, model updates, and evaluation processes, making it difficult for users and regulators to assess potential risks or determine whether to hold companies accountable. As platforms take initiative in designing content management policies and setting up algorithms, these engagements can largely influence or even

---

Electronic Journal.

[856] Kate Crawford and Trevor Paglen, 'Excavating AI: The Politics of Images in Machine Learning Training Sets' (2021) 36 AI & SOCIETY 1105.

[857] Rohit Girdhar and others, 'Emu Video: Factorizing Text-To-Video Generation by Explicit Image Conditioning' [2023] ArXiv (Cornell University).

[858] Mona Elswah, Aliya Bhatia and Dhanaraj Thakur, 'Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages' (Center for Democracy and Technology28 June 2025) <https://cdt.org/insights/content-moderation-in-the-global-south-a-comparative-study-of-four-low-resource-languages/> accessed 11 July 2025.

[859] Ben Bradford and others, 'Report of the Facebook Data Transparency Advisory Group' (2019) <https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf> accessed 10 July 2025.

[860] Paul M Barrett, 'NYU Stern Center for Business & Human Rights Safeguarding AI: Addressing the Risks of Generative Artificial Intelligence' (Nyu.edu25 June 2025) <https://bhr.stern.nyu.edu/publication/safeguarding-ai-addressing-the-risks-of-generative-artificial-intelligence/> accessed 10 July 2025.

determine the content of information that users can access and share[861]. The content review policies designed by platforms cannot be reduced to a case-by-case adjudication of personal expression, but rather have a systemic impact on the discursive ecology of the entire platform[862]. Besides, while giving platforms the flexibility to set review standards, proportionality also brings accountability challenges because it allows platforms to decide the benchmark for removal at their own discretion, and this discretion is not always transparent. This shows that in the field of AI-generated content, the platform has too much discretion and lacks clear boundaries, further exacerbating the fragmentation of legal governance[863]. Especially when users face drastically different governance policies when uploading the same expressions on multiple platforms, such differences may pose substantial challenges to the protection of user rights. Because US laws (such as Section 230) do not force platforms to disclose how they regulate their domain, platforms are not required to publish their disinformation identification standards, third-party fact-checking cooperation mechanisms, or their content review algorithms[864]. This makes it difficult for external supervision to implement and assess whether the platform has truly fulfilled its risk prevention responsibilities.

4.4 China: Fluctuations Caused by Special Actions Affect Enforcement

4.4.1 Institutional Framework for China's Online Disinformation Governance

China's Internet disinformation governance system is characterized by the completeness and strict implementations, forming a legal framework based on the

---

[861] Vicki C Jackson, 'Constitutional Law in an Age of Proportionality' (2015) 124 Yale Law Journal <https://www.yalelawjournal.org/article/constitutional-law-in-an-age-of-proportionality> accessed 2 December 2024.

[862] JAMEEL JAFFER, 'Facebook and Free Speech Are Different Things' (knightcolumbia.org2019) <https://knightcolumbia.org/content/facebook-and-free-speech-are-different-things> accessed 10 July 2025.

[863] Matthias C Kettemann and Wolfgang Schulz, 'Setting Rules for 2.7 Billion: A (First) Look into Facebook's Norm-Making System; Results of a Pilot Study' (2020) 1 Ssoar.info 34 <https://www.ssoar.info/ssoar/handle/document/71724>.

[864] Carly Miller, 'Can Congress Mandate Meaningful Transparency for Tech Platforms' (Stanford.edu2 June 2021) <https://fsi.stanford.edu/news/meaningful-transparency-0> accessed 6 August 2025.

Cybersecurity Law, the Data Security Law, and the Personal Information Protection Law, and supported by the Provisions on the “Ecological Governance of Network Information Contents”, the “Interim Measures for the Administration of Generative Artificial Intelligence Services”, and other special regulations. This system explicitly requires online platforms to fulfill their management responsibility of "review before release" for user-generated content and to conduct comprehensive reviews by equipping review teams that are commensurate with their business scale. According to the 2023 development report of the Cyberspace Administration of China, relevant website platforms have closed 127,878 illegal and irregular accounts under the law and contract, showing the strong action of the governance system[865].

In terms of governance measures for disinformation, China adopts a governance framework that combines regular legal supervision with platform cooperation in special operations[866]. On the one hand, a long-term and effective supervision mechanism is established through basic laws such as the Cybersecurity Law, DSL, and PIPL; on the other hand, through the "Qinglang" special operation, online platforms are required to cooperate and carry out centralized rectification of prominent problems, such as banning illegal accounts and deleting disinformation.

4.4.2 Allocation of Liabilities in China’s Disinformation Governance Framework

In terms of governance structure, China has established a collaborative framework in which the government plays a leading role, and online platforms bear the main responsibility.

In the division of responsibilities among government regulatory agencies, the Cyberspace Administration of China (CAC), as the core regulatory body, mainly bears the responsibility of overall coordination, while various departments participate in collaboration based on their division of responsibilities. The CAC is responsible for

---

865 State Internet Information Office, ‘National Informatization Development Report(2023)’ (2023) <https://www.gov.cn/lianbo/bumen/202409/content_6973030.htm> accessed 10 July 2025.

866 Juan Zhang, ‘A Comparative Study of Fake News Governance Measures in China and Abroad’ (Fx361.com2020) <https://m.fx361.com/news/2020/0718/6879933.html> accessed 12 July 2025.

formulating work norms for online information governance, guiding and urging internet service providers to formulate and improve content moderation rules, and conducting daily supervision and inspection[867]. Different administrative departments, according to their respective job functions, carry out collaborative governance in several ways. For example, the State Administration for Market Regulation is responsible for the governance of false propaganda in the field of Internet advertising.

Online platforms, as direct channels for the dissemination of disinformation, bear the primary responsibility for content review[868]. China's information governance system has established countermeasures to quickly address AI-generated disinformation. The "Interim Measures for the Administration of Generative Artificial Intelligence Services" [869] implemented in August 2023, is one of the earliest departmental regulations in the world that specifically regulates generative AI. Article 4 clearly stipulates that the provision and use of generative AI services shall not generate false and harmful information, and Article 17 requires that AI-generated content be marked prominently[870]. The "Provisions on Information Governance on Cyber Violence," promulgated in 2024, stipulates that AI-generated disinformation is one of the manifestations of cyber violence in the form of rumors and slander, and is one of the important governance objects of this departmental paper. Through Articles 2, 5, and 10, disinformation is included in the cyber violence governance framework, and through Articles 11 and 12, platforms are required to use AI tools in combination with manual review to strengthen the identification and monitoring of cyber violence information. These specialized administrative regulations and departmental rules provide an institutional basis for platforms to govern AI-generated disinformation and reflect

---

[867] Cyberspace Administration of China, Provisions on Information Governance on Cyber Violence (promulgated 1 June 2024, effective August 2024) https://www.gov.cn/gongbao/2024/issue_11526/202408/content_6969181.html accessed 13 July.

[868] Peng Lu, Lvjun Zhou and Xiaoguang Fan, 'Platform Governance and Sociological Participation' (2023) 10 The Journal of Chinese Sociology.

[869] Interim Measures for the Management of Generative Artificial Intelligence Services (生成式人工智能服务管理暂行办法) (promulgated 10 July 2023, effective 15 August 2023).

[870] Xiaodong Guo, 'Risks of Generative Artificial Intelligence and Its Inclusive Legal Governance' (2023) 25 Journal of Beijing Institute of Technology(Socoal Sciences Edition) 93.

China's rapid response capability in Internet governance.

Furthermore, the simultaneous development of technological detection tools has supported online platforms in identifying and detecting disinformation[871]. Significant advances in generative AI technology have significantly reduced the cost for malicious users to exploit this technology to generate disinformation. In this context, platforms' content review systems are facing unprecedented pressure.

China encourages the adoption of a technical approach that combines AI recognition with manual review to promote the development and application of deep synthesis detection technology[872]. According to data from Zellers et al., current detectors have an accuracy rate of approximately 73% in identifying AI-generated fake news with moderate training data, and this accuracy can be increased to 92% when using the generator model itself as a detector[873]. However, these tools still struggle to distinguish between human-written and AI-generated content[874], and their detection accuracy would drop by 20% to 50% when content is translated or manually paraphrased. These data demonstrate that current detection tools still need to improve their performance in detecting text or content that has been obfuscated to varying degrees. In content moderation practices, platforms can combine deleted disinformation with user reports, conduct detailed categorization based on factors such as domain and region, and establish their own disinformation feature databases to improve their disinformation identification and handling mechanisms[875]. This technical governance capability effectively enhances the system's adaptability to new types of disinformation.

---

[871] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society <https://journals.sagepub.com/doi/10.1177/2053951719897945>.

[872] Li Wang, 'Development and Prospect of False Information Detection on Social Media' (2022) 53 Journal of Taiyuan University of Technology.

[873] Rowan Zellers and others, 'Defending against Neural Fake News' (arXiv.org29 May 2019) <https://arxiv.org/abs/1905.12616> accessed 14 April 2025.

[874] Vivian van Oijen, 'AI-Generated Text Detectors: Do They Work? | SURF Communities' (communities. surf.nl31 March 2023) <https://communities.surf.nl/en/ai-in-education/article/ai-generated-text-detectors-do-they-work> accessed 10 July 2025.

[875] Cyberspace Administration of China, Provisions on the Administration of Deep Synthesis of Internet-based Information Services (promulgated 25 November 2022, effective 10 January 2023) https://www.cac.gov.cn/2022-12/11/c_1672254275974116.htm accessed 11 July 2025.

4.4.3 The Impact of Policy Implementation Fluctuations on Platform Content Moderation: Focus on Special Actions

The mode of combining normalized supervision with special actions adopted by China's Internet governance has had a significant periodic impact on the platform content review system. While this governance model ensures policy enforcement, it also causes continuous fluctuations in platform moderation standards, technical systems, and users' reactions.

First, the volatility of review standards and regulatory focus has led to deviations in the actual implementation of online platforms, such as the accidental deletion of users' content. The most direct impact of special actions is the high-frequency adjustment of platform content review standards. During the normalized supervision period, platforms mainly formulate relatively stable review rules based on basic laws and regulations such as the "Regulations on the Ecological Governance of Network Information Content"; while special actions often put forward new regulatory requirements for specific types of content (such as protection of minors, algorithm governance, or cyber violence), forcing platforms to update their review standards frequently.

To actively respond to these special actions, different platforms or their supervisory agencies may expand the definition of "disinformation " and thus present more stringent review standards. For example, in the 2023 "Qinglang" special action, some social media platforms included "comments on hot social issues that may cause negative emotions among netizens" in the scope of moderation [876], far exceeding the requirements of the Cybersecurity Law that disinformation should be deleted only if it disrupts social and economic order or infringes others' rights. This expansion of interpretation forces online platforms to adopt a predictive censorship strategy and pre-filter a large amount of content in the gray area.

---

[876] Jialai Fan, 'Douyin: Information on Events That May Trigger "Cyber Violence" and "Opening of Boxes" Will Be Included in the "Controversial" Hot Spots for Analysis and Treatment' (Thepaper.cnMay 2025) <https://m.thepaper.cn/detail/30886170> accessed 14 July 2025.

In addition, inconsistent standards across regions further exacerbate the difficulty of horizontal management of disinformation. China exercises territorial jurisdiction over online information moderation, and different regional cybersecurity and informatization departments have different interpretations of policies, leading to horizontal entities with law enforcement functions having their own understanding of the policy[877]. For example, in response to uncertain information about covid-19, Province A may require online platforms to restrict the publishments of all unverified information, while Province B only requires platforms to label uncertain content as "doubtful"[878]. Different treatments for the same information have led to the emergence of regionalized review, which requires dedicated personnel to coordinate local standards, which also increases administrative costs.

The time lag in the transmission of vertical policies reduces content moderation efficiency, causing disinformation cannot be dealt with in a timely manner. Usually, the notice and specific requirements of the special action are first issued to the platform headquarters and then transmitted to the specific review team. In this process, the auditors may not fully understand the specific implementation standards after the special action is launched, resulting in many illegal contents not being handled in time, thus causing information distortion[879].

Secondly, the technical system of online platforms faces unstable cyclical changes, showing incompatibility of technical adaptation.The technical system used by online platforms for content moderation shows significant path dependence characteristics when facing special operational requirements. AI models used for disinformation detection require stable training data and testing cycles. China's special actions usually

---

[877] Nianping Chen, 'Solving the New Contradiction of "Threading a Thousand Threads with One Needle"——Review and Prospect of Research on the Reform of Grass-Roots Comprehensive Administrative Law Enforcement' (2024) 6 Study on Party and Government.

[878] Yi Zhang and Chen Chen, 'Which Province or City Responded More Decisively and Effectively to the Epidemic? We Did Some Data Analysis' (Baidu.com2020) <https://baijiahao.baidu.com/s?id=1662292560776616747&wfr=spider&for=pc> accessed 14 July 2025.

[879] Guoming Yu and others, 'The Manifestation, Governance and Effect Evaluation of False Information on the Internet' (2024) 02 Youth Journalist.

target harmful information in new presentation forms, such as disinformation generated by AI tools. The applicability of large language models (LLMs) in content review is highly dependent on the annotation consistency, context coverage, and scale of training data[880]. Unlike the previous practice of converting policies into guidelines and then having them reviewed by humans or trained by models, the platform can now directly input policies through prompts, allowing review behaviors to be flexibly adjusted and quickly adapted[881]. However, Sudden concept drift in the data (i.e., sudden changes in the data) can cause the performance of machine learning applications to drop rapidly[882]. Besides, this model can currently only withstand small adjustments, such as changes to limited words, and larger adjustments will still lead to fluctuations and misrepresentations in the output results[883]. Therefore, when the platform needs to adapt to new detection targets and the training data is not yet mature, the accuracy of AI tools in reviewing disinformation may drop sharply.

Finally, the frequent changes in the platform's content review policies caused by the special actions will make users (especially content creators) more cautious when creating and uploading content. Ultimately lead to a decrease in the number of attractive and high-quality content on the platform and reduce the platform's competitiveness in the commercial market.

The cyclical pressure of the special action forces platforms to adjust their content review standards according to different requirements, resulting in the review and recommendation mechanism not being fully understood by users, which in turn forces creators to constantly test and adjust the content they publish to avoid being marked as

---

[880] Huan Ma and others, 'Adapting Large Language Models for Content Moderation: Pitfalls in Data Engineering and Supervised Fine-Tuning' (arXiv.org2023) <https://arxiv.org/abs/2310.03400> accessed 14 April 2025.

[881] Todor Markov and others, 'A Holistic Approach to Undesired Content Detection in the Real World' (2023) 37 Proceedings of the ... AAAI Conference on Artificial Intelligence 15009.

[882] Lennart Justen and others, 'No Time like the Present: Effects of Language Change on Automated Comment Moderation' (2022) 01 2022 IEEE 24th Conference on Business Informatics (CBI) 40 <https://ieeexplore.ieee.org/document/9944746> accessed 14 July 2025.

[883] Konstantina Palla and others, 'Policy-As-Prompt: Rethinking Content Moderation in the Age of Large Language Models' (arXiv.org2025) <https://arxiv.org/abs/2502.18695> accessed 14 April 2025.

false information and deleted[884]. For example, content producers in short video and live e-commerce regard algorithms as "weather vanes" and respond to the instability of platform traffic mechanisms by testing titles, times, and topics in real time[885]. Another study pointed out that anchors feel anxious about fluctuations in algorithms or review rules, and actively avoid sensitive topics and turn to low-risk content[886]. Although these adjustments meet regulatory requirements in the short term, they may distort market mechanisms and innovation incentives in the long run.

Content creators active on online platforms are also considered creative labor[887], and their content output is unstable due to changes in audience preferences[888], frequently changing rules after the digitalization of the cultural industry, and changes in platform algorithms. The platform adjusts the settings of its recommendation system due to changes in the focus of content review, which exacerbates the volatility of creators' views[889]. In this case, content creators may choose to upload the same or similar content to multiple platforms. Even if a platform chooses to delete or not allow it to be published after content review, the creators' publication on other platforms will not be interfered with. Such a multi-platform publishing strategy will lead to an increase in the homogeneity of different platforms and a significant reduction in differentiation, thereby reducing their respective unique competitive advantages. Moreover, such fluctuations may cause small and medium-sized creators to reduce the number of works or even exit the platform, resulting in less diversity in platform content and a decline in

---

884 Jenna Burrell, 'How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms' (2015) 3 SSRN Electronic Journal.

885 Luzhou Li and Kui Zhou, 'When Content Moderation Is Not about Content: How Chinese Social Media Platforms Moderate Content and Why It Matters' (2024) 27 New Media & Society 6150.

886 Caoyang Shen and Oliver L. Haimson, 'The Virtual Jail: Content Moderation Challenges Faced by Chinese Queer Content Creators on Douyin' (2025) CHI '25, Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems 177:1, 177:2. <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/196552/chi25-924.pdf>.

887 Brooke Erin Duffy and others, 'The Nested Precarities of Creative Labor on Social Media' (2021) 7 Social Media + Society 1, 2.

888 Xinlu Wang and Shule Cao, 'Harnessing the Stream: Algorithmic Imaginary and Coping Strategies for Live-Streaming E-Commerce Entrepreneurs on Douyin' (2024) 11 The Journal of Chinese Sociology 10.

889 Meng Liang, 'The End of Social Media? How Data Attraction Model in the Algorithmic Media Reshapes the Attention Economy' (2022) 44 Media, Culture & Society.

user satisfaction.

4.5 Conclusion

This chapter focuses on the difficulties and enforcement challenges faced by online platforms in regulating AI-generated disinformation under the laws and regulations of three jurisdictions. These challenges share some commonalities: a lack of clarity and specificity in legal provisions, which leaves online platforms with significant discretion in fulfilling their obligations, leading to inefficient enforcement of legal obligations. Also, enforcement challenges in the three jurisdictions also present distinct challenges due to differing legal provisions and enforcement priorities. While the EU's DSA allows platforms to agree on content moderation rights through their Terms of Services with users, in practice, different online platforms have varying requirements for content moderation due to their business priorities and platform functionality. Consequently, the same information posted on different platforms may receive different review results, significantly impact the interests of information publishers and diminish users' willingness to share information. Furthermore, the DSA's ambiguity in its definition of systemic risk may lead to uncertain in platforms' identification of disinformation that pose systemic risk, preventing them from effectively mitigating risk. While Section 230 of the CDA in the US grants platforms broad immunity for user-generated content, it does not restrict platform's authority to agree with users on content review. However, in practice, content review rules vary between platforms, and malicious users exploit these differences to evade regulation by spreading content across platforms. China has comprehensive laws and administrative regulations clarifying platform liabilities, but their actual implementation is significantly influenced by national policies. This thesis, citing special action as an example, illustrates the fluctuations in platform review priorities caused by this special action. These fluctuations not only make it technically difficult for platforms to adapt to these changes, but also make creators more cautious, thereby reducing the frequency with which they publish their works.

5. Cross-Jurisdictional Approaches to Regulating AI-Generated Disinformation

Overview

In this chapter, I will put forward two practical suggestions for promoting cross-border collaborative governance of AI-generated disinformation based on promoting procedural fairness, given the widespread cross-border spread of disinformation.

5.1 Cross-Border Cooperation in the Governance of Disinformation: Challenges and Recommendations

The growing prevalence of cross-border disinformation is partly due to the legal and regulatory frameworks of selected jurisdictions. While these frameworks pursue legitimate domestic objectives, such as safeguarding free speech in the US, ensuring data protection in the EU[890], or maintaining information security in China, they can also indirectly facilitate or even accelerate the cross-border spread of disinformation or misleading content. Against this backdrop, given the diverse political, legal, and cultural contexts, it is unrealistic to expect convergence on substantive regulatory standards across jurisdictions. While the previous chapters highlighted the differences in the substantive standards for online platform regulation in different jurisdictions, this chapter focuses on the procedural safeguards that could be strengthened to mitigate the harm of AI-generated disinformation. A more feasible approach is to focus on strengthening procedural safeguards within existing systems, such as by establishing and improving effective reporting and appeal mechanisms and regularly publishing transparency reports on disinformation moderation[891]. By empowering users to

[890] Giovanni De Gregorio, 'Digital Constitutionalism and Freedom of Expression', Digital Constitutionalism in Europe Reframing Rights and Powers in the Algorithmic Society (Cambridge University Press 2022) <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/72ACEF48324D180E95BBD456E52E9C96/9781316512777c5_157-215.pdf/digital_constitutionalism_and_freedom_of_expression.pdf> accessed 10 July 2025.

[891] Mark MacCarthy, 'Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry' (papers.ssrn.com12 February 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3615726> accessed 4 August 2025.

challenge moderation decisions and ensuring accountability for content governance, such mechanisms not only enhance fairness within domestic frameworks but also help foster trust and cooperation in cross-border responses to disinformation.

5.1.1 The Current State of Cross-Border Dissemination of Disinformation and the Legal Basis for Collaborative Regulation

One of the most significant challenges in regulating AI-generated disinformation lies in its ability to spread across national borders. While traditional forms of information dissemination are often limited by physical or jurisdictional constraints, digital disinformation, once generated in a particular jurisdiction, is immediately disseminated globally through online platforms. For example, the existence of the Austrian Data Protection Act[892] demonstrates that by the late 20th century, Austrian service providers were importing data from foreign clients and, accordingly, exporting data back to clients, enabling direct access to foreign databases from Austria[893].

This creates a significant gap between the global nature of disinformation and the territorial nature of national legal systems. Scholars have highlighted how the internet continually challenges territorial sovereignty, as states legislate based on their territorial jurisdictions, while digital communications effortlessly transcend borders[894]. The rapid development of generative AI technology has exacerbated this gap, exploiting differences in national legal frameworks and enforcement capabilities to enable the rapid spread of large amounts of disinformation beyond the country of origin[895].

This mismatch has led to jurisdictional gaps. While the EU has attempted to extend the extraterritorial reach of its laws and regulations, such as the GDPR and the DSA, this coverage is carefully structured around the EU's internal market. Article 3 of the GDPR explicitly states that its scope extends beyond data controllers and processors located

---

[892] Austrian Data Protection Act (Datenschutzgesetz, DSG) 1978, § 33.
[893] Austrian Data Protection Act 1978 (Österreichisches Datenschutzgesetz, DSG), § 34(2).
[894] Jack Goldsmith, 'Who Controls the Internet? Illusions of a Borderless World' (2007) 23 Strategic Direction.
[895] Leo SF Lin, 'Organizational Challenges in US Law Enforcement's Response to AI-Driven Cybercrime and Deepfake Fraud' (2025) 14 Laws 46 <https://www.mdpi.com/2075-471X/14/4/46>.

in the EU to include the offering of goods or services to data subjects located in the EU and the monitoring of data subjects' behavior within the EU[896]. This provision enables the EU to regulate foreign companies whose services target EU residents. For example, in the case of Google Spain v AEPD and Mario Costeja González[897], Google Inc. and Google Spain were deemed to be a single economic unit and, therefore, data controllers within the GDPR. Similarly, the DSA extends its scope beyond the EU's borders, requiring intermediary service providers that are classified as VLOPs and VLOSEs to comply with obligations related to systemic risk assessment, content moderation, and independent audit, regardless of where these intermediary service providers are established, as long as the recipient of the service is established or physically located in the EU[898]. While the DSA does not utilize the same explicit extraterritoriality of rules[899] as the GDPR, its obligations effectively apply to foreign providers of services in the EU, a mechanism that reflects the so-called "Brussels effect"[900]. The EU relies on the size and attractiveness of its internal market to push its regulatory standards beyond its borders, but the actual enforceability of these standards depends on specific jurisdictional mechanisms, such as the threat of significant fines, and the commercial incentives for foreign companies to comply to maintain access for EU consumers[901].

China has developed a model of strict, state-led regulation[902] of disinformation, characterized by universal platform obligations and direct state oversight. Through the Cybersecurity Law, the Data Security Law, the Personal Information Protection Law, and a series of regulations issued by the Cyberspace Administration of China (CAC)[903],

---

[896] Regulation (EU) 2016/679 (General Data Protection Regulation) [2016] OJ L119/1, art 3.
[897] Case C-131/12 Google Spain SL, Google Inc v Agencia Española de Protección de Datos (AEPD), Mario Costeja González EU:C:2014:317.
[898] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 2(1).
[899] Lena Hornkohl, 'The Extraterritorial Application of Statutes and Regulations in EU Law' (2022) 1 SSRN Electronic Journal.
[900] Anu Bradford, The Brussels Effect: How the European Union Rules the World (Oxford University Press 2020) <https://scholarship.law.columbia.edu/books/232/>.
[901] Christopher Kuner, 'Extraterritoriality and Regulation of International Data Transfers in EU Data Protection Law' (Social Science Research Network30 August 2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2644237> accessed 4 August 2025.
[902] Nikolay Bozhkov, 'China's Cyber Diplomacy: A Primer ' (Horizon2020) <https://eucyberdirect.eu/research/chinas-cyber-diplomacy-a-primer> accessed 10 July 2025.
[903] Rogier Creemers and Nicholas D Wright, 'The International and Foreign Policy Impact of China?S

the state requires platforms to undertake broad content moderation and disinformation control obligations, ensuring effective enforcement through administrative penalties and platform accountability[904]. However, due to the structural separation of the Chinese Internet from the global online sphere[905], the effectiveness of this approach is largely limited to the digital ecosystem within China[906]. However, unlike laws like the EU's GDPR or DSA, which have explicit extraterritorial application provisions, China's digital regulatory framework, due to potential trade barriers and practical enforcement difficulties[907], primarily applies domestically, lacking direct control over transnational platforms or overseas information flows[908]. China has constructed the Great Firewall of China[909] through technology and law to isolate the domestic Internet ecosystem from the global network by filtering external information and realizing strict control over content input. However, reports in commercialized media, retweets by users of Chinese platforms, and the direct registration of accounts by foreign entities for posting have impacted the strict regulation of content input. Lu and others[910] found that for Twitter and Weibo, the two major platforms with close to the same number of daily active users[911], only 20% of Twitter users[912] are located in the US, where the company

---

Artificial Intelligence and BigData Strategies' (Air University Press 2019) <https://www.jstor.org/stable/resrep19585.23> accessed 8 August 2025.

[904] Rogier Creemers, 'China's Conception of Cyber Sovereignty: Rhetoric and Realization' [2020] SSRN Electronic Journal.

[905] Samm Sacks, 'Beijing Wants to Rewrite the Rules of the Internet' (The Atlantic18 June 2018) <https://www.theatlantic.com/international/archive/2018/06/zte-huawei-china-trump-trade-cyber/563033/> accessed 10 July 2025.

[906] Justin Sherman, 'China's War for Control of Global Internet Governance' [2022] SSRN Electronic Journal.

[907] Jinhe Liu and Baoguo Cui, 'Data Localization and the Legitimacy and Trend of Data Defensivism' (2020) 12 Global Review89.

[908] Wanshu Cong, 'The Spatial Expansion of China's Digital Sovereignty', Digital Sovereignty in the BRICS Countries (Cambridge University Press 2024) <https://www.cambridge.org/core/books/digital-sovereignty-in-the-brics-countries/spatial-expansion-of-chinas-digital-sovereignty/D18D59288E33063AF1EBAF416FACD6A2> accessed 10 July 2025.

[909] Nicholas Gisonna, 'Great Firewall | History, China, Hong Kong, & Facts | Britannica' (www.britannica.com27 July 2023) <https://www.britannica.com/topic/Great-Firewall>.

[910] Yingdan Lu and others, 'How Information Flows from the World to China' (2022) 29 The International Journal of Press/Politics 305, 308.

[911] Weibo Corporation , 'Weibo Reports Second Quarter 2021 Unaudited Financial Results | Weibo Corporation' (*Weibo Corporation*2021) <http://ir.weibo.com/news-releases/news-release-details/weibo-reports-second-quarter-2021-unaudited-financial-results> accessed 3 September 2025.

[912] Lu and others (n 910) 309.

is headquartered, but almost all Weibo users have an IP address in the company's mainland China location. Besides, Chinese users must use virtual private networks (VPNs) or other censorship-avoiding technologies to access social media outlets like Twitter[913]. But even under these circumstances, content that is hotly debated on Twitter, especially information that matches the concerns of users in their home countries, is able to flow into Chinese internet platforms such as Weibo[914], regardless of whether its authenticity is certified by official media. In conclusion, given the relative isolation of China's Internet ecosystem from the global network, it is clear that the regulatory model is structurally deficient in its ability to control cross-border dissemination of disinformation, even though the model is highly effective in its own country.

Unlike the EU and China, which structure platform regulation primarily around statutory obligations, the US combines two complementary forms of protection, collectively fostering a permissionless environment for the spread of disinformation. On the one hand, the First Amendment provides a strong safeguard for free speech, shielding most speakers from government regulation, even when disseminating disinformation or misleading content, as long as it does not constitute unprotected speech such as incitement to incitement to imminent lawless action[915], true threat[916], or defamation motivated by actual malice[917]. On the other hand, Section 230 grants online platforms immunity from liability for disinformation by not treating them as publishers of user-generated information. These two protections operate at different levels: the First Amendment protects individuals who generate disinformation, while Section 230 protects platforms that disseminate it. Therefore, the interaction between the two in practice means that in the US, the creation and dissemination of disinformation are both legally protected, and platforms lack incentives to remove such content in the absence

---

913 Hal Roberts, Ethan Zuckerman and John G Palfrey, '2011 Circumvention Tool Evaluation' (SSRN Electronic Journal2011) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1940455> accessed 26 January 2025.

914 Lu and others (n 910) 310.

915 Brandenburg v Ohio, 395 US 444 (1969).

916 Virginia v Black, 538 US 343 (2003).

917 New York Times Co v Sullivan, 376 US 254 (1964).

of moderation policies[918]. Consequently, permissive domestic protections facilitate the cross-border spread of disinformation, exporting American free speech norms to jurisdictions with stricter regulatory environments[919].

As selected jurisdictions attempt to address this transnational phenomenon using legal tools designed for their own national contexts, this results in a fragmented regulatory environment. As Kuner points out in the context of data protection, even laws and regulations enacted with a strictly territorial scope can have significant global impacts in practice[920]. This insight is particularly relevant in the regulation of disinformation, as the transnational flow of online content means that measures taken in one jurisdiction often affect the information environment beyond its borders[921].

5.1.2 Strengthening Report and Appeal Mechanisms as Procedural Safeguards

While the EU, US, and China have adopted different regulatory philosophies regarding the governance of disinformation by online platforms, the comparative analysis presented above suggests that limited convergence at the procedural level is feasible[922]. Although the selected jurisdictions may find it difficult to reach consensus on substantive content moderation standards, they can still enhance cooperation in governing cross-border disinformation through the establishment of procedural safeguards.

First, online platforms should be required to provide accessible reporting mechanisms for users to flag disinformation(including unlawful or harmful content), as well as appeal mechanisms for users whose uploaded content is restricted or removed. The EU

---

[918] Ashley Johnson and Daniel Castro, 'Fact-Checking the Critiques of Section 230: What Are the Real Problems?' (*itif.org*22 February 2021) <https://itif.org/publications/2021/02/22/fact-checking-critiques-section-230-what-are-real-problems/> accessed 10 July 2025.

[919] Daphne Keller, 'Internet Platforms: Observations on Speech, Danger, and Money' (papers.ssrn.com13 June 2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3262936> accessed 4 August 2025.

[920] Christopher Kuner, 'Extraterritoriality and Regulation of International Data Transfers in EU Data Protection Law' (Social Science Research Network30 August 2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2644237> accessed 4 August 2025.

[921] Christopher Kuner, 'The Internet and the Global Reach of EU Law' (papers.ssrn.com1 February 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2890930> accessed 4 August 2025.

[922] Natali Helberger, Jo Pierson and Thomas Poell, 'Governing Online Platforms: From Contested to Cooperative Responsibility' (2018) 34 The Information Society 1, 4.

has institutionalized such safeguards under the DSA, which requires platforms to provide users with user-friendly reporting systems for illegal or harmful content, ensuring that platforms can promptly identify problematic content that might otherwise be overlooked, thereby promoting effective enforcement of content management standards [923]. Furthermore, platforms are obliged to establish internal complaint handling systems, explain the reasons for removing content, and promptly inform users of available remedies (Articles 17 and 20)[924]. Besides, Article 21 of the DSA explicitly provides that if users (including those who submit report notices) are dissatisfied with a platform's decision, including content removal, account restrictions, or rejection of a report, and have exhausted the platform's internal remedies, they may seek out-of-court dispute settlement [925]. This mechanism demonstrates that users have procedural remedies in platform content governance, enabling them to obtain authorized external remedies even without going to court.

In the US, requiring online platforms to establish complaint mechanisms for users does not conflict with the First Amendment's protection of free speech. The US Constitution protects free speech by restricting the actions of government entities, not private companies(including the online platforms)[926], thus their content moderation standards are not subject to the First Amendment. Complaint mechanisms do not further restrict or censor speech; rather, they provide users with an opportunity to challenge platform decisions. When a user's objection is addressed by an online platform, their content may be restored, increasing rather than narrowing the diversity of online opinion[927]. The contemporary speech governance structure has shifted to a triangular structure of "state-private online platforms-speakers"[928], where end users can influence the state and

---

[923] Luca Nannini and others, 'Beyond Phase-In: Assessing Impacts on Disinformation of the EU Digital Services Act' (2024) 5 AI and Ethics.

[924] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, arts 17 and 20.

[925] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, art 21.

[926] Manhattan Community Access Corp v Halleck, 139 S Ct 1921 (2019).

[927] Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (papers.ssrn.com30 June 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3639234> accessed 4 August 2025.

[928] Jack M Balkin, 'Free Speech Is a Triangle' (Ssrn.com2018)

online platforms through their speech and actions[929]. Within this framework, users exercise their power through the platform's complaint and appeal system, enhancing procedural fairness while also safeguarding constitutional free speech. In China, complaint mechanisms are legally mandated by a series of regulations, including the Cybersecurity Law (Article 47[930]) and the Regulations on the Governance of the Online Information Content Ecosystem (Article 16[931]), which require platforms to establish channels for users to report harmful content. However, "complaints" here primarily serve as initial feedback channels to platforms, without providing users with the right to request a secondary review or an independent complaints process. While laws and regulations don't explicitly require platforms to establish feasible appeal systems, some large platforms have already provided channels for consumers or users to file complaints. E-commerce platforms, such as Taobao[932], have relatively robust self-regulatory systems, providing an effective consumer complaint mechanism for resolving online disputes.

The above analysis demonstrates that online platforms in all three jurisdictions have the ability and motivation to establish and improve user information complaint mechanisms. The EU and China even explicitly require specific platforms to establish complaint and appeal systems in their laws and regulations. Effective remedies for erroneous removal constitute the cornerstone of platform liability and can be embedded in various regulatory systems without undermining their constitutional or political foundations[933]. While the scope of platform liability varies significantly across jurisdictions, three selected jurisdictions recognize that users should have some form of report and appeal mechanism after decisions on their platforms (such as account suspension or content deletion) are made. Therefore, procedurally strengthening complaint and appeal

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3186205> accessed 4 August 2025.

929 Jack M Balkin, 'Old School/New School Speech Regulation' (*Ssrn.com*2014) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2377526> accessed 4 August 2025.

930 Cybersecurity Law of the People's Republic of China 2016, art 47.

931 Provisions on the Governance of the Online Information Content Ecosystem, art 16.

932 Yang Lin, 'Self-Regulatory ODR in China's E-Commerce Market' (2025) 6 Amicus Curiae 358.

933 Helberger and others (n 922) 7.

mechanisms to handle user complaints regarding disinformation efficiently is both feasible and necessary in all three jurisdictions. User reporting and appeal mechanisms prioritize procedural fairness, making it feasible for the three jurisdictions to establish a cooperative framework in regulating AI-generated disinformation.

5.1.3 Enhancing Cross-Border Cooperation through Online Platform Transparency Reporting

Across jurisdictions, there is widespread agreement that platforms should be subject to some form of transparency obligation to enhance accountability for their content moderation practices[934]. Transparency reporting is considered a key governance tool, although the frequency of submission, the scope of scrutiny, and the level of rigor vary significantly across selected jurisdictions.

In the EU, the DSA introduced a comprehensive and enforceable system of transparency reporting. All online platforms must publish an annual transparency report (Article 24) detailing their proactive engagement in content moderation, the content of measures taken, the use of automated tools, and the outcome of internal complaints (Article 15)[935]. VLOPs and VLOSEs are subject to more stringent obligations, including systematic risk assessments and independent compliance audits, requiring them to proactively engage in the moderation of user-generated content and prevent the spread of disinformation[936]. Consequently, the EU has explicitly stipulated transparency reporting requirements for online platforms, mandating their regular publication within publicly accessible sections of online interfaces. These statutory obligations facilitate cross-jurisdictional cooperation by providing data support for the

---

[934] Josephine Wolff, 'Policy Approaches to Defining and Enforcing Responsibilities for Online Platforms', Defeating Disinformation (Cambridge University Press 2025) <https://www.cambridge.org/core/books/defeating-disinformation/policy-approaches-to-defining-and-enforcing-responsibilities-for-online-platforms/EFF7B8FAC2D22BD36CA860B97755679E> accessed 4 September 2025.

[935] Regulation (EU) 2022/2065 (Digital Services Act) [2022] OJ L277/1, arts 15 and 24.

[936] Julian Jaursch, 'Here Is Why Digital Services Coordinators Should Establish Strong Research and Data Units - DSA Observatory' (DSA Observatory10 March 2023) <https://dsa-observatory.eu/2023/03/10/here-is-why-digital-services-coordinators-should-establish-strong-research-and-data-units/> accessed 10 September 2025.

governance of cross-border disinformation dissemination.

In contrast, the US has no federal transparency reporting requirements. Section 230 of the CDA grants platforms immunity from liability but does not impose reporting or disclosure obligations. Instead, transparency is pursued through voluntary industry practices[937] (such as regular reports published by Meta, Google, and Twitter) and emerging state-level initiatives, including California's Content Moderation Transparency Act 2022[938]. These transparency reports typically include government requests for content removal, data access requests, and the removal of copyright-infringing content under the DMCA. For example, Google's Transparency Report regularly updates government requests for content removal and user data by country and request type[939]. Similarly, Meta released a transparency report detailing government requests for personal data, the amount of legally based content restrictions, and actions taken against content that violates its community standards, including for indicators such as hate speech, terrorist propaganda, and disinformation[940]. Twitter began publishing regular transparency reports on content removal in 2012, including reports received on false information, the number of egregious content removed, and the number of suspended accounts, to explain and provide feedback on Twitter's implementation of content moderation[941]. These initiatives reflect both a sense of social responsibility among online platforms[942] and a desire to improve their reputation by disclosing some of the results of their content moderation efforts, thereby reaping long-

---

[937] Amanda Reid, Evan Ringel and Shanetta M Pendleton, 'Transparency Reports as CSR Reports: Motives, Stakeholders, and Strategies' (2023) 20 Social Responsibility Journal 81.

[938] California Assembly Bill 587, Social Media Companies: Terms of Service (Content Moderation Transparency Act 2022), Cal Stat ch 8.

[939] Google, 'Google Transparency Report' (transparencyreport.google.com2025) <https://transparencyreport.google.com> accessed 10 September 2025.

[940] Meta Transparency Center, 'Integrity Reports, First Quarter 2025 | Transparency Center' (Meta.com2025) <https://transparency.meta.com/zh-cn/integrity-reports-q1-2025/> accessed 4 September 2025.

[941] Twitter, 'Evolving Our Twitter Transparency Report: Expanded Data and Insights' (X.com2018) <https://blog.x.com/en_us/topics/company/2018/evolving-our-twitter-transparency-report> accessed 4 September 2025.

[942] Amanda Reid, Shanetta M Pendleton and Lightning Ele Hulud JM Czabovsky, 'Social Media Transparency Reports: Longitudinal Content Analysis of News Coverage' (2024) 13 SSRN Electronic Journal 122, 123.

term benefits and maintaining loyal users[943]. Therefore, platforms' self-regulatory measures also demonstrate their willingness to disclose transparency reports directly to the public and users, thereby gaining users' trust and maintaining the platform's reputation[944].

While China mandates that platforms review and report illegal content, the regulations on how platforms should implement these measures are overly vague and lack transparency requirements. Article 30 of the Data Security Law [945] and Article 55 of the Personal Information Protection Law[946] require data security assessments or personal information protection assessments in the case of high-risk data processing or cross-border data transfers, but these reports are only submitted to the competent authorities and are not disclosed to the public. Under the Regulations on the Governance of the Online Information Content Ecosystem (Article 10)[947], platforms must establish content governance systems and report to the competent authorities on measures taken against illegal and harmful content. These obligations focus on state oversight rather than user accountability and do not require platforms to publish transparency reports[948]. However, these regulations ensure that online platforms maintain detailed governance data and produce reports for review by national regulators[949]. While these reports and filings are not publicly available, they could theoretically serve as official channels for providing necessary information to other national regulators in cross-border regulatory cooperation[950].

---

[943] Elettra Bietti, 'A Genealogy of Digital Platform Regulation' (papers.ssrn.com3 June 2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3859487> accessed 4 August 2025.
[944] Reid, Pendleton and Czabovsky (n 942) 124-125.
[945] Data Security Law of the People's Republic of China 2021, art 30.
[946] Personal Information Protection Law of the People's Republic of China 2021, art 55.
[947] Provisions on the Governance of the Online Information Content Ecosystem, art 10.
[948] Evelyn Douek, 'The Limits of International Law in Content Moderation' (2021) 6 UC Irvine Journal of International, Transnational, and Comparative Law <https://escholarship.org/uc/item/2857f1jq> accessed 4 September 2025.
[949] Jian Xu, 'Opening the "Black Box" of Algorithms: Regulation of Algorithms in China' (2024) 10 Communication Research and Practice 288.
[950] Weixing Shen and Yun Liu, 'China's Normative Systems for Responsible AI: From Soft Law to Hard Law' (Cambridge University Press2024) 150 <https://www.cambridge.org/core/books/cambridge-handbook-of-responsible-artificial-intelligence/chinas-normative-systems-for-responsible-ai/25A6636116359C1282F5874434CF467C> accessed 13 September 2025.

In summary, while the EU, the US, and China all require transparency in platform governance, the extent and content of the obligations they impose differ. The EU implements binding, user-facing reporting requirements, the US relies primarily on voluntary transparency measures by online platforms, while China merely mandates that online platforms provide internal reports on their information processing to higher authorities, lacking public accountability[951]. These three distinct institutional designs each have their own strengths and weaknesses, but they can complement each other in cross-border cooperation. The EU's transparency has established a verifiable basis for cross-border cooperation, the US's voluntary transparency reports provide flexibility and supplementary information, and China's centralized regulatory mechanism ensures information integrity. Despite differences in openness and accountability, they can serve as a common starting point for cross-border regulatory cooperation. Even if specific approaches differ, the parties can still establish a foundation for collaboration on this basis. Therefore, future cross-border cooperation does not require the complete alignment of transparency standards among the three parties; instead, it can achieve complementarity through the development of a minimal common framework.

5.2 Conclusion

Given the widespread cross-border dissemination of disinformation, I have analyzed the necessity of promoting cross-jurisdictional cooperative governance and the potential for exercising jurisdiction over disinformation disseminated abroad in accordance with its laws and regulations. I have recommended that the three jurisdictions strengthen reporting and appeal mechanisms to ensure procedural fairness and improve the effectiveness of disinformation governance through user reporting. Moreover, given that all three jurisdictions impose varying degrees of requirements on transparency reporting by online platforms, I have proposed to establish cross-jurisdictional cooperation on this basis to achieve complementary standards of

[951] Hao Xiaoming, Kewen Zhang and Huang Yu, 'The Internet and Information Control: The Case of China' (1996) 3 Javnost - the Public 117.

transparency.

6. Conclusion

This thesis has provided a comparative analysis of how the European Union, the United States, and China regulate AI-generated disinformation and how they assign liabilities to online platforms. These three jurisdictions all possess rapidly developing AI technologies, vast digital markets, and the potential to be affected by the spread of disinformation. This thesis aims to systematically analyze and compare the laws and regulations governing platform liability in these three jurisdictions. By examining the evolution of these laws and regulations, I have explored the respective legal development models of each jurisdiction and the factors influencing their development. Also, I have evaluated the effectiveness of these laws and regulations in practice and explored the causes of enforcement difficulties faced by online platforms.

Also, this thesis has provided a detailed analysis and answers to the research questions raised in the introduction.

First, this thesis has examined how national policies, legal traditions, and the balance of protected interests influence the content moderation obligations imposed by different jurisdictions. The research shows that while the EU's DSA imposes mandatory content moderation obligations only on the specific platforms (VLOPs), it also allows and encourages platforms to proactively develop content moderation policies and requires them to adhere to standards of transparency, fairness, and compliance. This demonstrates that DSA strikes a balance between protecting users' fundamental rights, encouraging technological innovation, and avoiding excessive regulation. In the United States, under the First Amendment and Section 230, laws and regulations favor the protection of free speech, granting platforms significant immunity against user-generated disinformation. However, statutory exceptions to Section 230 and the varying interpretations of Section 230 in case law demonstrate that this immunity is not absolute, but rather depends on the platform's contribution to the generation and dissemination of disinformation. In China, the government leads information regulation and administrative agencies serve as coordinators; national policies significantly influence

both its laws and regulations, as well as the content moderation policies of online platforms themselves. As a civil law system, China has progressively issued several departmental documents targeting AI-generated content, encouraging platforms to proactively regulate content and remove or block disinformation.

Secondly, this thesis set out to illustrate the platform liability attribution principles adopted by different jurisdictions and analyze the reasons for adopting these principles. The key to determining the platform liability principle lies in determining whether the platform's liability requires fault and the extent of fault, which in turn is closely related to the legal provisions of different jurisdictions. The choice of attribution principle determines whether the platform's duty of care also requires knowledge of the existence of disinformation as a triggering threshold.

Finally, this thesis sets out to offer feasible recommendations for promoting cross-jurisdictional collaborative regulations. Since regulatory measures to ensure procedural fairness generally do not involve ideological conflicts, and all three jurisdictions recognize the necessity of establishing platform reporting and appeal mechanisms and requiring platform transparency, there is no conflict of regulatory intent. In practical implementation, cross-jurisdictional cooperative regulation does not require the three jurisdictions to harmonize their content moderation standards. Instead, it involves progressively establishing a cross-jurisdictional collaborative mechanism through minimal coordination of procedural rules. Therefore, these two proposals are feasible in practice.

Bibliography

Table of Cases

United Kingdom

Donoghue v Stevenson [1932] AC 562 (HL)
Robinson v Chief Constable of West Yorkshire Police [2018] UKSC 4, [2018] AC 736

European Union

Google Spain v AEPD and Mario Costeja González, Case C-131/12, EU:C:2014:317
Joined Cases C-236/08 to C-238/08 Google France SARL and Google Inc v Louis Vuitton Malletier SA and Others [2010] ECR I-2417
L'Oréal SA v eBay International AG (C-324/09) [2011] ECR I-6011

United States

A&M Records Inc v Napster Inc, 239 F 3d 1004 (9th Cir 2001)
Anderson v TikTok Inc, No 22-3061 (3rd Cir, 27 August 2024)
Barnes v Yahoo! Inc, 570 F 3d 1096 (9th Cir 2009)
Blumenthal v Drudge, 992 F Supp 44 (D DC 1998)
Brandenburg v Ohio, 395 US 444 (1969)
Doe v MySpace Inc, 528 F 3d 413 (5th Cir 2008)
Force v Facebook Inc, 934 F 3d 53 (2d Cir 2019)
Global-Tech Appliances Inc v SEB SA, 563 US 754 (2011)
Gonzalez v Google LLC, 143 S Ct 1191 (2023)
Malwarebytes Inc v Enigma Software Group USA LLC, 946 F 3d 1040 (9th Cir 2020)
Manhattan Community Access Corp v Halleck, 139 S Ct 1921 (2019)
New York Times Co v Sullivan, 376 US 254 (1964)
Twitter Inc v Taamneh, 598 US ___ (2023)
Virginia v Black, 538 US 343 (2003)
Zeran v America Online Inc, 129 F 3d 327 (4th Cir 1997)

Table of Legislation

European Union

Artificial Intelligence Act (Regulation (EU) 2024/1689)
Code of Practice on Disinformation (2018)
Code of Practice on Disinformation (Strengthened 2022)
Directive 85/374/EEC on Liability for Defective Products [1985] OJ L210/29
Directive 2000/31/EC on Electronic Commerce [2000] OJ L178/1 (E-commerce Directive)
Directive 2024/2853/EU on Liability for Defective Products [2024] OJ L, pending publication
European Commission, Tackling Online Disinformation: A European Approach COM (2018) 236 final
General Data Protection Regulation (Regulation (EU) 2016/679) (GDPR)
Principles of European Tort Law (European Group on Tort Law, 2005)
Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act) [2022] OJ L277/1
Regulation (EU) 2022/1925 on Contestable and Fair Markets in the Digital Sector (Digital Markets Act) [2022] OJ L265/1
Austrian Data Protection Act (Datenschutzgesetz, DSG) 1978, § 33

United States

California Assembly Bill No 730, Ch 493 (2019)
California Assembly Bill No 2013, Ch 817 (2024)
California Assembly Bill No 2355, Ch 260 (2024)
Communications Decency Act, 47 USC § 230 (1996)
Digital Millennium Copyright Act, 17 USC § 512 (1998)
Florida Senate Bill 7072, Social Media Platforms, Regular Session 2021 (enacted 24 May 2021) https://www.flsenate.gov/Session/Bill/2021/7072 accessed 29 April 2025
Minnesota Stat § 211B.075 (2023)
New Mexico House Bill 182, 56th Leg, 1st Sess (2024)
Wire Fraud Act, 18 USC § 1343

China

Administrative Provisions on Deep Synthesis in Internet-based Information Services (《互联网信息服务深度合成管理规定》) (promulgated 2022, effective 10 January 2023).
Civil Code of the People's Republic of China (《中华人民共和国民法典》) (promulgated 28 May 2020, effective 1 January 2021).
Cybersecurity Law of the People's Republic of China (《中华人民共和国网络安全法》) (promulgated 7 November 2016, effective 1 June 2017).
Data Security Law of the People's Republic of China (《中华人民共和国数据安全法》) (promulgated 10 June 2021, effective 1 September 2021).

Interim Measures for the Management of Generative Artificial Intelligence Services (《生成式人工智能服务管理暂行办法》) (promulgated 13 July 2023, effective 15 August 2023).
Personal Information Protection Law of the People's Republic of China (《中华人民共和国个人信息保护法》) (promulgated 20 August 2021, effective 1 November 2021).
Provisions on Administration of Algorithmic Recommendation in Internet Information Services (《互联网信息服务算法推荐管理规定》) (promulgated 31 December 2021, effective 1 March 2022).
Provisions on the Governance of the Online Information Content Ecosystem (《网络信息内容生态治理规定》) (promulgated 15 December 2019, effective 1 March 2020).

Secondary Sources

Ahmed A and Khan MN, 'AI and Content Moderation: Legal and Ethical Approaches to Protecting Free Speech and Privacy' (*ResearchGate*2 September 2024) <https://www.researchgate.net/publication/383661951_AI_and_Content_Moderation_Legal_and_Ethical_Approaches_to_Protecting_Free_Speech_and_Privacy> accessed 5 October 2025

AI Index Steering Committee (Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terfa Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi and Sukrut Oak), 'The AI Index 2025 Annual Report'(Institute for Human-Centered AI, Stanford University April 2025) <https://hai.stanford.edu/ai-index/2025-ai-index-report> accessed 4 June 2025

Akers J and others, 'Technology-Enabled Disinformation: Summary, Lessons, and Recommendations' [2019] arXiv:1812.09383 [cs] <https://arxiv.org/abs/1812.09383> accessed 14 April 2025

Alam I ibne, 'AI Detectors Are Broken: Here's the Proof - Imtiaz Ibne Alam - Medium' (*Medium*26 May 2025) <https://imtiazibnealam.medium.com/ai-detectors-are-broken-heres-the-proof-e9556d94f634> accessed 12 July 2025

Albara AA, 'The Concept of Gatekeeping in Information Science: A Philosophical Reflection' (2018) 8 Global Journal of Information Technology: Emerging Technologies 16

Alessia Zornetta, 'Is the Digital Services Act Truly a Transparency Machine?' (*Tech Policy Press*11 July 2024) <https://www.techpolicy.press/is-the-digital-services-act-truly-a-transparency-machine/>

Algorithmic Accountability & New Technology, 'Enhancing Algorithmic Transparency' (*Center for News, Technology & Innovation*2024) <https://innovating.news/article/enhancing-algorithmic-transparency/> accessed 2 December 2024

Ali M and others, 'Discrimination through Optimization' (2019) 3 Proceedings of the ACM on Human-Computer Interaction 1 <https://dl.acm.org/doi/10.1145/3359301>

Allen C, Wallach W and Smit I, 'Why Machine Ethics?' (2006) 21 IEEE Intelligent Systems 12

Al-Maamari A, ‘Between Innovation and Oversight: A Cross-Regional Study of AI Risk Management Frameworks in the EU, U.S., UK, and China’ (*arXiv.org*2025) <https://arxiv.org/abs/2503.05773> accessed 4 June 2025

Ananny M, ‘Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance’ (*knightcolumbia.org*2019) <https://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance> accessed 8 July 2025

Ananthaswamy A, ‘The Physics Principle That Inspired Modern AI Art | Quanta Magazine’ (*Quanta Magazine*5 January 2023) <https://www.quantamagazine.org/the-physics-principle-that-inspired-modern-ai-art-20230105/> accessed 5 December 2024

Anantrasirichai N and Bull D, ‘Artificial Intelligence in the Creative Industries: A Review’ (2021) 55 Artificial Intelligence Review 589 <https://link.springer.com/article/10.1007/s10462-021-10039-7>

Anderljung M and Hazell J, ‘Protecting Society from AI Misuse: When Are Restrictions on Capabilities Warranted? | GovAI’ (*Governance.ai*2023) <https://www.governance.ai/research-paper/protecting-society-from-ai-misuse-when-are-restrictions-on-capabilities-warranted> accessed 22 November 2024

Anderson M and Susan Leigh Anderson, ‘Machine Ethics: Creating an Ethical Intelligent Agent’ (2019) 28 AI Magazine 15 <https://www.aaai.org/ojs/index.php/aimagazine/article/view/2065>

Angelopoulos C, ‘On Online Platforms and the Commission’s New Proposal for a Directive on Copyright in the Digital Single Market’ (papers.ssrn.com1 January 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2947800> accessed 3 February 2025

Angrist J and others, ‘Inputs and Impacts in Charter Schools: KIPP Lynn’ (2010) 100 American Economic Review: Papers & Proceedings 1 <https://users.nber.org/~dynarski/KIPP_Lynn.pdf>

Archibugi D, Filippetti A and Frenz M, ‘The Impact of the Economic Crisis on Innovation: Evidence from Europe’ (2013) 80 Technological Forecasting and Social Change 1247

Arcila BB, ‘Systemic Risks in the DSA and Its Enforcement’ (*DSA Decoded*2024) <https://www.dsadecoded.com/systemic-risks-in-the-dsa-and-its-enforcement> accessed 4 August 2025

Arora A and others, ‘Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go’ (2023) 56 ACM Computing Surveys 1

Artime O and others, ‘Effectiveness of Dismantling Strategies on Moderated vs. Unmoderated Online Social Platforms’ (2020) 10 Scientific Reports

Associated Press, ‘New Bipartisan Bill Would Require Online Identification, Labeling of AI-Generated Videos and Audio’ (*US News & World Report*2024) <https://www.usnews.com/news/us/articles/2024-03-21/new-bipartisan-bill-would-require-online-identification-labeling-of-ai-generated-videos-and-audio> accessed 10 July 2025

Ayer B, ‘Symbolic AI vs Machine Learning in Natural Language Processing’ (*Inbenta*4 March 2020) <https://www.inbenta.com/articles/symbolic-ai-vs-machine-learning-in-natural-language-processing/> accessed 5 December 2024

Bailey KC, ‘Regulating ISPs in the Age of Technology Exceptionalism | Texas Law Review’ (*Texas Law Review*4 May 2020) <https://texaslawreview.org/regulating-isps-in-the-age-of-technology-exceptionalism/> accessed 3 July 2025

Bain & Company, ‘Chinese Retailers Invest in Generative AI to Boost Performance’ (*Bain*2024) <https://www.bain.com/about/media-center/press-releases/2024/chinese-retailers-invest-in-generative-ai-to-boost-performance/> accessed 5 October 2025

Balkı B, ‘Online Platforms - Concurrences’ (*www.concurrences.com*2025) <https://www.concurrences.com/en/dictionary/online-platforms> accessed 25 March 2025

Balkin JM, ‘Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society’ (2004) 79 SSRN Electronic Journal

Balkin JM, ‘Fixing Social Media’s Grand Bargain’ (*papers.ssrn.com*15 October 2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3266942> accessed 4 August 2025

Balkin JM, ‘Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation’ (*papers.ssrn.com*9 September 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3038939> accessed 4 August 2025

Balkin JM, ‘Free Speech Is a Triangle’ (*Ssrn.com*2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3186205> accessed 4 August 2025

Balkin JM, ‘Information Fiduciaries and the First Amendment’ (*papers.ssrn.com*3 February 2016) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2675270> accessed 14 April 2025

Balkin JM, ‘Old School/New School Speech Regulation’ (*Ssrn.com*2014) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2377526> accessed 4 August 2025

Balland P-A and others, ‘Generative AI and Foundation Models in the EU: Uptake, Opportunities, Challenges, and a Way Forward ’ (2025) <https://cdn.ceps.eu/wp-content/uploads/2025/03/EESC_report_Generative-AI-and-founding-models-in-the-EU.pdf> accessed 5 October 2025

Banker E, ‘A Review of Section 230’S Meaning & Application Based on More than 500 Cases’ (*Archiveia.org*2020) <https://archiveia.org/publications/a-review-of-section-230s-meaning-application-based-on-more-than-500-cases/> accessed 29 April 2025

Bansal N and others, ‘Real-Time Advanced Computational Intelligence for Deep Fake Video Detection’ (2023) 13 Applied Sciences 3095

Bar C von, *The Common European Law of Torts: Volume One*, vol. 2 (Oxford University Press 1998) <https://academic.oup.com/book/37023> accessed 7 October 2025

Barker K, Jurasz O and Law School S, ‘Online Harms White Paper Consultation Response’ (2019) <https://oro.open.ac.uk/69840/1/Barker%20&%20Jurasz%20-%20Online%20Harms%20White%20Paper%20Consultation%20Response%20(2019)%20.pdf> accessed 18 June 2025

Barlow JP, ‘A Declaration of the Independence of Cyberspace’ (*Electronic Frontier Foundation*8 February 1996) <https://www.eff.org/cyberspace-independence> accessed 13 September 2025

Barrett PM, ‘NYU Stern Center for Business & Human RightsSafeguarding AI: Addressing the Risks of Generative Artificial Intelligence’ (*Nyu.edu*25 June 2025) <https://bhr.stern.nyu.edu/publication/safeguarding-ai-addressing-the-risks-of-generative-artificial-intelligence/> accessed 10 July 2025

Barzilai-Nahon K, ‘Toward a Theory of Network Gatekeeping: A Framework for Exploring Information Control’ (2008) 59 Journal of the American Society for Information Science and Technology 1493

Bassini M, ‘Fundamental Rights and Private Enforcement in the Digital Age’ (2019) 25 European Law Journal 182

Bateman J and Jackson D, ‘Countering Disinformation Effectively: An Evidence-Based Policy Guide’ (*Carnegie Endowment for International Peace*31 January 2024) <https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en> accessed 2 December 2024

BBC NEWS, ‘Meta Requires Political Advertisers to Mark When Deepfakes Used’ *BBC News* (9 November 2023) <https://www.bbc.co.uk/news/technology-67366311> accessed 18 August 2025

BBC News, ‘TikTok and Twitch Face Fines under New Ofcom Rules’ *BBC News* (5 October 2021) <https://www.bbc.co.uk/news/technology-58809169> accessed 18 August 2025

BBC Trending, ‘“Pizzagate”: The Fake Story That Shows How Conspiracy Theories Spread’ (*BBC News*2 December 2016) <https://www.bbc.co.uk/news/blogs-trending-38156985> accessed 5 December 2024

Beauvais C, ‘Fake News: Why Do We Believe It?’ (2022) 89 Joint Bone Spine

Belenguer L, ‘AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry’ (2022) 2 AI and Ethics

Belen-Saglam R and others, ‘A Systematic Literature Review of the Tension between the GDPR and Public Blockchain Systems’ (2023) 4 Blockchain: Research and Applications 100129 <https://www.sciencedirect.com/science/article/pii/S2096720923000040>

Bello P and Bringsjord S, ‘On How to Build a Moral Machine’ (2012) 32 Topoi 251

Bender E and others, ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ’ [2021] FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610 <https://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf> accessed 2021

Benditt TM, ‘Law and the Balance of Interests’ (1975) 3 Social Theory and Practice 321 <https://www.jstor.org/stable/23557739>

Berger I and Torres T, ‘What Is Trust and Safety? A Comprehensive Guide’ (*Activefence.com*2024) <https://www.activefence.com/what-is-trust-and-safety/> accessed 30 May 2025

Bergmann D and Stryker C, ‘What Is a Variational Autoencoder? | IBM’ (*www.ibm.com*12 June 2024) <https://www.ibm.com/think/topics/variational-autoencoder> accessed 5 December 2024

Bietti E, ‘A Genealogy of Digital Platform Regulation’ (*papers.ssrn.com*3 June 2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3859487> accessed 4 August 2025

Blevins JF, ‘The Use and Abuse of “Light-Touch” Internet Regulation’ (2018) 99 SSRN Electronic Journal 177

Bode K, ‘FCC, State Action Nets an Amazing 80% Reduction in Auto Warranty Scam Robocalls’ (*Techdirt*25 August 2022) <https://www.techdirt.com/2022/08/25/fcc-state-action-nets-an-amazing-80-reduction-in-auto-warranty-scam-robocalls/> accessed 1 May 2025

Bonadio E, ‘Trade Marks in Online Marketplaces: The CJEU’s Stance in L’Oreal v. EBay’ (*Ssrn.com*7 March 2012) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2017741> accessed 4 August 2025

Bond-Taylor S and others, ‘Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models’ (2021) 44 IEEE Transactions on Pattern Analysis and Machine Intelligence 1 <https://arxiv.org/abs/2103.04922>

Bontcheva K and others, ‘Contributing Authors Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities Editor’ (2024) <https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf> accessed 2 December 2024

Bontridder N and Poullet Y, ‘The Role of Artificial Intelligence in Disinformation’ (2021) 3 Data & Policy <https://www.cambridge.org/core/journals/data-and-policy/article/role-of-artificial-intelligence-in-disinformation/7C4BF6CA35184F149143DE968FC4C3B6>

Bowers J and Zittrain J, ‘Answering Impossible Questions: Content Governance in an Age of Disinformation’ (2020) 1 Harvard Kennedy School Misinformation Review

Boyd A and others, ‘The Value of AI Guidance in Human Examination of Synthetically-Generated Faces’ (*arXiv.org*2022) <https://arxiv.org/abs/2208.10544> accessed 4 June 2025

Bozhkov N, ‘China’s Cyber Diplomacy: A Primer ’ (*Horizon*2020) <https://eucyberdirect.eu/research/chinas-cyber-diplomacy-a-primer> accessed 10 July 2025

Bradford A, ‘The Chinese State-Driven Regulatory Model’, *Digital Empires: the Global Battle to Regulate Technology* (Oxford University Press 2023) <https://academic.oup.com/book/46736/chapter-abstract/418514383?redirectedFrom=fulltext&login=false>

Bradford A, *The Brussels Effect: How the European Union Rules the World* (Oxford University Press 2020) <https://scholarship.law.columbia.edu/books/232/>

Bradford B and others, ‘Report of the Facebook Data Transparency Advisory Group’ (2019) <https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf> accessed 10 July 2025

Britton A, ‘The Interplay between Section 230 Immunity and the Allow States and Victims to Fight Online Sex Trafficking Act of 2018 - Weintraub Tobin’ (*Weintraub Tobin*10 November 2022) <https://www.weintraub.com/2022/11/the-interplay-between-section-230-immunity-and-the-allow-states-and-victims-to-fight-online-sex-trafficking-act-of-2018/> accessed 27 September 2025

Brown M, ‘Rethinking Section 230: Fostering Transparency, Accountability, and User Protection Online – Denver Journal of International Law & Policy’ (*Djilp.org*17 November 2024) <https://djilp.org/rethinking-section-230-fostering-transparency-accountability-and-user-protection-online/?utm_.com> accessed 29 April 2025

Brown S, ‘Machine Learning, Explained’ (*MIT Sloan*21 April 2021) <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>

Brown TB and others, ‘Language Models Are Few-Shot Learners’ (2020) 4 arxiv.org 1 <https://arxiv.org/abs/2005.14165>

Brownstein M and Jennifer Mather Saul, *Implicit Bias and Philosophy* (Oxford University Press 2016)

Brownstein M, ‘Attributionism and Moral Responsibility for Implicit Bias’ (2015) 7 Review of Philosophy and Psychology 765

Buchanan B and others, ‘Truth, Lies, and Automation’ (Center for Security and Emerging TechnologyMay 2021) <https://cset.georgetown.edu/publication/truth-lies-and-automation/> accessed 5 January 2025

Buchanan BG, ‘A (Very) Brief History of Artificial Intelligence’ (2005) 26 AI Magazine 53 <https://www.aaai.org/ojs/index.php/aimagazine/article/view/1848>

Büchi M, Festic N and Latzer M, ‘The Chilling Effects of Digital Dataveillance: A Theoretical Model and an Empirical Research Agenda’ (2022) 9 Big Data & Society

Buckland MK, ‘Information as Thing’ (1991) 42 Journal of the American Society for Information Science

Buiten MC, ‘OUP Accepted Manuscript’ (2020) 9 Journal of Antitrust Enforcement 270

Buiten MC, ‘The Digital Services Act: From Intermediary Liability to Platform Regulation’ (2021) 12 Journal of Intellectual Property, Information Technology and E-Commerce Law 361 <https://www.jipitec.eu/jipitec/article/view/331>

Bumble, ‘Bumble’s Community Guidelines | Bumble’ (*Bumble*2025) <https://bumble.com/en/guidelines?.com> accessed 21 June 2025

Buonocore T, ‘Man Is to Doctor as Woman Is to Nurse: The Gender Bias of Word Embeddings’ (*Medium*8 March 2019) <https://medium.com/towards-data-science/gender-bias-word-embeddings-76d9806a0e17> accessed 11 February 2025

Burgsdorff C von and Kramer L, ‘Increased Liability due to the New EU Product Liability Directive: What Does This Mean for the Medical and Pharmaceutical Industry?’ (*Ibanet.org*2025) <https://www.ibanet.org/increased-liability-eu-product-directive> accessed 9 April 2025

Burrell J, ‘How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms’ (2016) 3 Big Data & Society 1

Busch C, ‘Platform Responsibility in the European Union’, *Defeating Disinformation* (Cambridge University Press 2025) <https://www.cambridge.org/core/books/defeating-disinformation/platform-responsibility-in-the-european-union/AA3D55C57B0F6A7C18F5CAEF25146557>

Busch C, ‘Regulating the Expanding Content Moderation Universe: A European Perspective on Infrastructure Moderation’ (2022) 27 UCLA Journal of Law & Technology

Busch C, ‘Self-Regulation and Regulatory Intermediation in the Platform Economy’ (*Ssrn.com*30 November 2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3309293> accessed 22 April 2025

Cane P and Goudkamp J, *Atiyah's Accidents, Compensation and the Law* (Cambridge University Press 2018) 85

Cao J and others, 'Consumers' Risk Perception, Market Demand, and Firm Innovation: Evidence from China' (2024) 19 PloS One

Cao Y and others, 'A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT' (2023) 37 arXiv (Cornell University)

Capper S, 'Cyber Criminals: Being Anonymous Online' (*Darkinvader.io*2024) <https://www.darkinvader.io/blogs/cyber-criminals-being-anonymous-online> accessed 5 December 2024

Castelvecchi D, 'Can We Open the Black Box of AI?' (2016) 538 Nature 20 <https://www.nature.com/articles/doi:10.1038/538020a>

Castillo APD, 'Exposing Generative AI | Etui' (*Etui*2024) <https://www.etui.org/publications/exposing-generative-ai> accessed 12 February 2025

Castle R, 'Lord Atkin and the Neighbour Test: Origins of the Principles of Negligence in Donoghue v Stevenson' (2003) 7 Ecclesiastical Law Journal 210 <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/CBCF36E5E5998EB037E232CAAE3317ED/S0956618X00005214a.pdf/lord_atkin_and_the_neighbour_test_origins_of_the_principles_of_negligence_in_donoghue_v_stevenson.pdf>

Cerwick M, 'Censoring Social Media: Texas HB 20' (*Vanderbilt University*2021) <https://www.vanderbilt.edu/jetlaw/2021/10/06/censoring-social-media-texas-hb-20/>

Chander A, 'How Law Made Silicon Valley' (*Ssrn.com*15 August 2013) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2340197> accessed 7 July 2025

Chase PH, Fellow S and The German Marshall Fund of the United States, 'The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem †' (2019) <https://www.ivir.nl/publicaties/download/EU_Code_Practice_Disinformation_Aug_2019.pdf>

Chatterjee S, 'Rules for Independent Audits under the EU's Digital Services Act (DSA)' (*Holisticai.com*2023) <https://www.holisticai.com/blog/rules-for-independent-audits-digital-services-act> accessed 1 July 2025

Chen B and Chen J, ‘China’s Legal Practices Concerning Challenges of Artificial General Intelligence’ (2024) 13 Laws 60 <https://www.mdpi.com/2075-471X/13/5/60>

Chen J and Shi C, ‘Proactive Governance by Official Administrators on Chinese Social Media Platforms: Boundary Discourse and Governance Legitimacy’ [2025] Media Culture & Society

Chen N, ‘Solving the New Contradiction of “Threading a Thousand Threads with One Needle”——Review and Prospect of Research on the Reform of Grass-Roots Comprehensive Administrative Law Enforcement’ (2024) 6 Study on Party and Government

Chen Q, ‘China’s Emerging Approach to Regulating General-Purpose Artificial Intelligence: Balancing Innovation and Control | Asia Society’ (*asiasociety.org*7 February 2024) <https://asiasociety.org/policy-institute/chinas-emerging-approach-regulating-general-purpose-artificial-intelligence-balancing-innovation-and> accessed 5 October 2025

Cheng E, ‘Big Chinese Companies like Alibaba Show That AI-Powered Ads Are Giving Shopping a Boost’ (*CNBC*16 May 2025) <https://www.cnbc.com/2025/05/16/chinese-companies-like-alibaba-see-more-consumption-helped-by-ai-ads.html> accessed 1 October 2025

Cherry K, ‘How Does Implicit Bias Influence Behavior?’ (*Verywell Mind*2023) <https://www.verywellmind.com/implicit-bias-overview-4178401> accessed 18 January 2025

Chesterman S, ‘Lawful but Awful: Evolving Legislative Responses to Address Online Misinformation, Disinformation, and Mal-Information in the Age of Generative AI’ (*arXiv.org*2025) <https://arxiv.org/abs/2505.15067?utm.com> accessed 28 May 2025

Cho CY and Zhu L, ‘Defining and Regulating Online Platforms’ (*Congress.gov*2025) <https://www.congress.gov/crs-product/R47662> accessed 25 March 2025

Chun J, Schroeder C and Elkins K, ‘Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US’ (*arXiv.org*2024) <https://arxiv.org/abs/2410.21279> accessed 2 November 2024

Church P and Pehlivan CN, ‘The Digital Services Act (DSA): A New Era for Online Harms and Intermediary Liability’ (2023) 4 Global Privacy Law Review <https://kluwerlawonline.com/journalarticle/Global+Privacy+Law+Review/4.1/GPLR2023005> accessed 3 August 2025

Church Z, ‘Study: False News Spreads Faster than the Truth | MIT Sloan’ (*MIT Sloan*8 March 2018) <https://mitsloan.mit.edu/ideas-made-to-matter/study-false-news-spreads-faster-truth> accessed 2 December 2024

Cinelli M and others, ‘The COVID-19 Social Media Infodemic’ (2020) 10 Scientific Reports <https://www.nature.com/articles/s41598-020-73510-5>

Citron D and Wittes B, ‘Fordham Law Review the Internet Will Not Break: Denying Bad Samaritans § 230 Immunity’ (2017) <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5435&context=flr> accessed 4 August 2025

Citron D and Wittes B, ‘The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity’ (2017) 86 Fordham Law Review 401 <https://ir.lawnet.fordham.edu/flr/vol86/iss2/3/> accessed 4 August 2025

Coaston J, ‘Alex Jones Banned from YouTube, Facebook, and Apple, Explained’ (*Vox*6 August 2018) <https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-youtube-conspiracy-theories> accessed 7 July 202

Cohen H, ‘CRS Report for Congress Freedom of Speech and Press: Exceptions to the First Amendment’ (2009) <https://resources.saylor.org/wwwresources/archived/site/wp-content/uploads/2014/01/POLSC401-3.1-FreedomofSpeechandPress-PublicDomain.pdf>

Cohen T, ‘Regulating Manipulative Artificial Intelligence’ (2023) 20 SCRIPTed 203 <https://script-ed.org/article/regulating-manipulative-artificial-intelligence/>

Cohent M and Sundararajantt A, ‘Self-Regulation and Innovation in the Peer-To-Peer Sharing Economy’ (2015) <https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1039&context=uclrev_online>

Coleman JL, ‘Fault and Strict Liability’, *Risks and Wrongs* (Oxford University Press 2002)

Colomina C, Margalef S and Youngs R, ‘The Impact of Disinformation on Democratic Processes and Human Rights in the World ’ (2021) <https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf> accessed 24 January 2025

Comito C, Caroprese L and Zumpano E, ‘Multimodal Fake News Detection on Social Media: A Survey of Deep Learning Techniques’ (2023) 13 Social Network Analysis and Mining

Cong W, ‘The Spatial Expansion of China’s Digital Sovereignty’, *Digital Sovereignty in the BRICS Countries* (Cambridge University Press 2024) <https://www.cambridge.org/core/books/digital-sovereignty-in-the-brics-countries/spatial-expansion-of-chinas-digital-sovereignty/D18D59288E33063AF1EBAF416FACD6A2> accessed 10 July 2025

Cook T, ‘Online Intermediary Liability in the European Union’ (2012) 17 Journal of Intellectual Property Rights 157 <https://www.researchgate.net/publication/295162458_Online_intermediary_liability_in_the_European_Union>

Corbitt BJ, Thanasankit T and Yi H, ‘Trust and E-Commerce: A Study of Consumer Perceptions’ (2003) 2 Electronic Commerce Research and Applications 203

Cordella A and Gualdi F, ‘Regulating Generative AI: The Limits of Technology-Neutral Regulatory Frameworks. Insights from Italy’s Intervention on ChatGPT’ (2024) 41 Government Information Quarterly 101982

Cornell Law School, ‘Legal Information Institute (Cornell LII)’ (*LII / Legal Information Institute*2024) <https://www.law.cornell.edu/wex/tort>

Crawford K and Paglen T, ‘Excavating AI: The Politics of Images in Machine Learning Training Sets’ (2021) 36 AI & SOCIETY 1105

Creemers R and Wright ND, ‘The International and Foreign Policy Impact of China?S Artificial Intelligence and BigData Strategies’ (Air University Press 2019) <https://www.jstor.org/stable/resrep19585.23> accessed 8 August 2025

Creemers R, ‘China’s Conception of Cyber Sovereignty: Rhetoric and Realization’ [2020] SSRN Electronic Journal

Crémer J, De Montjoye Y-A and Schweitzer H, ‘Digital Era a Report by Competition Policy Competition’ (2019) <https://euagenda.eu/upload/publications/untitled-257961-ea.pdf> accessed 8 August 2025

Dai J and Qin Y, ‘The Ideological Risks of Generative Artificial Intelligence such as ChatGPT and Its Response’ (2023) 29 Journal of Chongqing University (Social Science Edition)

Danry V, Leong J, Pataranutaporn P, Tandon P, Liu Y, Shilkrot R, Punpongsanon P, Weissman T, Maes P and Sra M, ‘AI-Generated Characters: Putting Deepfakes to Good Use’ in CHI Conference on Human Factors in Computing Systems Extended Abstracts (ACM 2022).

Davie G, ‘Gatekeeping Theory’ (*Mass Communication Theory*2 November 2018) <https://masscommtheory.com/theory-overviews/gatekeeping-theory/> accessed 13 May 2025

de Almeida Leite EM and Ramos Leite MA, ‘Platform Liability, Free Speech, and Market Fairness: Transatlantic Legal Responses to Commercial Defamation and Digital Competition’ [2025] International Review of Law, Computers & Technology 1

De Gregorio G and Radu R, ‘Digital Constitutionalism in the New Era of Internet Governance’ (2022) 30 International Journal of Law and Information Technology

De Gregorio G, ‘The Rise of Digital Constitutionalism in the European Union’ (2021) 19 International Journal of Constitutional Law 41

Del Vicario M and others, ‘The Spreading of Misinformation Online’ (2016) 113 Proceedings of the National Academy of Sciences 554 <https://www.pnas.org/doi/full/10.1073/pnas.1517441113>

Deng H and Wang X, ‘Risks and Responses to the Disinformation Governance of Generative AI’ [2024] Theory Monthly

Dennehy F, ‘Almost 90% of Young People Exposed to Harmful Content on Social Media’ (*The Alan Turing Institute*2023) <https://www.turing.ac.uk/news/almost-90-young-people-exposed-harmful-content-social-media> accessed 18 August 2025

Derner E and Batistič K, ‘Beyond the Safeguards: Exploring the Security Risks of ChatGPT’ (*arXiv.org*13 May 2023) <https://arxiv.org/abs/2305.08005> accessed 4 June 2025

Devlin J and others, ‘BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding’ (*ArXiv*11 October 2018) <https://arxiv.org/abs/1810.04805>

Dhariwal P and Nichol A, ‘Diffusion Models Beat GANs on Image Synthesis’ [2021] arXiv:2105.05233 [cs, stat] <https://arxiv.org/abs/2105.05233>

Dhiman B and Singh P, ‘Exploding AI-Generated Deepfakes and Misinformation: A Threat to Global Concern in the 21st Century’ (*SSRN*7 December 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4651093> accessed 23 November 2024

Dholakia U, ‘Algorithmic Transparency and Consumer Disclosure’ [2023] Springer eBooks 135

Di Domenico G and Ding Y, 'Between Brand Attacks and Broader Narratives: How Direct and Indirect Misinformation Erode Consumer Trust' (2023) 54 Current Opinion in Psychology 101716 <https://www.sciencedirect.com/science/article/pii/S2352250X23001616#:~:text=Fake%20customer%20reviews%20constitute%20another> accessed 16 November 2023

Di Sotto S and Viviani M, 'Health Misinformation Detection in the Social Web: An Overview and a Data Science Approach' (2022) 19 International Journal of Environmental Research and Public Health 2173

Diakopoulos N, 'Algorithmic Accountability' (2014) 3 Digital Journalism 398

Diamond L and Schell O, *China's Influence and American Interests* (Hoover Press 2019)

Dickinson GM, 'An Interpretive Framework for Narrower Immunity under Section 230 of the Communications Decency Act' (*arXiv.org*2023) <https://arxiv.org/abs/2306.04461>

Dimitrieska S, 'Generative Artificial Intelligence and Advertising' (2024) 6 Trends in Economics, Finance and Management Journal 23 <https://tefmj.ibupress.com/uploads/2024/07/ibu_journal_tefmj-3.pdf>

Dinwoodie GB, 'A Comparative Analysis of the Secondary Liability of Online Service Providers' (*Ssrn.com*17 May 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2997891> accessed 24 August 2025

Director T, 'From Immunity to Regulation: Turning Point of Internet Intermediary Regulatory Agenda' (*The Journal of Law and Technology at Texas*8 October 2016) <https://jolttx.com/2016/10/08/immunity-regulation-turning-point-internet-intermediary-regulatory-agenda/> accessed 15 April 2025

Directorate-General for Communications Networks and Content and Technology (European Commission), 'A Multi-Dimensional Approach to Disinformation : Report of the Independent High Level Group on Fake News and Online Disinformation.' (*Europa.eu*30 April 2018) <https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1/language-en> accessed 17 May 2025

DiResta R, 'AI-Generated Text Is the Scariest Deepfake of All' (*Wired*31 July 2020) <https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/> accessed 2 November 2024

DLA Piper, ‘What Is the Digital Services Act? | DLA Piper’ (*DLA Piper*2023) <https://www.dlapiper.com/en/insights/publications/2023/06/what-is-the-digital-services-act> accessed 4 June 2025

Dondl W and Zunke M, ‘How to Protect Your Machine Learning Models | Thales’ (*cpl.thalesgroup.com*2024) <https://cpl.thalesgroup.com/blog/software-monetization/how-to-protect-your-machine-learning-models> accessed 5 December 2024

Douek E, ‘Governing Online Speech: From “Posts-As-Trumps” to Proportionality and Probability’ (2020) 121 SSRN Electronic Journal

Douek E, ‘The Limits of International Law in Content Moderation’ (2021) 6 UC Irvine Journal of International, Transnational, and Comparative Law <https://escholarship.org/uc/item/2857f1jq> accessed 4 September 2025

Douek E, ‘What Kind of Oversight Board Have You given Us? | the University of Chicago Law Review’ (*Uchicago.edu*2025) <https://lawreview.uchicago.edu/online-archive/what-kind-oversight-board-have-you-given-us>

Draper D, ‘Section 230- Are Online Platforms Publishers, Distributors, or Neither? | Bipartisan Policy Center’ (*bipartisanpolicy.org*13 March 2023) <https://bipartisanpolicy.org/blog/section-230-online-platforms/> accessed 5 December 2024

Duffield W, ‘Circumventing Section 230: Product Liability Lawsuits Threaten Internet Speech’ (*Cato Institute*26 January 2021) <https://www.cato.org/policy-analysis/circumventing-section-230-product-liability-lawsuits-threaten-internet-speech?> accessed 19 October 2025

Duffy BE and others, ‘The Nested Precarities of Creative Labor on Social Media’ (2021) 7 Social Media + Society 1

Eder N, ‘Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation’ (2024) 25 German Law Journal 1

Eennnaam J van , ‘The New Platform Liability: From the E-Commerce Directive to the Digital Services Act Regulation (“ DSA ”)’ (*WiseMen Advocaten*2023) <https://www.wisemen.nl/en/news/the-new-platform-liability-from-the-e-commerce-directive-to-the-digital-services-act-regulation-dsa-/> accessed 13 May 2025

Ehlen T, ‘The Digital Services Act: New Liability Rules?’ (*Passle*10 July 2023) <https://technologyquotient.freshfields.com/post/102iiyf/the-digital-services-act-new-liability-rules>

Eisenstein J, Ahmed A and Xing EP, ‘Sparse Additive Generative Models of Text’ [2011] International Conference on Machine Learning 1041

Elizabeth Carney, ‘Protecting Internet Freedom at the Expense of Facilitating Online Child Sex Trafficking? An Explanation as to Why CDA’s Section 230 Has No Place in a New NAFTA’ (2019) 68 Catholic University Law Review 353 <https://scholarship.law.edu/lawreview/vol68/iss2/8/>

Elswah M, Bhatia A and Thakur D, ‘Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages’ (*Center for Democracy and Technology*28 June 2025) <https://cdt.org/insights/content-moderation-in-the-global-south-a-comparative-study-of-four-low-resource-languages/> accessed 11 July 2025

Emerson TI, ‘The Doctrine of Prior Restraint’ (1955) 20 Law and Contemporary Problems 648 <https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=2658&context=lcp> accessed 10 July 2020

Emes CS, ‘Exploring New Frontiers in Digital Governance: Addressing the Ambiguities of Risk-Based Regulation Approach for Platforms’ <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5242418> accessed 4 August 2025

ESAFETY COMMISSIONER, ‘Anonymity and Identity Shielding’ (*ESafety Commissioner*25 July 2023) <https://www.esafety.gov.au/industry/tech-trends-and-challenges/anonymity> accessed 11 March 2025

EU DisinfoLab, ‘When Platforms Make Mistakes, Users Need Redress’ (*EU DisinfoLab*2022) <https://www.disinfo.eu/publications/when-platforms-make-mistakes-users-need-redress/> accessed 10 July 2025

European Commission, ‘Digital Services Act: Commission Launches Transparency Database | Shaping Europe’s Digital Future’ (*digital-strategy.ec.europa.eu*26 September 2023) <https://digital-strategy.ec.europa.eu/en/news/digital-services-act-commission-launches-transparency-database>

European Commission, ‘Generative AI Set to Transform EU Economy but Requires Further Policy Action’ (2025) <https://digital-strategy.ec.europa.eu/en/news/generative-ai-set-transform-eu-economy-requires-further-policy-action> accessed 5 October 2025

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' (*Europa.eu*2021) <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>

European Commission, 'Supervision of the Designated Very Large Online Platforms and Search Engines under DSA | Shaping Europe's Digital Future' (*digital-strategy.ec.europa.eu*2025) <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses> accessed 18 August 2025

European Commission, 'The Digital Markets Act: Ensuring Fair and Open Digital Markets' (*European Commission*2022) <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en>

European Commission, 'The Digital Services Act Package | Shaping Europe's Digital Future' (*digital-strategy.ec.europa.eu*2022) <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>

European Parliament, 'AT a GLANCE: Digital Issues in Focus, European Parliamentary Research Service' (2024) <https://www.europarl.europa.eu/RegData/etudes/ATAG/2024/760392/EPRS_ATA(2024)760392_EN.pdf> accessed 2 November 2024

Fallis D, 'What Is Disinformation?' (2015) 63 Library Trends 401

Fan J, 'Douyin: Information on Events That May Trigger "Cyber Violence" and "Opening of Boxes" Will Be Included in the "Controversial" Hot Spots for Analysis and Treatment' (*Thepaper.cn*May 2025) <https://m.thepaper.cn/detail/30886170> accessed 14 July 2025

Faris R and others, 'Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election' (*papers.ssrn.com*1 August 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3019414> accessed 4 August 2025

Fathaigh RÓ, Helberger N and Appelman N, 'The Perils of Legally Defining Disinformation' (2021) 10 Internet Policy Review <https://policyreview.info/articles/analysis/perils-legally-defining-disinformation>

Fatima A, 'How Has Artificial Intelligence Evolved from Symbolic AI to Deep Learning?' (*Machine Mindscape*31 January 2024) <https://machinemindscape.com/artificial-intelligence-to-deep-learning-history-concepts/> accessed 21 January 2025

Federal Communications Commission , 'FCC SETTLES CASE against PROVIDER THAT TRANSMITTED SPOOFED AI-GENERATED ROBOCALLS for

ELECTION INTERFERENCE in NEW HAMPSHIRE Lingo Telecom to Pay $1 Million Civil Penalty and Implement First-of-Their-Kind Compliance Terms Secured by the FCC' (2024) <https://docs.fcc.gov/public/attachments/DOC-404951A1.pdf> accessed 1 May 2025

Federal Communications Commission, 'Disclosure and Transparency of Artificial Intelligence-Generated Content in Political Advertisements' (*Federal Register*5 August 2024) <https://www.federalregister.gov/documents/2024/08/05/2024-16977/disclosure-and-transparency-of-artificial-intelligence-generated-content-in-political-advertisements>

Federal Communications Commission, 'FCC & State AGs Warn of Student Loan Debt Scam Robocalls & Robotexts' (2023) <https://www.fcc.gov/document/fcc-state-ags-warn-student-loan-debt-scam-robocalls-robotexts> accessed 1 May 2025

Federal Communications Commission, 'FCC Settles Spoofed AI-Generated Robocalls Case' (*Fcc.gov*21 August 2024) <https://www.fcc.gov/document/fcc-settles-spoofed-ai-generated-robocalls-case> accessed 13 September 2025

Federal Communications Commission, 'Federal Communications Commission FCC 24-74 NOTICE of PROPOSED RULEMAKING' (2024) <https://docs.fcc.gov/public/attachments/FCC-24-74A1.pdf> accessed 1 May 2025

Felin T and Holweg M, 'Theory Is All You Need: AI, Human Cognition, and Decision Making' [2024] Social Science Research Network <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4737265> accessed 18 May 2024

Felipe Romero Moreno, 'Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content' (2024) 38 International Review of Law Computers & Technology 1

Fellow J, Researcher R and Hanson L, 'Algorithmic Accountability: A Primer' (2018) <https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL.pdf> accessed 18 January 2025

Fernández M, Bellogín A and Cantador I, 'Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation' [2021] arXiv:2103.14748 [cs] <https://arxiv.org/abs/2103.14748>

Ferrara E and others, 'The Rise of Social Bots' (2016) 59 Communications of the ACM 96 <https://dl.acm.org/doi/abs/10.1145/2818717>

Ferrer X and others, ‘Bias and Discrimination in AI: A Cross-Disciplinary Perspective’ (2021) 40 IEEE Technology and Society Magazine 72 <https://ieeexplore.ieee.org/abstract/document/9445793>

Feuerriegel S and others, ‘Generative AI’ (2023) 66 Business & Information Systems Engineering 111 <https://link.springer.com/article/10.1007/s12599-023-00834-7>

Fichtner L, ‘Content Moderation and the Quest for Democratic Legitimacy’ (2024) 4 Weizenbaum Journal of the Digital Society <https://ojs.weizenbaum-institut.de/index.php/wjds/article/view/2_2_2>

Filippi D and Lavayssière X, ‘Blockchain Technology: Toward a Decentralized Governance of Digital Platforms?’ (*Ssrn.com*5 December 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3760483> accessed 4 August 2025

Finck M, ‘Digital Co-Regulation: Designing a Supranational Legal Framework for the Platform Economy’ (*papers.ssrn.com*20 June 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2990043> accessed 4 August 2025

Fishman JP, ‘Section 230 as First Amendment Rule’ (2018) 131 Harvard Law Review <https://harvardlawreview.org/print/vol-131/section-230-as-first-amendment-rule/>

Fleming M, ‘How AI Is Boosting Disinformation’ (*www.linkedin.com*2024) <https://www.linkedin.com/pulse/how-ai-boosting-disinformation-melissa-fleming-vbdpe/> accessed 24 January 2025

Floridi L and others, ‘AI4People—an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’ (2018) 28 Minds and Machines 689 <https://link.springer.com/article/10.1007/s11023-018-9482-5>

Freshfields Bruckhaus Deringer, ‘2025 Data Law Trends | Freshfields’ (2025) <https://www.freshfields.com/en/our-thinking/campaigns/data-trends-2025>

Frosio G and Geiger C, ‘Taking Fundamental Rights Seriously in the Digital Services Act’s Platform Liability Regime’ (2023) 29 European Law Journal 31

Frosio G, ‘From the E-Commerce Directive to the Digital Services Act’ (*SSRN*2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4914816> accessed 4 August 2025

Frosio G, ‘Platform Responsibility in the Digital Services Act: Constitutionalising, Regulating and Governing Private Ordering’ (*Queen’s University Belfast*October

2023) <https://pure.qub.ac.uk/en/publications/platform-responsibility-in-the-digital-services-act-constitutiona> accessed 30 August 2025

Frosio GF, 'Reforming Intermediary Liability in the Platform Economy: A European Digital Single Market Strategy' (*Northwestern Pritzker School of Law Scholarly Commons*2017) <https://scholarlycommons.law.northwestern.edu/nulr_online/251/> accessed 18 September 2025

Fuchs C, *Social Media: A Critical Introduction* (SAGE Publications Ltd 2014)

Fung B, 'FCC Issues Historic $300 Million Fine against the Largest Robocall Scam It Has Ever Investigated' (*CNN*4 August 2023) <https://www.cnn.com/2023/08/04/tech/fcc-robocall-scam-biggest-fine/index.html> accessed 13 September 2025

Gary J and Soltani A, 'First Things First: Online Advertising Practices and Their Effects on Platform Speech' (*knightcolumbia.org*2019) <https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech> accessed 13 September 2025

Gheisari S and others, 'A Combined Convolutional and Recurrent Neural Network for Enhanced Glaucoma Detection' (2021) 11 Scientific Reports

Ghosh D and Scott B, 'Digital Deceit II: A Policy Agenda to Fight Disinformation on the Internet' (*Shorenstein Center*2 October 2018) <https://shorensteincenter.org/digital-deceit-ii-policy-agenda-fight-disinformation-internet/> accessed 25 March 2025

Gillespie T, 'Content Moderation, AI, and the Question of Scale' (2020) 7 Big Data & Society 1 <https://doi.org/10.1177/2053951720943234>

Gillespie T, 'Platforms Intervene' (2015) 1 Social Media + Society

Gillespie T, Boczkowski PJ and Foot KA, 'The Relevance of Algorithms' in Gillespie T and others (eds), Media Technologies: Essays on Communication, Materiality, and Society (MIT Press 2013)

Gillespie T, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*(Yale University Press 2018)

Ginsburg JC, 'Separating the Sony Sheep from the Grokster Goats: Reckoning the Future Business Plans of Copyright-Dependent Technology Entrepreneurs' (2008) 50 Arizona Law Review <https://journals.librarypublishing.arizona.edu/arizlrev/article/id/7378/> accessed 15 October 2025

Girdhar R and others, ‘Emu Video: Factorizing Text-To-Video Generation by Explicit Image Conditioning’ [2023] ArXiv (Cornell University)

Gisonna N, ‘Great Firewall | History, China, Hong Kong, & Facts | Britannica’ (*www.britannica.com*27 July 2023) <https://www.britannica.com/topic/Great-Firewall>

Giulia Giannasi, ‘Risk in the Digital Services Act and AI Act: Implications for Media Freedom, Pluralism, and Disinformation - Centre for Media Pluralism and Media Freedom’ (*Centre for Media Pluralism and Media Freedom*27 May 2025) <https://cmpf.eui.eu/risk-in-the-digital-services-act-and-ai-act-implications-for-media-freedom-pluralism-and-disinformation/> accessed 18 August 2025

Goldberg J, School H and Zipursky B, ‘Fordham Law Review Fordham Law Review the Strict Liability in Fault and the Fault in Strict Liability the Strict Liability in Fault and the Fault in Strict Liability Recommended Citation Recommended Citation’ (2016) <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5250&context=flr> accessed 4 August 2025

Goldman E, ‘Why Section 230 Is Better than the First Amendment’ (2019) 95 SSRN Electronic Journal

Goldsmith J, ‘Who Controls the Internet? Illusions of a Borderless World’ (2007) 23 Strategic Direction

Golpayegani D, Pandit HJ and Lewis D, ‘To Be High-Risk, or Not to Be—Semantic Specifications and Implications of the AI Act’s High-Risk AI Applications and Harmonised Standards’ [2023] Zenodo (CERN European Organization for Nuclear Research)

Goodfellow I and others, ‘Generative Adversarial Networks’ (2020) 63 Communications of the ACM 139

Google, ‘Google Transparency Report’ (*transparencyreport.google.com*2025) <https://transparencyreport.google.com> accessed 10 September 2025

Gori P, ‘The Strengthened Code of Practice on Disinformation – Many Stakeholders, One Goal - MediaLaws’ (*MediaLaws*9 January 2023) <https://www.medialaws.eu/the-strengthened-code-of-practice-on-disinformation-many-stakeholders-one-goal/> accessed 10 June 2025

Gorwa R, Binns R and Katzenbach C, ‘Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance’ (2020) 7 Big Data & Society <https://journals.sagepub.com/doi/10.1177/2053951719897945>

Gosztonyi G, ‘How the European Union Had Tried to Tackle Fake News and Disinformation with Soft Law and What Changed with the Digital Services Act?’ (2024) 3 Frontiers in Law 102

Gosztonyi G, ‘The Contribution of the Court of Justice of the European Union to a Better Understanding the Liability and Monitoring Issues Regarding Intermediary Service Providers’ (2020) 59 Annales Universitatis Scientiarum Budapestinensis De Rolando Eötvös Nominatae. Sectio Iuridica 133

Gozalo-Brizuela R and Garrido-Merchán EC, ‘A Survey of Generative AI Applications’ (*arXiv.org*14 June 2023) <https://arxiv.org/abs/2306.02781>

Graber CB, ‘Bottom-up Constitutionalism: The Case of Net Neutrality’ (2017) 7 SSRN Electronic Journal

Grant Thornton Ireland, ‘Determining Gatekeepers under the Digital Markets Act’ (*Grant Thornton Ireland*22 October 2024) <https://www.grantthornton.ie/insights/factsheets/determining-gatekeepers-under-the-digital-markets-act/> accessed 4 April 2025

Greene J, ‘Free Speech on Public Platforms’ in Lee C Bollinger and Geoffrey R Stone (eds), Social Media, Freedom of Speech, and the Future of our Democracy (Oxford University Press 2022).

Greene J, ‘Rights as Trumps?’ (2018) 132 Harv. L. Rev. 28 <https://scholarship.law.columbia.edu/faculty_scholarship/2920/>

Gregorio GD and Demková S, ‘The Enforcement Dilemmas in Europe’s Digital Rulebook’ (*Tech Policy Press*19 May 2025) <https://www.techpolicy.press/the-enforcement-dilemmas-in-europes-digital-rulebook/> accessed 10 July 2025

Gregorio GD, ‘Digital Constitutionalism and Freedom of Expression’, *Digital Constitutionalism in Europe Reframing Rights and Powers in the Algorithmic Society* (Cambridge University Press 2022) <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/72ACEF48324D180E95BBD456E52E9C96/9781316512777c5_157-215.pdf/digital_constitutionalism_and_freedom_of_expression.pdf> accessed 10 July 2025

Griffin R, ‘Governing Platforms through Corporate Risk Management: The Politics of Systemic Risk in the Digital Services Act’ [2025] European Law Open 1 <https://www.cambridge.org/core/journals/european-law-open/article/governing-platforms-through-corporate-risk-management-the-politics-of-systemic-risk-in-the-digital-services-act/287159FD68134232851133FEFF451D42> accessed 15 July 2025

GRUR International, ‘Liability of E-Commerce Platform Operators for Taking Necessary Measures’ (2023) 72 GRUR International 566

Guerra A, Luppi B and Parisi F, ‘Do Presumptions of Negligence Incentivize Optimal Precautions?’ (2022) 54 European Journal of Law and Economics 349

Guo X, ‘Risks of Generative Artificial Intelligence and Its Inclusive Legal Governance’ (2023) 25 Journal of Beijing Institute of Technology(Socoal Sciences Edition) 93

Gupta P and others, ‘Generative AI: A Systematic Review Using Topic Modelling Techniques’ (2024) 8 Data and Information Management 100066 <https://www.sciencedirect.com/science/article/pii/S2543925124000020>

Hamilton IA, ‘Why It’s Totally Unsurprising That Amazon’s Recruitment AI Was Biased against Women’ (Business Insider13 October 2018) <https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10> accessed 18 August 2025

Hamilton JF, Bodle R and Korin E, Explorations in Critical Studies of Advertising (Routledge, Taylor & Francis Group 2019)

Harmeling T, ‘The Role of the Gatekeepers under the DMA Regulation’ (Consent Management Platform (CMP) Usercentrics2023) <https://usercentrics.com/knowledge-hub/role-of-gatekeepers-under-digital-markets-act/> accessed 18 August 2025

Hay Bruce L and Spier Kathryn E, ‘Burdens of Proof in Civil Litigation: An Economic Perspective’ (1997) 26 The Journal of Legal Studies 413

Helberger N, Pierson J and Poell T, ‘Governing Online Platforms: From Contested to Cooperative Responsibility’ (2018) 34 The Information Society 1

Heldt AP, ‘EU Digital Services Act: The White Hope of Intermediary Regulation’ [2022] Palgrave Macmillan 69

Henderson P, Hashimoto T and Lemley M, ‘Where’s the Liability in Harmful AI Speech?’ (*arXiv.org*2023) <https://arxiv.org/abs/2308.04635> accessed 17 April 2025

Heppell F, Bakir ME and Bontcheva K, ‘Lying Blindly: Bypassing ChatGPT’s Safeguards to Generate Hard-To-Detect Disinformation Claims’ (*arXiv.org*2024) <https://arxiv.org/abs/2402.08467> accessed 14 February 2025

Hickey KJ, ‘Digital Millennium Copyright Act (DMCA) Safe Harbor Provisions for Online Service Providers: A Legal Overview’ (*Congress.gov*2025) <https://www.congress.gov/crs-product/IF11478>

Hinton GE, Osindero S and Teh Y-W, 'A Fast Learning Algorithm for Deep Belief Nets' (2006) 18 Neural Computation 1527

Hoehndorf R and Queralt-Rosinach N, 'Data Science and Symbolic AI: Synergies, Challenges and Opportunities' (2017) 1 Data Science 27

Holdsworth J and Scapicchio M, 'Deep Learning' (*Ibm.com*17 June 2024) <https://www.ibm.com/think/topics/deep-learning> accessed 18 August 2025

Holznagel D, 'Shortcomings of the First DSA Audits — and How to Do Better - DSA Observatory' (*DSA Observatory - a Hub of Expertise on the DSA package.*11 June 2025) <https://dsa-observatory.eu/2025/06/11/shortcomings-of-the-first-dsa-audits-and-how-to-do-better/> accessed 30 June 2025

Honigberg B, 'The Existential Threat of AI-Enhanced Disinformation Operations' (*Just Security*8 July 2022) <https://www.justsecurity.org/82246/the-existential-threat-of-ai-enhanced-disinformation-operations/> accessed 5 January 2025

Hornkohl L, 'The Extraterritorial Application of Statutes and Regulations in EU Law' (2022) 1 SSRN Electronic Journal

Horsley JP, 'Behind the Facade of China's Cyber Super-Regulator' (*DigiChina*8 August 2022) <https://digichina.stanford.edu/work/behind-the-facade-of-chinas-cyber-super-regulator/>

Huang J and others, 'LARGE LANGUAGE MODELS CANNOT SELF-CORRECT REASONING YET' (2024) <https://arxiv.org/pdf/2310.01798>

Huang T, 'Content Moderation by LLM: From Accuracy to Legitimacy' (2025) 58 Artificial Intelligence Review

Huang Z and others, 'Understanding Self-Attention Mechanism via Dynamical System Perspective' (*arXiv.org*2023) <https://arxiv.org/abs/2308.09939> accessed 14 April 2025

Huscroft G, Miller BW and Webber G, 'Proportionality and the Rule of Law: Rights, Justification, Reasoning Introduction' (papers.ssrn.com8 May 2014) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2434504> accessed 20 September 2025

Husovec M, 'Amicus Curiae Delfi as v Estonia' (*Scribd*2025) <https://www.scribd.com/document/232759055/Amicus-Curiae-Delfi-AS-v-Estonia> accessed 16 August 2025

Husovec M, ‘Introduction to Liability Framework ’, *Principles of the Digital Services Act* (Oxford University Press 2024) <https://academic.oup.com/book/58088> accessed 29 November 2024

Husovec M, ‘The Digital Services Act’s Red Line: What the Commission Can and Cannot Do about Disinformation’ (2024) 16 Journal of Media Law 1

Husovec M, ‘The DSA as a Co-Regulatory System’ [2024] Oxford University Press eBooks 443 <https://academic.oup.com/oxford-law-pro/book/58088/chapter/478884189#515183574> accessed 28 June 2025

Hutchinson T and Duncan N, ‘Defining and Describing What We Do: Doctrinal Legal Research’ (2012) 17 Deakin Law Review 83

Ignatidou S, ‘Deepfakes, Shallowfakes and Speech Synthesis: Tackling Audiovisual Manipulation’ (*European Science-Media Hub*4 December 2019) <https://sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/> accessed 2 December 2024

International Telecommunication Union Development Sector, ‘Measuring Digital Development Facts and Figures 2023’ (2023) <https://www.itu.int/itu-d/reports/statistics/wp-content/uploads/sites/5/2023/11/Measuring-digital-development-Facts-and-figures-2023-E.pdf> accessed 25 March 2025

Internet Development Research Institution of Peking University, ‘China’s General Security Perception Report (2021) ’ (*Iqilu.com*2021) <https://news.iqilu.com/china/gedi/2021/1228/5031290.shtml> accessed 2 December 2024

Jackson F, ‘20% of Generative AI “Jailbreak” Attacks Are Successful’ (*TechRepublic*9 October 2024) <https://www.techrepublic.com/article/genai-jailbreak-report-pillar-security/> accessed 2 December 2024

Jackson VC, ‘Constitutional Law in an Age of Proportionality’ (2015) 124 Yale Law Journal <https://www.yalelawjournal.org/article/constitutional-law-in-an-age-of-proportionality> accessed 2 December 2024

JAFFER J, ‘Facebook and Free Speech Are Different Things’ (*knightcolumbia.org*2019) <https://knightcolumbia.org/content/facebook-and-free-speech-are-different-things> accessed 10 July 2025

Jahangir R, ‘The EU’s Code of Practice on Disinformation Is Now Part of the Digital Services Act. What Does It Mean?’ (*Tech Policy Press*25 February 2025) <https://www.techpolicy.press/the-eus-code-of-practice-on-disinformation-is-now-part-of-the-digital-services-act-what-does-it-mean/> accessed 10 July 2025

Janakbari PZJ, ‘Detection and Mitigation of Deepfake Attacks in Cybersecurity : Leveraging Computer Vision and Deep Learning’ (*Theseus.fi*2025) <https://www.theseus.fi/handle/10024/894608> accessed 10 July 2025

Janger EJ and Twerski AD, ‘Functional Tort Principles for Internet Platforms: Duty, Relationship, and Control | Yale Journal of Law & Technology’ (2025) 26 Yale Journal of Law & Technology <https://yjolt.org/functional-tort-principles-internet-platforms-duty-relationship-and-control>

Jankowich A, ‘EULAw: The Complex Web of Corporate Rule-Making in Virtual Worlds’ (2019) 8 Tulane Journal of Technology & Intellectual Property <https://journals.tulane.edu/TIP/article/view/2500> accessed 13 May 2025

Jaursch J, ‘Here Is Why Digital Services Coordinators Should Establish Strong Research and Data Units - DSA Observatory’ (*DSA Observatory*10 March 2023) <https://dsa-observatory.eu/2023/03/10/here-is-why-digital-services-coordinators-should-establish-strong-research-and-data-units/> accessed 10 September 2025

Jimenez L, ‘Tech Regulation Digest: Sunsetting Section 230—the Future of Content Moderation, Ads, and AI | Milken Institute’ (*Milken Institute*3 March 2025) <https://milkeninstitute.org/content-hub/collections/articles/tech-regulation-digest-sunsetting-section-230-future-content-moderation-ads-and-ai>

Jin X and others, ‘Assessing the Perceived Credibility of Deepfakes: The Impact of System-Generated Cues and Video Characteristics’ (2023) 27 New Media & Society

Johnson A and Castro D, ‘Fact-Checking the Critiques of Section 230: What Are the Real Problems?’ (*itif.org*22 February 2021) <https://itif.org/publications/2021/02/22/fact-checking-critiques-section-230-what-are-real-problems/> accessed 10 July 2025

Johnson A and Castro D, ‘Overview of Section 230: What It Is, Why It Was Created, and What It Has Achieved’ (*itif.org*22 February 2021) <https://itif.org/publications/2021/02/22/overview-section-230-what-it-why-it-was-created-and-what-it-has-achieved/> accessed 10 July 2025

Jones GH, Jones BH and Little P, ‘Reputation as Reservoir: Buffering against Loss in Times of Economic Crisis’ (2000) 3 Corporate Reputation Review 21

Jóźwiak M, ‘The DSA’s Systemic Risk Framework: Taking Stock and Looking Ahead’ (*Dsa-observatory.eu*2025) <https://dsa-observatory.eu/2025/05/27/the-dsas-systemic-risk-framework-taking-stock-and-looking-ahead/> accessed 15 July 2025

Justen L and others, ‘No Time like the Present: Effects of Language Change on Automated Comment Moderation’ (2022) 01 2022 IEEE 24th Conference on

Business Informatics (CBI) 40 <https://ieeexplore.ieee.org/document/9944746> accessed 14 July 2025

Kalyan KS, 'A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4' (*arXiv.org*2023) <https://arxiv.org/abs/2310.12321> accessed 14 April 2025

Karaş Z, 'Effects of AI-Generated Misinformation and Disinformation on the Economy' (2024) 12 Düzce Üniversitesi Bilim Ve Teknoloji Dergisi

KARNOWSKI S, 'Elon Musk's X Sues to Overturn Minnesota Political Deepfakes Ban' (*ABC News*25 April 2025) <https://abcnews.go.com/Technology/wireStory/elon-musks-sues-overturn-minnesota-political-deepfakes-ban-121173206>

Katz E, 'Ask the Expert: What Meta's New Fact-Checking Policies Mean for Misinformation and Hate Speech' (*MSUToday | Michigan State University*27 January 2025) <https://msutoday.msu.edu/news/2025/ask-the-expert-what-meta-new-fact-checking-policies-mean-for-misinformation-and-hate-speech>

Katz R and Tushman M, 'An Investigation into the Managerial Roles and Career Paths of Gatekeepers and Project Supervisors in a Major R & D Facility' (1981) 11 R&D Management 103

Kaur A and others, 'Deepfake Video Detection: Challenges and Opportunities' (2024) 57 Artificial Intelligence Review

Kaushal R and others, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database' (*arXiv.org*2024) <https://arxiv.org/abs/2404.02894> accessed 11 June 2025

Keating GC, 'The Idea of Fairness in the Law of Enterprise Liability' (1997) 95 Michigan Law Review 1266

Keeton P and Prosser WL, *Prosser and Keeton on Torts* (West Pub Co 1984)

Keller D, 'Internet Platforms: Observations on Speech, Danger, and Money' (*papers.ssrn.com*13 June 2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3262936> accessed 4 August 2025

Keller D, 'Systemic Duties of Care and Intermediary Liability' (*Stanford CIS*29 May 2020) <https://cyberlaw.stanford.edu/blog/2020/05/systemic-duties-care-and-intermediary-liability/> accessed 5 December 2024

Keller D, ‘Who Do You Sue? State and Platform Hybrid Power over Online Speech’ (*Lawfare*2019) <https://www.lawfaremedia.org/article/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech?.com> accessed 4 August 2025

Kennedy S, ‘Holding “Governance” Accountable’ (*Sheila Kennedy*3 May 2005) <https://sheilakennedy.net/2005/05/holding-governance-accountable/> accessed 3 July 2025

Kettemann MC and Schulz W, ‘Setting Rules for 2.7 Billion: A (First) Look into Facebook’s Norm-Making System; Results of a Pilot Study’ (2020) 1 Ssoar.info 34 <https://www.ssoar.info/ssoar/handle/document/71724>

Kim G, Baldi P and McAleer S, ‘Language Models Can Solve Computer Tasks’ (*arXiv.org*16 November 2023) <https://arxiv.org/abs/2303.17491> accessed 14 April 2025

Kim JJH and others, ‘Generative AI Can Effectively Manipulate Data’ (2024) 5 AI and Ethics

Kingma DP and Welling M, ‘Auto-Encoding Variational Bayes’ (*arXiv.org*20 December 2013) <https://arxiv.org/abs/1312.6114> accessed 14 April 2025

Klonick K, ‘The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression’ (*papers.ssrn.com*30 June 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3639234> accessed 4 August 2025

Klonick K, ‘The New Governors: The People, Rules, and Processes Governing Online Speech’ (*Harvard Law Review*10 April 2018) <https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/>

Kofman A, Tseng F and Weigel M, ‘The Hate Store: Amazon’s Self-Publishing Arm Is a Haven for White Supremacists’ *ProPublica* (16 April 2020) <https://www.propublica.org/article/the-hate-store-amazons-self-publishing-arm-is-a-haven-for-white-supremacists> accessed 7 July 202

Kohl U, ‘Toxic Recommender Algorithms: Immunities, Liabilities and the Regulated Self-Regulation of the Digital Services Act and the Online Safety Act’ (2024) 16 Journal of Media Law 1

Komendantova N and Erokhin D, ‘Artificial Intelligence Tools in Misinformation Management during Natural Disasters’ (2025) 25 Public Organization Review

Kosseff J, *The Twenty-Six Words That Created the Internet* (Cornell University Press 2019)

Kraakman RH, 'Gatekeepers: The Anatomy of a Third-Party Enforcement Strategy' (1986) 2 The Journal of Law, Economics, and Organization <https://academic.oup.com/jleo/article/2/1/53/873299>

Krack N, Dutkiewicz L and De Meyere J, 'Generative Artificial Intelligence and Disinformation' (*SSRN*2025) <https://ssrn.com/abstract=5192993> accessed 12 September 2025

Kreps S, McCain RM and Brundage M, 'All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation' (2020) 9 Journal of Experimental Political Science 1

Kuczerawy A, 'Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative' (*Ssrn.com*2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2560257> accessed 25 February 2019

Kuczerawy A, 'The Good Samaritan That Wasn't: Voluntary Monitoring under the (Draft) Digital Services Act' (*Verfassungsblog*12 January 2021) <https://verfassungsblog.de/good-samaritan-dsa/> accessed 5 December 2024

Kuczerawy A, 'To Monitor or Not to Monitor? The Uncertain Future of Article 15 of the E-Commerce Directive' (*CITIP Blog*10 July 2019) <https://www.law.kuleuven.be/citip/blog/to-monitor-or-not-to-monitor-the-uncertain-future-of-article-15-of-the-e-commerce-directive/> accessed 5 December 2024

Kumar AK, 'Situating Automated Decision-Making Jurisprudence within Data Protection Frameworks: A Study of Intersections between GDPR and EU Artificial Intelligence Act- Part II' (*Law School Policy Review*16 May 2024) <https://lawschoolpolicyreview.com/2024/05/16/situating-automated-decision-making-jurisprudence-within-data-protection-frameworks-a-study-of-intersections-between-gdpr-and-eu-artificial-intelligence-act-part-ii/> accessed 26 September 2025

Kumar K, 'Understanding Transformers: A Deep Dive into NLP's Core Technology' (*Analytics Vidhya*16 April 2024) <https://www.analyticsvidhya.com/blog/2024/04/understanding-transformers-a-deep-dive-into-nlps-core-technology/> accessed 5 December 2024

Kuner C, 'Extraterritoriality and Regulation of International Data Transfers in EU Data Protection Law' (*Social Science Research Network*30 August 2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2644237> accessed 4 August 2025

Kuner C, ‘The Internet and the Global Reach of EU Law’ (*papers.ssrn.com*1 February 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2890930> accessed 4 August 2025

Lagioia F and others, ‘AI in Search of Unfairness in Consumer Contracts: The Terms of Service Landscape’ (2022) 45 Journal of Consumer Policy 481

Lally C and Stepney ES, ‘Disinformation: Sources, Spread and Impact’ (*POST*16 October 2024) <https://post.parliament.uk/research-briefings/post-pn-0719/> accessed 2 December 2024

Lang M, ‘Blocked and Banned by Social Media: When Is It Censorship?’ (*SFChronicle.com*31 August 2016) <https://www.sfchronicle.com/business/article/Blocked-and-banned-by-social-media-When-is-it-9193998.php> accessed 8 July 2025

Larsen B and Dignum V, ‘AI Value Alignment: Aligning AI with Human Values’ (*World Economic Forum*17 October 2024) <https://www.weforum.org/stories/2024/10/ai-value-alignment-how-we-can-align-artificial-intelligence-with-human-values/> accessed 18 January 2025

Laskowski N, Tucci L and Craig L, ‘What Is Artificial Intelligence (AI)?’ (*TechTarget*2022) <https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence>

Laux J, Wachter S and Mittelstadt B, ‘Taming the Few: Platform Regulation, Independent Audits, and the Risks of Capture Created by the DMA and DSA’ (2021) 43 Computer Law & Security Review 105613

Law Teacher, ‘Donoghue v Stevenson [1932] Doctrine of Negligence’ (*Lawteacher.net*7 March 2018) <https://www.lawteacher.net/cases/donoghue-v-stevenson.php>

Lazar A and others, ‘Going Gray, Failure to Hire, and the Ick Factor’ [2017] Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing

Lazer DMJ and others, ‘The Science of Fake News’ (2018) 359 Science 1094 <https://www.science.org/doi/10.1126/science.aao2998>

LeCun Y, Bengio Y and Hinton G, ‘Deep Learning’ (2015) 521 Nature 436 <https://www.nature.com/articles/nature14539>

Lee E, ‘Moderating Content Moderation: A Framework for Nonpartisanship in Online Governance’ (2020) 70 American University Law Review

Lee NT, Resnick P and Barton G, ‘Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms’ (*Brookings*22 May 2019) <https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/> accessed 18 January 2025

Leerssen P, ‘The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems’ (*papers.ssrn.com*24 February 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3544009> accessed 4 August 2025

Leffer L, ‘Your Personal Information Is Probably Being Used to Train Generative AI Models’ (Scientific American19 October 2023) <https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/> accessed 10 November 2024

Lepri B and others, ‘Fair, Transparent, and Accountable Algorithmic Decision-Making Processes’ (2017) 31 Philosophy & Technology 611 <https://link.springer.com/article/10.1007/s13347-017-0279-x>

Levi SD and others, ‘California Enacts New Laws to Combat AI-Generated Deceptive Election Content’ (*Skadden.com*27 September 2024) <https://www.skadden.com/insights/publications/2024/09/california-enacts-new-laws>

Lewandowsky S, Ecker UKH and Cook J, ‘Beyond Misinformation: Understanding and Coping with the “Post-Truth” Era’ (2017) 6 Journal of Applied Research in Memory and Cognition 353 <https://www.sciencedirect.com/science/article/abs/pii/S2211368117300700>

Li B, ‘IAPP’ (*Iapp.org*2024) <https://iapp.org/news/a/china-issues-the-regulations-on-network-data-security-management-what-s-important-to-know>

Li J-B and others, ‘Chinese Public’s Panic Buying at the Beginning of COVID-19 Outbreak: The Contribution of Perceived Risk, Social Media Use, and Connection with Close Others’ (2021) 41 Current Psychology

Li L and Zhou K, ‘When Content Moderation Is Not about Content: How Chinese Social Media Platforms Moderate Content and Why It Matters’ (2024) 27 New Media & Society 6150

Li X and others, ‘A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules’ (2025) 53 The Annals of Statistics

Li Y, Chang M-C and Lyu S, 'In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking' (2018) <https://arxiv.org/pdf/1806.02877> accessed 2 November 2024

Liang M, 'The End of Social Media? How Data Attraction Model in the Algorithmic Media Reshapes the Attention Economy' (2022) 44 Media, Culture & Society

Lin LSF, 'Organisational Challenges in US Law Enforcement's Response to AI-Driven Cybercrime and Deepfake Fraud' (2025) 14 Laws 46 <https://www.mdpi.com/2075-471X/14/4/46>

LIN Y, 'Self-Regulatory ODR in China's E-Commerce Market' (2025) 6 Amicus Curiae 358

Lipinski TA, Buchanan EA and Britz JJ, 'Sticks and Stones and Words That Harm: Liability vs. Responsibility, Section 230 and Defamatory Speech in Cyberspace' (2002) 4 Ethics and Information Technology 143

Liu BF, Bartz L and Duke N, 'Communicating Crisis Uncertainty: A Review of the Knowledge Gaps' (2016) 42 Public Relations Review 479

Liu J and Cui B, 'Data Localization and the Legitimacy and Trend of Data Defensivism' (2020) 12 Global Review 89

Liu J, 'Internet Censorship in China: Looking through the Lens of Categorisation' (2024) 0 Journal of Current Chinese Affairs

Liu J, 'Regulatory Framework for New Risks of Large Generative AI Models' [2024] Administrative Law Review 17

Liu R and Liu F, 'Analysis of the Human-Machine Collaborative Governance Network of Short Video Content Ecology' (2024) 12 Youth Journalist

Liu S and others, 'DSCLAP: Domain-Specific Contrastive Language-Audio Pre-Training' [2024] arXiv (Cornell University)

Liu X, 'Normative Construction of Platform Criminal Liability in the Governance of Deepfake Technology' (2025) 16 Advances in Social Behavior Research 40

Liu Y, 'Research on Algorithm Bias and Its Regulation Approach' (2019) 40 Law Science Magazine

Llansó E, 'Artificial Intelligence, Content Moderation, and Freedom of Expression †' (2020) <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 18 January 2025

Llansó EJ, ‘No Amount of “AI” in Content Moderation Will Solve Filtering’s Prior-Restraint Problem’ (2020) 7 Big Data & Society

London Intercultural Academy, ‘The Story of ELIZA: The AI That Fooled the World’ (*London Intercultural Academy*2024) <https://liacademy.co.uk/the-story-of-eliza-the-ai-that-fooled-the-world/>

Longpre S and others, ‘A Large-Scale Audit of Dataset Licensing and Attribution in AI’ (2024) 6 Nature Machine Intelligence 975

Longpre S and others, ‘Data Authenticity, Consent, & Provenance for AI Are All Broken: What Will It Take to Fix Them?’ (*arXiv.org*2024) <https://arxiv.org/abs/2404.12691> accessed 14 April 2025

Lu A, ‘Generative Artificial Intelligence: Exploring Risk, Regulation,and Governance Models ’ (*Cnki.net*2024) <http://kns.cnki.net/kcms/detail/50.1180.C.20240628.0838.004.html> accessed 23 November 2024

Lu P, Zhou L and Fan X, ‘Platform Governance and Sociological Participation’ (2023) 10 The Journal of Chinese Sociology

Lu S, ‘Algorithmic Opacity, Private Accountability, and Corporate Social Disclosure in the Age of Artificial Intelligence’ (2020) 23 Vanderbilt Journal of Entertainment & Technology Law 99 <https://scholarship.law.vanderbilt.edu/jetlaw/vol23/iss1/3/>

Lu Y and others, ‘How Information Flows from the World to China’ (2022) 29 The International Journal of Press/Politics

MA A, ‘Digital Legislation: Convergence or Divergence of Models? A Comparative Look at the European Union, China and the United States’ (2024) <https://server.www.robert-schuman.eu/storage/en/doc/questions-d-europe/qe-769-en.pdf> accessed 31 May 2025

Ma H and others, ‘Adapting Large Language Models for Content Moderation: Pitfalls in Data Engineering and Supervised Fine-Tuning’ (arXiv.org2023) <https://arxiv.org/abs/2310.03400> accessed 14 April 2025

Ma R and Kou Y, ‘“I’m Not Sure What Difference Is between Their Content and Mine, Other than the Person Itself”’ (2022) 6 Proceedings of the ACM on Human-Computer Interaction 1

Ma X, ‘Establishing an Indirect Liability System for Digital Copyright Infringement in China: Experience from the United States’ Approach - NYU Journal of Intellectual

Property & Entertainment Law' (*NYU Journal of Intellectual Property & Entertainment Law*4 May 2015) <https://jipel.law.nyu.edu/vol-4-no-2-3-ma/>

MacCarthy M, 'Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry' (*papers.ssrn.com*12 February 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3615726> accessed 4 August 2025

Madiega T, 'Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act | Think Tank | European Parliament' (*Europa.eu*2020) <https://www.europarl.europa.eu/thinktank/en/document/%20EPRS_IDA(2020)649404> accessed 16 April 2025

Mahon JE, 'The Definition of Lying and Deception' [2008] plato.stanford.edu <https://plato.stanford.edu/entries/lying-definition/?ref=ameasuredapproach.blog> accessed 5 December 2024

Maras M-H and Alexandrou A, 'Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos' (2019) 23 The International Journal of Evidence & Proof 255

Markesinis BS, 'Tort - Tort Law and Alternative Methods of Compensation', *Encyclopædia Britannica* (2019) <https://www.britannica.com/topic/tort/Tort-law-and-alternative-methods-of-compensation>

Markov T and others, 'A Holistic Approach to Undesired Content Detection in the Real World' (2023) 37 Proceedings of the ... AAAI Conference on Artificial Intelligence 15009

Marsden C, Meyer T and Brown I, 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?' (2019) 36 Computer Law & Security Review

Martin MS and Veatch S, 'AB 2013: New California AI Law Mandates Disclosure of GenAI Training Data | Perkins Coie' (*Perkinscoie.com*7 October 2024) <https://perkinscoie.com/insights/update/ab-2013-new-california-ai-law-mandates-disclosure-genai-training-data>

Martineau K, 'What Is Generative AI?' (*IBM Research Blog*20 April 2023) <https://research.ibm.com/blog/what-is-generative-AI> accessed 5 December 2024

Martin-Rodriguez F, Garcia-Mojon R and Fernandez-Barciela M, 'Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks' (2023) 23 Sensors 9037 <https://www.mdpi.com/1424-8220/23/22/9037>

Martins B and Morar D, 'Online Content Moderation Lessons from Outside the US' (*Brookings*17 June 2020) <https://www.brookings.edu/articles/online-content-moderation-lessons-from-outside-the-u-s/> accessed 5 December 2024

Marwick AE, 'Why Do People Share Fake News? A Sociotechnical Model of Media Effects' (*Georgetown Law Technology Review*21 July 2018) <https://georgetownlawtechreview.org/why-do-people-share-fake-news-a-sociotechnical-model-of-media-effects/GLTR-07-2018/> accessed 13 September 2025

Mast T, 'Platform Law as EU Law' (2024) 73 GRUR International 607 <https://doi.org/10.1093/grurint/ikae072> accessed 17 October 2025

Matias JN, Hounsel A and Feamster N, 'Software-Supported Audits of Decision-Making Systems: Testing Google and Facebook's Political Advertising Policies' (2022) 6 Proceedings of the ACM on Human-Computer Interaction 1

Matsakis L, 'Facebook's Targeted Ads Are More Complex than It Lets On' (*Wired*25 April 2018) <https://www.wired.com/story/facebooks-targeted-ads-are-more-complex-than-it-lets-on/> accessed 13 September 2025

Matthias A, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata' (2004) 6 Ethics and Information Technology 175

Matula C, 'Any Safe Harbor in a Storm: SESTA-FOSTA and the Future of § 230 of the Communications Decency Act' (2019) 18 Duke Law & Technology Review 353 <https://scholarship.law.duke.edu/dltr/vol18/iss1/24/> accessed 14 October 2025

McAfee A and Brynjolfsson E, 'Big Data: The Management Revolution' (*Harvard Business Review*October 2012) <https://hbr.org/2012/10/big-data-the-management-revolution> accessed 5 December 202

Mcafee A and Brynjolfsson E, *Machine, Platform, Crowd: Harnessing Our Digital Future* (W W Norton & Company 2017) <https://books.google.com.hk/books?hl=en&lr=&id=zh1DDQAAQBAJ&oi=fnd&pg=PA1905&dq=Machine> accessed 5 December 202

McCarthy J and others, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955' (1955) 27 AI Magazine 12 <https://www.aaai.org/ojs/index.php/aimagazine/article/view/1904>

McKinsey & Company, ‘The State of AI in 2023: Generative AI’s Breakout Year’ (*McKinsey & Company*1 August 2023) <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year> accessed 12 September 2025

McSherry C, ‘User Generated Content and the Fediverse: A Legal Primer’ (*Electronic Frontier Foundation*20 December 2022) <https://www.eff.org/deeplinks/2022/12/user-generated-content-and-fediverse-legal-primer> accessed 15 April 2025

Mctear M, *Conversational AI : Dialogue Systems, Conversational Agents, and Chatbots* (Morgan & Claypool Publishers 2020) 20

Mekacher A, Falkenberg M and Baronchelli A, ‘The Systemic Impact of Deplatforming on Social Media’ (2023) 2 PNAS Nexus

Meng L, ‘From Algorithm Bias to Algorithm Discrimination: Research on the Responsibility of Algorithmic Discrimination’ (2022) 24 Journal of Northeastern University(Social Science)

Meßmer A-K and Degeling M, ‘Auditing Recommender Systems -- Putting the DSA into Practice with a Risk-Scenario-Based Approach’ (*arXiv.org*2023) <https://arxiv.org/abs/2302.04556?.com> accessed 30 June 2025

Meta Newsroom, ‘Standing against Hate’ (*Meta Newsroom*27 March 2019) <https://about.fb.com/news/2019/03/standing-against-hate/> accessed 8 July 2025

Meta Transparency Center, ‘Integrity Reports, First Quarter 2025 | Transparency Center’ (*Meta.com*2025) <https://transparency.meta.com/zh-cn/integrity-reports-q1-2025/> accessed 4 September 2025

Meyer JG and others, ‘ChatGPT and Large Language Models in Academia: Opportunities and Challenges’ (2023) 16 BioData Mining

Meyer T and Pershan C, ‘Room for Improvement. Analysing Redress Policy on Facebook, Instagram, YouTube and Twitter - EU DisinfoLab’ (*EU DisinfoLab*2025) <https://www.disinfo.eu/publications/room-for-improvement-analysing-redress-policy-on-facebook-instagram-youtube-and-twitter> accessed 9 July 2025

Micova SB and Calef A, ‘Elements for Effective Systemic Risk Assessment under the DSA’ [2023] SSRN Electronic Journal

Miller C, ‘Can Congress Mandate Meaningful Transparency for Tech Platforms’ (*Stanford.edu*2 June 2021) <https://fsi.stanford.edu/news/meaningful-transparency-0> accessed 6 August 2025

Miller N, ‘Navigating the Web 2.0 Safe Harbour - Internet for Lawyers Newsletter’ (*Internet for Lawyers Newsletter*September 2008) <https://www.infolaw.co.uk/newsletter/2008/09/navigating-the-web-20-safe-harbour/> accessed 24 August 2025

Mirsky Y and Lee W, ‘The Creation and Detection of Deepfakes: A Survey’ (2021) 54 ACM Computing Surveys 1

Mitchell TM, *Machine Learning* (Mcgraw Hill 2020)

Mnasri M, ‘Recent Advances in Conversational NLP : Towards the Standardization of Chatbot Building’ (2019) <https://arxiv.org/pdf/1903.09025> accessed 14 April 2025

Montagnani ML and Trapova A, ‘New Obligations for Internet Intermediaries in the Digital Single Market—Safe Harbors in Turmoil?’ (2019) 22 Joural of Internet Law

Moore M, ‘Causation in the Law’ (*Stanford.edu*3 October 2019) <https://plato.stanford.edu/archIves/fall2024/entries/causation-law/#LawsExplDefiCaus> accessed 13 August 2025

Morita-Jaeger M and others, ‘Interoperability of Data Governance Regimes: Challenges for Digital Trade Policy | CITP’ (*Citp.ac.uk*2024) <https://citp.ac.uk/publications/interoperability-of-data-governance-regimes-challenges-for-digital-trade-policy> accessed 10 July 2025

Mosseri A, ‘Working to Stop Misinformation and False News - about Facebook’ (*About Facebook*6 April 2017) <https://about.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/> accessed 11 March 2025

Mulligan C, ‘Technological Intermediaries and Freedom of the Press’ (2013) 66 SSRN Electronic Journal

Mündges S and Park K, ‘But Did They Really? Platforms’ Compliance with the Code of Practice on Disinformation in Review’ (2024) 13 Internet Policy Review <https://policyreview.info/articles/analysis/platforms-compliance-code-of-practice-on-disinformation-review>

Murikah W, Nthenge JK and Musyoka FM, ‘Bias and Ethics of AI Systems Applied in Auditing - a Systematic Review’ (2024) 25 Scientific African <https://www.sciencedirect.com/science/article/pii/S2468227624002266>

Naitali A and others, ‘Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions’ (2023) 12 Computers 216 <https://www.mdpi.com/2530758>

Namirial Focus, ‘AI and Machine Learning: How Computers and AI Evolve Together’ (*Focus Namirial EN*27 September 2023) <https://focus.namirial.com/en/ai-machine-learning/#google_vignette> accessed 21 February 2025

Nanayakkara NWOKDSP, ‘Application of Artificial Intelligence in Marketing Mix: A Conceptual Review’ (*papers.ssrn.com*19 November 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3862936> accessed 4 August 2025

Nannini L and others, ‘Beyond Phase-In: Assessing Impacts on Disinformation of the EU Digital Services Act’ (2024) 5 AI and Ethics

Napoli PM, ‘Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers’ (2014) 39 SSRN Electronic Journal 756

Natale S, ‘If Software Is Narrative: Joseph Weizenbaum, Artificial Intelligence and the Biographies of ELIZA’ (2018) 21 New Media & Society 712 <https://doi.org/10.1177%2F1461444818804980> accessed 29 June 2020

Nathan J, ‘How Generative AI Is Becoming a Prime Target for Cyberattacks’ (*Medium*11 October 2024) <https://medium.com/@johnnathans/how-generative-ai-is-becoming-a-prime-target-for-cyberattacks-4b9fe760b1a0> accessed 5 December 202

National Academies of Sciences, Engineering, and Medicine, *Section 230 Protections: Can Legal Revisions or Novel Technologies Limit Online Misinformation and Abuse?* (National Academies Press 2021)

Neekhara P and others, ‘Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples’ [2020] arXiv:2002.12749 [cs] <https://arxiv.org/abs/2002.12749>

Newman N, Levy DAL and Nielsen RK, ‘Reuters Institute Digital News Report 2023’ [2023] SSRN Electronic Journal <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf> accessed 11 March 2025

News I, ‘TikTok Failed to Detect Disinformation on Adverts to Voting ahead of Irish General Election’ (*ITV News*29 November 2024) <https://www.itv.com/news/2024-11-29/tiktok-failed-to-detect-disinformation-relating-to-irish-general-election> accessed 8 July 2025

Niehoff DL and others, ‘New Product Liability Directive 2024/2853: New Product Liability Risks for Products in the EU’ (*Taylorwessing.com*6 January 2025)

<https://www.taylorwessing.com/en/insights-and-events/insights/2025/01/di-new-product-liability-directive> accessed 7 April 2025

Nielsen RK, 'How to Respond to Disinformation While Protecting Free Speech' (*Reuters Institute for the Study of Journalism*19 February 2021) <https://reutersinstitute.politics.ox.ac.uk/news/how-respond-disinformation-while-protecting-free-speech> accessed 11 March 2025

Nilsson N and Nielson D, 'SHAKEY the ROBOT' (1984) <https://www.cs.sfu.ca/~vaughan/teaching/415/papers/shakey.pdf> accessed 5 December 2024

Nilsson NJ, *Artificial Intelligence : A New Synthesis* (Kaufmann 2003) 4

Noble SU, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press 2018)

Nolan D, 'The Duty of Care after Robinson v Chief Constable of West Yorkshire Police' (*Ssrn.com*2 September 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3414361&> accessed 23 August 2025

Nunziato DC, 'The Digital Services Act and the Brussels Effect on Platform Content Moderation | Chicago Journal of International Law' (*cjil.uchicago.edu*2024) <https://cjil.uchicago.edu/print-archive/digital-services-act-and-brussels-effect-platform-content-moderation> accessed 6 June 2025

NW 1615 LS, Suite 800Washington and Inquiries D 20036USA202-419-4300 | M-8-8 | F-4-4 | M, 'Internet/Broadband Fact Sheet' (*Pew Research Center: Internet, Science & Tech*2024) <https://www.pewresearch.org/internet/fact-sheet/internet-broadband/#who-uses-the-internet> accessed 5 December 2024

O'Hara K, 'What Is Section 230 and Why Should I Care? - Internet Society' (*Internet Society*24 February 2023) <https://www.internetsociety.org/blog/2023/02/what-is-section-230-and-why-should-i-care-about-it/?gad_source=1&gad_campaignid=138051697&gbraid=0AAAADqyrA_4bQLTagzsVLldz6FPqNC8o&gclid=Cj0KCQjwzrzABhD8ARIsANlSWNPuD-ha2mduJQKarP5lEoWDCMu38Dw0stTXMsSsR_hrdriwEB7YHnsaAui_EALw_wcB> accessed 29 April 2025

O'Sullivan A, 'Section 230 Isn't an Aberration, It's a Distillation of Common Law Trends' (*Mercatus Center*16 July 2019) <https://www.mercatus.org/economic-insights/expert-commentary/section-230-isnt-aberration-its-distillation-common-law-trends> accessed 19 October 2025

Oliver A and others, ‘Realistic Evaluation of Deep Semi-Supervised Learning Algorithms’ (2018) 31 ArXiv (Cornell University) 3235

OpenAI, ‘GPT-4 Technical Report’ (OpenAI 2023) <https://cdn.openai.com/papers/gpt-4.pdf>

OpenAI, ‘Introducing ChatGPT’ (*Openai.com*30 November 2022) <https://openai.com/index/chatgpt/>

Ortutay B, ‘What You Should Know about Section 230, the Rule That Shaped Today’s Internet’ (*PBS NewsHour*21 February 2023) <https://www.pbs.org/newshour/politics/what-you-should-know-about-section-230-the-rule-that-shaped-todays-internet>

Oruç TH, ‘The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in Light of Recent Developments: Is It Still Necessary to Maintain It?’ (2022) 13 JIPITEC–Journal of Intellectual Property, Information Technology and E-Commerce Law 176 <https://www.jipitec.eu/jipitec/article/view/354> accessed 19 October 2025

Ouyang L and others, ‘Training Language Models to Follow Instructions with Human Feedback’ (2022) 35 Advances in Neural Information Processing Systems 27730 <https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html>

Palla K and others, ‘Policy-As-Prompt: Rethinking Content Moderation in the Age of Large Language Models’ (*arXiv.org*2025) <https://arxiv.org/abs/2502.18695> accessed 14 April 2025

Papadogiannakis E and others, ‘Before & After: The Effect of EU’s 2022 Code of Practice on Disinformation’ (*ACM Digital Library*22 April 2025) 1577 <https://dl.acm.org/doi/10.1145/3696410.3714898>

Papaevangelou C and Fabio V, ‘Trading Nuance for Scale? Platform Observability and Content Governance under the DSA’ (2025) 14 Internet Policy Review <https://policyreview.info/articles/analysis/platform-observability-and-content-governance>

Papazoglou A, ‘What Would John Stuart Mill Do—to Fix Facebook?’ (*The New Republic*28 January 2019) <https://newrepublic.com/article/152939/john-stuart-mill-doto-fix-facebook> accessed 7 July 202

Park K and Culloty E, ‘BEYOND PERFORMATIVE TRANSPARENCY: LESSONS LEARNED from the EU CODE of PRACTICE on DISINFORMATION’ [2023] Selected Papers of Internet Research

Park S and Nan X, ‘Generative AI and Misinformation: A Scoping Review of the Role of Generative AI in the Generation, Detection, Mitigation, and Impact of Misinformation’ [2025] AI & SOCIETY

Pasquale F, ‘The Black Box Society: The Secret Algorithms That Control Money and Information’ (2016) 45 Contemporary Sociology: a Journal of Reviews 367

Patel Y and others, ‘Deepfake Generation and Detection: Case Study and Challenges’ (2023) 11 IEEE Access 143296

Pech L, ‘THE CONCEPT of CHILLING EFFECT ITS UNTAPPED POTENTIAL to BETTER PROTECT DEMOCRACY, the RULE of LAW, and FUNDAMENTAL RIGHTS in the EU’ (2021) <https://www.opensocietyfoundations.org/uploads/c8c58ad3-fd6e-4b2d-99fa-d8864355b638/the-concept-of-chilling-effect-20210322.pdf> accessed 11 March 2025

Pedersen CS and H, ‘Platform Liability Trends around the Globe: Taxonomy and Tools of Intermediary Liability’ (*Electronic Frontier Foundation*25 May 2022) <https://www.eff.org/deeplinks/2022/05/platform-liability-trends-around-globe-taxonomy-and-tools-intermediary-liability>

Pengpai News, ‘Humans Start “Bullying” ChatGPT: Death Threats to Make Them Answer Banned Questions’ (*The Paper*2023) <https://baijiahao.baidu.com/s?id=1757148822155117011&wfr=spider&for=pc> accessed 11 March 2025

Pernot-Leplay E and Pernot-Leplay E, ‘AI Law in China, EU & U.S.: Comparative Analysis’ (*Pernot-Leplay*11 August 2024) <https://pernot-leplay.com/ai-regulation-china-eu-us-comparison/>

Perreault G, ‘Gatekeeping’ [2022] The SAGE Encyclopedia of Journalism

Pershan C, ‘Cutting through the Jargon - Independent Audits in the Digital Services Act’ (*Mozilla Foundation*30 January 2023) <https://www.mozillafoundation.org/en/blog/cutting-through-the-jargon-independent-audits-in-the-digital-services-act/> accessed 30 June 2025

Petersen N, *Proportionality and Judicial Activism* (Cambridge University Press 2017)

Peukert A, ‘The Regulation of Disinformation: A Critical Appraisal’ (2024) 16 Journal of Media Law 1

Pi Y, ‘Missing Value Chain in Generative AI Governance China as an Example’ (*arXiv.org*2024) <https://arxiv.org/abs/2401.02799> accessed 26 September 2025

Pielemeier JS, ‘Disentangling Disinformation: What Makes Regulating Disinformation so Difficult?’ (*Ssrn.com*17 January 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3629541> accessed 8 March 2025

Piotr Machnikowski, ‘The Principles of European Tort Law and Product Liability’ (2024) 15 Journal of European Tort Law 31

Plunkett JC, ‘The Historical Foundations of the Duty of Care’ (2015) 41 bridges.monash.edu <https://bridges.monash.edu/articles/journal_contribution/The_Historical_Foundations_of_the_Duty_of_Care/10065659>

Press TP, ‘Florida Social Media Platforms Bill - SB.7072 | TechPolicy.Press’ (*Tech Policy Press*) <https://www.techpolicy.press/tracker/florida-social-media-platforms-bill-sb-7072/>

Pykes K, ‘Variational Autoencoders: How They Work and Why They Matter’ (*Datacamp.com*13 August 2024) <https://www.datacamp.com/tutorial/variational-autoencoders> accessed 5 December 2024

Qian T, ‘The Knowledge Standard for the Internet Intermediary Liability in China’ (2011) 20 International Journal of Law and Information Technology 1

Quelle D and others, ‘Lost in Translation -- Multilingual Misinformation and Its Evolution’ (*arXiv.org*2023) <https://arxiv.org/abs/2310.18089> accessed 14 February 2025

Quintais J, ‘The New Copyright in the Digital Single Market Directive: A Critical Look’ (*papers.ssrn.com*14 October 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3424770> accessed 4 August 2025

Quintais JP, Appelman N and Fathaigh RÓ, ‘Using Terms and Conditions to Apply Fundamental Rights to Content Moderation’ (2023) 24 German Law Journal 1

Rafi MA, Feng Y and Jeon H, ‘Revealing Secrets from Pre-Trained Models’ (*arXiv.org*2022) <https://arxiv.org/abs/2207.09539> accessed 10 February 2025

Raji ID and others, ‘Closing the AI Accountability Gap: Defining an End-To-End Framework for Internal Algorithmic Auditing’ [2020] Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 33 <https://dl.acm.org/doi/abs/10.1145/3351095.3372873>

Rakhimov M, Javliev S and Nasimov R, ‘Parallel Approaches in Deep Learning: Use Parallel Computing’, *Proceedings of the 7th International Conference on Future Networks and Distributed Systems (ICFNDS 2023)* (Association for Computing Machinery (ACM) 2023)

Ramesh A and others, ‘Hierarchical Text-Conditional Image Generation with CLIP Latents’ (2022) <https://arxiv.org/pdf/2204.06125> accessed 14 April 2025

Rana V and Woods P, ‘How to Help AI Developers Understand the Societal Implications of Their Creations’ (*LSE Business Review*15 January 2024) <https://blogs.lse.ac.uk/businessreview/2024/01/15/how-to-help-ai-developers-understand-the-societal-implications-of-their-creations/> accessed 5 December 2024

Ranjith Kumar Gatla and others, ‘Advancements in Generative AI: Exploring Fundamentals and Evolution’ [2024] 2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC) <https://ieeexplore.ieee.org/document/10594003?denied=> accessed 18 August 2024

Ray PP, ‘ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope’ (2023) 3 Internet of Things and Cyber-Physical Systems 121 <https://www.sciencedirect.com/science/article/pii/S266734522300024X>

Regnier L, ‘AI Platform Launches in UK to Combat Harmful Content Online | Startups Magazine’ (*Startups Magazine*2025) <https://startupsmagazine.co.uk/index.php/article-ai-platform-launches-uk-combat-harmful-content-online> accessed 13 May 2025

Reichman JH, Dinwoodie GB and Samuelson P, ‘A Reverse Notice and Takedown Regime to Enable Public Interest Uses of Technically Protected Copyrighted Works’ (*Ssrn.com*2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1007817> accessed 7 August 2025

Reid A, Pendleton SM and Czabovsky LEHJ, ‘Social Media Transparency Reports: Longitudinal Content Analysis of News Coverage’ (2024) 13 SSRN Electronic Journal 122

Reid A, Ringel E and Pendleton SM, ‘Transparency Reports as CSR Reports: Motives, Stakeholders, and Strategies’ (2023) 20 Social Responsibility Journal 81

Richart JL, ‘A New Legal Framework for Online Platforms in the European Union (and Beyond)’ (2024) 59 Review of European and Comparative Law

Rizzoli A, ‘Training Data Quality: Why It Matters in Machine Learning’ (*www.v7labs.com*2022) <https://www.v7labs.com/blog/quality-training-data-for-machine-learning-guide> accessed 5 December 2024

RKB LAW SOLICITORS, ‘What Is the Difference between Public and Private Law? | RKB Law Kent’ (*RKB | Solicitors | Law* |13 June 2019) <https://rkb-law.co.uk/what-is-the-difference-between-public-and-private-law/>

Robb A, ‘Pizzagate: Anatomy of a Fake News Scandal’ (*Rolling Stone*16 November 2017) <https://www.rollingstone.com/feature/anatomy-of-a-fake-news-scandal-125877/> accessed 2 December 2024

Roberts H, Zuckerman E and Palfrey JG, ‘2011 Circumvention Tool Evaluation’ (SSRN Electronic Journal2011) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1940455>accessed 26 January 2025

Röder M, Both A and Hinneburg A, ‘Exploring the Space of Topic Coherence Measures’ [2015] Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM ’15 399 <http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf>

Romana F and Santiago F, ‘A Comparative Analysis of Artificial Intelligence Regulatory Law in Asia, Europe, and America’ (2024) 204 SHS Web of Conferences

Romero DM, Meeder B and Kleinberg J, ‘Differences in the Mechanics of Information Diffusion across Topics’ [2011] Proceedings of the 20th International Conference on World Wide Web - WWW ’11

Roth E, ‘The EU’s Digital Services Act Goes into Effect Today: Here’s What That Means’ (*The Verge*25 August 2023) <https://www.theverge.com/23845672/eu-digital-services-act-explained>

Roumeliotis KI and Tselikas ND, ‘ChatGPT and Open-AI Models: A Preliminary Review’ (2023) 15 Future Internet 192 <https://www.mdpi.com/1999-5903/15/6/192/htm>

Roynette K, ‘Drawing the Line of the Scope of the Duty of Care in American Negligence and French Fault-Based Tort Liability’ (2015) 8 Journal of Civil Law Studies <https://digitalcommons.law.lsu.edu/jcls/vol8/iss1/4/>

Rozenshtein A, ‘Interpreting the Ambiguities of Section 230’ (*Yale Journal on Regulation*17 April 2024) <https://www.yalejreg.com/bulletin/interpreting-the-ambiguities-of-section-230/> accessed 19 October 2024

Rozenshtein A, ‘The Real Takeaway from the Enjoining of the Florida Social Media Law’ (*Lawfare*2021) <https://www.lawfaremedia.org/article/real-takeaway-enjoining-florida-social-media-law> accessed 19 October 2024

Rustad M and Koenig T, ‘The Case for a CDA Section 230 Notice-And-Takedown Duty’ (2023) 23 Nevada Law Journal <https://scholars.law.unlv.edu/nlj/vol23/iss2/7/> accessed 4 August 2025

Sacks S, ‘Beijing Wants to Rewrite the Rules of the Internet’ (*The Atlantic*18 June 2018) <https://www.theatlantic.com/international/archive/2018/06/zte-huawei-china-trump-trade-cyber/563033/> accessed 10 July 2025

Saha B, Rani N and Shukla SK, ‘Generative AI in Financial Institution: A Global Survey of Opportunities, Threats, and Regulation’ (*arXiv.org*2025) <https://arxiv.org/abs/2504.21574> accessed 4 June 2025

Sakaguchi K and others, ‘WinoGrande: An Adversarial Winograd Schema Challenge at Scale’ (2020) 34 Proceedings of the AAAI Conference on Artificial Intelligence 8732

Salawu S, He Y and Lumsden J, ‘Approaches to Automated Detection of Cyberbullying: A Survey’ (2017) 11 IEEE Transactions on Affective Computing 1

Sanchez J, ‘Opinion | the Future of Fake News Is Being Pioneered in Homemade Porn’ (NBC News8 February 2018) <https://www.nbcnews.com/think/opinion/thanks-ai-future-fake-news-may-be-easily-faked-video-ncna845726> accessed 26 November 2024

Sandotra N and Arora B, ‘A Comprehensive Evaluation of Feature-Based AI Techniques for Deepfake Detection’ (2023) 36 Neural Computing and Applications 3860

Sarker IH, ‘Machine Learning: Algorithms, Real-World Applications and Research Directions’ (2021) 2 SN Computer Science 1 <https://link.springer.com/article/10.1007/s42979-021-00592-x>

Sartor G, ‘DIRECTORATE GENERAL for INTERNAL POLICIES POLICY DEPARTMENT A: ECONOMIC and SCIENTIFIC POLICY Providers Liability: From the ECommerce Directive to the Future IN-DEPTH ANALYSIS’ (2017) <https://www.europarl.europa.eu/RegData/etudes/IDAN/2017/614179/IPOL_IDA(2017)614179_EN.pdf>

Savin A, ‘The EU Digital Services Act: Towards a More Responsible Internet’ (*papers.ssrn.com*16 February 2021)

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3786792> accessed 4 August 2025

Schaffner B and others, '"Community Guidelines Make This the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies' [2024] ArXiv (Cornell University)

Schlagwein D and Willcocks LP, '"ChatGPT et Al.": The Ethics of Using (Generative) Artificial Intelligence in Research and Science' (2023) 38 Journal of Information Technology 232

Schmidhuber J, 'Deep Learning in Neural Networks: An Overview' (2015) 61 Neural Networks 85

Schmidt RM, 'Recurrent Neural Networks (RNNs): A Gentle Introduction and Overview' [2019] ArXiv (Cornell University)

Schmitt V and others, 'The Role of Explainability in Collaborative Human-AI Disinformation Detection' (ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)3 June 2024) <https://dl.acm.org/doi/10.1145/3630106.3659031> accessed 5 October 2025

Schuett J, 'Risk Management in the Artificial Intelligence Act' (2023) 15 European Journal of Risk Regulation 1 <https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/risk-management-in-the-artificial-intelligence-act/2E4D5707E65EFB3251A76E288BA74068>

Schulze H and others, 'Far-Right Conspiracy Groups on Fringe Platforms: A Longitudinal Analysis of Radicalization Dynamics on Telegram' (2022) 28 Convergence: the International Journal of Research into New Media Technologies

Sebastian G, 'Exploring Ethical Implications of ChatGPT and Other AI Chatbots and Regulation of Disinformation Propagation' [2023] SSRN Electronic Journal

Sedova K and others, 'AI and the Future of Disinformation Campaigns' (*Center for Security and Emerging Technology*December 2021) <https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/> accessed 24 January 2025

Segura-Bedmar I and Alonso-Bartolome S, 'Multimodal Fake News Detection' (2022) 13 Information 284

Sekwenz M-T and others, 'Doing Audits Right? The Role of Sampling and Legal Content Analysis in Systemic Risk Assessments and Independent Audits in the

Digital Services Act' (*arXiv.org*2025) <https://arxiv.org/abs/2505.03601> accessed 1 July 2025

Shaheen L, 'Section 230'S Immunity for Generative Artificial Intelligence' (*SSRN.com*15 December 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4886463> accessed 12 October 2025

Shamsi K, 'Dangerous Social Media Trends: Can Social Media Platforms Be Held Liable?' (*Brooklaw.edu*2025) <https://sports-entertainment.brooklaw.edu/media/dangerous-social-media-trends-can-social-media-platforms-be-held-liable/?utm_.com> accessed 6 May 2025

Shao A, 'Beyond Misinformation: A Conceptual Framework for Studying AI Hallucinations in (Science) Communication' (*arXiv.org*2025) <https://arxiv.org/abs/2504.13777> accessed 2 May 2025

Shao L and others, 'Artificial Intelligence Generated Content (AIGC) in Medicine: A Narrative Review' (2024) 21 Mathematical biosciences and engineering 1672

Shazeer N and others, 'OUTRAGEOUSLY LARGE NEURAL NETWORKS: THE SPARSELY-GATED MIXTURE-OF-EXPERTS LAYER' (2017) <https://arxiv.org/pdf/1701.06538> accessed 14 April 2025

Sheehan M, 'China's AI Regulations and How They Get Made' (*Carnegie Endowment for International Peace*2023) <https://carnegieendowment.org/research/2023/07/chinas-ai-regulations-and-how-they-get-made?lang=en> accessed 5 December 2024

Shen C and Haimson O, 'The Virtual Jail: Content Mod-Eration Challenges Faced by Chinese Queer Content Creators on Douyin' [2025] Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems 2 <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/196552/chi25-924.pdf>

Shen W and Liu Y, 'China's Normative Systems for Responsible AI: From Soft Law to Hard Law' (*Cambridge University Press*2024) 150 <https://www.cambridge.org/core/books/cambridge-handbook-of-responsible-artificial-intelligence/chinas-normative-systems-for-responsible-ai/25A6636116359C1282F5874434CF467C> accessed 13 September 2025

Shen X and others, '"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models' (*arXiv.org*7 August 2023) <https://arxiv.org/abs/2308.03825> accessed 14 April 2025

Shepardson D, ‘Lingo Telecom Agrees to $1 Million Fine over AI-Generated Biden Robocalls’ *Reuters* (21 August 2024) <https://www.reuters.com/technology/artificial-intelligence/lingo-telecom-agrees-1-million-fine-over-ai-generated-biden-robocalls-2024-08-21/> accessed 13 September 2025

Sherman J, ‘China’s War for Control of Global Internet Governance’ [2022] SSRN Electronic Journal

Shim Y and Jhaver S, ‘Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms’ (*Arxiv.org*2025) <https://arxiv.org/html/2409.08498v1#bib.bib98> accessed 6 June 2025

Shin D, ‘How Do People Judge the Credibility of Algorithmic Sources?’ (2021) 37 AI & SOCIETY

Shivani Tufchi, Yadav A and Ahmed T, ‘A Comprehensive Survey of Multimodal Fake News Detection Techniques: Advances, Challenges, and Opportunities’ (2023) 12 International Journal of Multimedia Information Retrieval

Shoaib M and others, ‘Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models’ (2023) <https://arxiv.org/pdf/2311.17394> accessed 2 January 2025

Silva R, ‘Exploring Feature Extraction with CNNs - TDS Archive - Medium’ (*Medium*25 November 2023) <https://medium.com/towards-data-science/exploring-feature-extraction-with-cnns-345125cefc9a> accessed 25 February 2025

Silverman D, Kaltenthaler K and Dagher M, ‘Seeing Is Disbelieving: The Depths and Limits of Factual Misinformation in War’ (2021) 65 International Studies Quarterly 798

Singhal M and others, ‘SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice’ [2022] arXiv:2206.14855 [cs] <https://arxiv.org/abs/2206.14855>

Smith G and Brake J, ‘South Korea Confronts a Deepfake Crisis’ (*East Asia Forum*18 November 2024) <https://eastasiaforum.org/2024/11/19/south-korea-confronts-a-deepfake-crisis/> accessed 13 December 2024

Smith SJ and Marja Elsinga, *International Encyclopedia of Housing and Home* (Elsevier 2012)

SmythOS, ‘Understanding the Limitations of Symbolic AI: Challenges and Future Directions’ (*SmythOS*13 November 2024) <https://smythos.com/artificial-intelligence/symbolic-ai/symbolic-ai-limitations/> accessed 5 December 2024

Solsman JE, ‘A Deepfake Bot Is Creating Nudes out of Regular Photos’ (CNET2020) <https://www.cnet.com/news/privacy/deepfake-bot-on-telegram-is-violating-women-by-forging-nudes-from-regular-pics/> accessed 15 January 2025

Somers M, ‘Deepfakes, Explained’ (*MIT Sloan*21 July 2020) <https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained> accessed 5 December 2024

Song X, Liu M and Chen J, ‘Comprehensive Impact Analysis of GPT-4: High-Quality Economic Development and National Security Prevention’ (2023) 38 Journal of Guangdong University of Finance & Economics 100

Soundappan N and Sayed P, ‘FCC Cracks down on AI-Powered Robocalls’ (*Harvard Journal of Law & Technology*23 February 2024) <https://jolt.law.harvard.edu/digest/fcc-cracks-down-on-ai-powered-robocalls> accessed 13 September 2025

Spitale G, Biller-Andorno N and Germani F, ‘AI Model GPT-3 (Dis)Informs Us Better than Humans’ (2023) 9 Science Advances

Stanford University, ‘The 2025 AI Index Report’ (2025) <https://hai.stanford.edu/ai-index/2025-ai-index-report>

Stanton K, ‘Professional Negligence: Duty of Care Methodology in the Twenty First Century’ (2006) 22 Tottel’s Journal of Professional Negligence 134 <https://research-information.bris.ac.uk/en/publications/professional-negligence-duty-of-care-methodology-in-the-twenty-fi> accessed 21 August 2025

Stapleton J, ‘Duty of Care Factors: A Selection from the Judicial Menus’, *The Law of Obligations: Essays in Celebration of John Fleming* (Oxford University Press 1998) <https://academic.oup.com/book/50576/chapter-abstract/420948118?redirectedFrom=fulltext> accessed 18 September 2025

Stark D and Pais I, ‘Algorithmic Management in the Platform Economy’ (2020) 14 Sociologica 47 <https://sociologica.unibo.it/article/view/12221>

State Internet Information Office, ‘National Informatisation Development Report(2023)’ (2023) <https://www.gov.cn/lianbo/bumen/202409/content_6973030.htm> accessed 10 July 2025

Stawowy A, ‘User Content Moderation under the Digital Services Act – 10 Key Takeaways – Law Firm Traple Konarski Podrecki and Partners’ (*Law Firm Traple Konarski Podrecki and Partners*24 October 2023) <https://www.traple.pl/en/user-

content-moderation-under-the-digital-services-act-10-key-takeaways/> accessed 18 September 2025

Steinfeld CW and Fulk J, *Organizations and Communication Technology* (SAGE Publishing 1990)

Stengel R, 'Revoking the Law That Protects Twitter Could Backfire on Trump' (*Vanity Fair*June 2020) <https://www.vanityfair.com/news/2020/06/revoking-the-law-that-protects-twitter-could-backfire-on-trump?srsltid=AfmBOoq1vOrVHagyPDsW-qT6mz0pNCK6XWApw6N4cAWEz5hhN5CwKuSS> accessed 28 May 2025

Stepney E and Lally C, 'Disinformation: Sources, Spread and Impact Overview' (2024) <https://researchbriefings.files.parliament.uk/documents/POST-PN-0719/POST-PN-0719.pdf> accessed 2 December 2024

Stern MJ, 'The 5th Circuit's Reinstatement of Texas' Internet Censorship Law Could Break Social Media' (*Slate Magazine*12 May 2022) <https://slate.com/technology/2022/05/texas-internet-censorship-social-media-first-amendment-fifth-circuit.html>

Striepe M and Cunningham C, 'Gatekeepers, Guides and Ghosts: Intermediaries Impacting Access to Schools during COVID-19' (2022) 17 Ethnography and Education 1

Stroud M, 'These Six Lawsuits Shaped the Internet' (*The Verge*19 August 2014) <https://www.theverge.com/2014/8/19/6044679/the-six-lawsuits-that-shaped-the-internet>

Stryker C, 'What Is Multimodal AI? | IBM' (*Ibm.com*15 July 2024) <https://www.ibm.com/think/topics/multimodal-ai> accessed 5 December 2024

StudyBounty, 'Doe v. MySpace: The Case That Changed Social Media' (*StudyBounty*22 July 2022) <https://studybounty.com/doe-v-myspace-the-case-that-changed-social-media-research-paper> accessed 10 April 2025

Su C, Li Z and Qiao Q, 'Internet Platform Governance: A Comparison of PRC Law and EU Law - KWM' (Kwm.com2022) <https://www.kwm.com/cn/en/insights/latest-thinking/internet-platform-governance-a-comparison-of-prc-law-and-eu-law.html> accessed 20 September 2025

Sulimov D, 'Prompt-Efficient Fine-Tuning for GPT-like Deep Models to Reduce Hallucination and to Improve Reproducibility in Scientific Text Generation Using Stochastic Optimisation Techniques' (*arXiv.org*2024) <https://arxiv.org/abs/2411.06445> accessed 31 January 2025

Sullivan D and Pielemeier J, ‘Unpacking “Systemic Risk” under the EU’s Digital Service Act’ (*Tech Policy Press*19 July 2023) <https://www.techpolicy.press/unpacking-systemic-risk-under-the-eus-digital-service-act/> accessed 26 June 2025

Sullivan D, ‘Systemic Risk Assessments Hold Clues for EU Platform Enforcement’ (*Lawfare*2025) <https://www.lawfaremedia.org/article/systemic-risk-assessments-hold-clues-for-eu-platform-enforcement> accessed 15 July 2025

Sun J and Yusufu P, ‘ChatGPT’s Risk Overlay and Legal Response to Data Compliance’ (2023) 7 Law and Modernization

Suzor N and Gillett R, ‘Self-Regulation and Discretion’ [2022] Palgrave Global Media Policy and Business 259

Suzor N, ‘Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms’ (2018) 4 Social Media + Society

Suzor NP and others, ‘What Do We Mean When We Talk about Transparency? Toward Meaningful Transparency in Commercial Content Moderation’ (2019) 13 International Journal of Communication 18 <https://ijoc.org/index.php/ijoc/article/view/9736>

Suzor NP, *Lawless: The Secret Rules That Govern Our Digital Lives* (Cambridge University Press 2019) <https://www.cambridge.org/core/books/lawless/8504E4EC8A74E539D701A04D3EE8D8DE>

Swatton P and Leblanc M, ‘What Are Deepfakes and How Can We Detect Them?’ (*The Alan Turing Institute*2023) <https://www.turing.ac.uk/blog/what-are-deepfakes-and-how-can-we-detect-them> accessed 2 November 2024

Swire B, Ecker UKH and Lewandowsky S, ‘The Role of Familiarity in Correcting Inaccurate Information.’ (2017) 43 Journal of Experimental Psychology: Learning, Memory, and Cognition 1948

Szczepaniak A, ‘Leo Rover Blog - What Was the World’s First Mobile Intelligent Robot?’ (*www.leorover.tech*2023) <https://www.leorover.tech/post/what-was-the-worlds-first-mobile-intelligent-robot> accessed 5 December 2024

Tanner B, ‘EU Code of Practice on Disinformation’ (*Brookings*5 August 2022) <https://www.brookings.edu/articles/eu-code-of-practice-on-disinformation/>

Tao Q, ‘Legal Framework of Online Intermediaries’ Liability in China’ (2012) 14 Info 59

Thomas Wesley Allen, ‘Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within the R&D Organization’ (1984) 1 RePEc: Research Papers in Economics

Tiernan M and Sluiter G, ‘The European Union’s Digital Services Act and Secondary Criminal Liability for Online Platform Providers: A Missed Opportunity for Fair Criminal Accountability?’ (*SSRN Electronic Journal*2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4731220> accessed 5 December 2024

Tolson B, ‘Generative AI and Data Privacy: The Challenge of PII Use in Training Data Sets’ (*Smarsh*11 June 2024) <https://www.smarsh.com/blog/thought-leadership/generative-AI-and-data-privacy-the-challenge-of-PII-use-in-training-data-sets> accessed 5 December 2024

Toptchiyska D, ‘Legal Aspects of Content Moderation on Social Media Platforms: A Comparative Perspective’ (*Ssrn.com*22 May 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4901501> accessed 4 August 2025

Trandabăț D and Gifu D, ‘Discriminating AI-Generated Fake News’ (2023) 225 Procedia Computer Science 3822 <https://www.sciencedirect.com/science/article/pii/S1877050923015363>

Trujillo A, Fagni T and Cresci S, ‘The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media’ (*arXiv.org*2023) <https://arxiv.org/abs/2312.10269> accessed 14 April 2025

Tsamados A and others, ‘The Ethics of Algorithms: Key Problems and Solutions’ (2021) 37 AI & Society 215 <https://link.springer.com/article/10.1007/s00146-021-01154-8>

Tsertekidis G and Polyzoidis P, ‘Leveraging Artificial Intelligence in the Field of Social Policy against Social Inequalities: The Current Landscape’ (2024) 3 Journal of Politics and Ethics in New Technologies and AI <https://ejournals.epublishing.ekt.gr/index.php/jpentai/article/view/38831> accessed 19 September 2025

Turilli M and Floridi L, ‘The Ethics of Information Transparency’ (2009) 11 Ethics and Information Technology 105 <https://link.springer.com/article/10.1007/s10676-009-9187-9>

Turilli M, ‘Ethical Protocols Design’ (2007) 9 Ethics and Information Technology 49

Turing AM, ‘Computing Machinery and Intelligence’, *Parsing the Turing Test* (Mind 2007) <https://www.csee.umbc.edu/courses/471/papers/turing.pdf> accessed 5 December 2024

Tushnet R, ‘A Hobgoblin Comes for Internet Regulation’ (*VerfBlog (short for Verfassungsblog)*19 February 2024) <https://verfassungsblog.de/a-hobgoblin-comes-for-internet-regulation/> accessed 28 June 2025

Tusikov N, *Chokepoints : Global Private Regulation on the Internet* (Oakland, California University of California Press 2017)

Twitter, ‘Evolving Our Twitter Transparency Report: Expanded Data and Insights’ (*X.com*2018) <https://blog.x.com/en_us/topics/company/2018/evolving-our-twitter-transparency-report> accessed 4 September 2025

Tyagi S and Yadav D, ‘A Detailed Analysis of Image and Video Forgery Detection Techniques’ (2022) 39 The Visual Computer

UK P, ‘EU Study on the Legal Analysis of a Single Market for the Information Society - Publications Office of the EU’ (*Publications Office of the EU*2023) <https://op.europa.eu/en/publication-detail/-/publication/a856513e-ddd9-45e2-b3f1-6c9a0ea6c722> accessed 16 April 2025

Ullrich C, ‘Standards for Duty of Care? Debating Intermediary Liability from a Sectoral Perspective’ (2017) 8 Journal of Intellectual Property, Information Technology and E-Commerce Law 111 <https://www.jipitec.eu/jipitec/article/view/197> accessed 18 September 2025

Urban JM and Quilter L, ‘Efficient Process or “Chilling Effects”? Takedown Notices under Section 512 of the Digital Millennium Copyright Act’ (*Ssrn.com*23 May 2006) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2210935> accessed 4 August 2025

Urman A and Makhortykh M, ‘How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms’ (2023) 47 Telecommunications Policy 102477

Vaccari C and Chadwick A, ‘Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News’ (2020) 6 Social Media + Society <https://journals.sagepub.com/doi/10.1177/2056305120903408>

Valcke P, Aleksandra Kuczerawy and Pieter-Jan Ombelet, ‘Did the Romans Get It Right? What Delfi, Google, EBay, and UPC TeleKabel Wien Have in Common’ (2017) 31 Law, Governance and Technology Series 101

Valgaeren E and Fischer C, ‘Online Platforms and Uploading of Protected Works: A Priori No Liability for Operators of Online Platforms’ (*Stibbe*15 July 2021) <https://www.stibbe.com/publications-and-insights/online-platforms-and-uploading-of-protected-works-a-priori-no-liability> accessed 13 May 2025

Van Dijck J and Poell T, ‘Understanding Social Media Logic’ (2013) 1 Media and Communication 2

van J, *Platform Mechanisms* (Oxford University Press 2018) 31

van Oijen V, ‘AI-Generated Text Detectors: Do They Work? | SURF Communities’ (*communities.surf.nl*31 March 2023) <https://communities.surf.nl/en/ai-in-education/article/ai-generated-text-detectors-do-they-work> accessed 10 July 2025

Vasileios Mezaris and others, *Video Verification in the Fake News Era* (Springer International Publishing 2019)

Vaswani A and others, ‘Attention Is All You Need’ (*arXiv*12 June 2017) <https://arxiv.org/abs/1706.03762> accessed 14 April 2025

Veale M and Borgesius FZ, ‘Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach’ (2021) 22 Computer Law Review International 97

Velásquez N and others, ‘Online Hate Network Spreads Malicious COVID-19 Content Outside the Control of Individual Social Media Platforms’ (2021) 11 Scientific Reports 11549 <https://www.nature.com/articles/s41598-021-89467-y>

Vincent J, ‘Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women’ (*The Verge*10 October 2018) <https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report> accessed 18 January 2025

Vincent J, ‘Deepfake Detection Algorithms Will Never Be Enough’ (*The Verge*27 June 2019) <https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work>

Vinhas O and Bastos M, ‘The WEIRD Governance of Fact-Checking and the Politics of Content Moderation’ (2023) 27 New Media & Society

von Eschenbach WJ, ‘Transparency and the Black Box Problem: Why We Do Not Trust AI’ (2021) 34 Philosophy & Technology 1607 <https://link.springer.com/article/10.1007/s13347-021-00477-0>

Vosoughi S, Roy D and Aral S, ‘The Spread of True and False News Online’ (2018) 359 Science 1146

Vranckaert K, ‘Disinformation as a Cyber Threat under EU Law: Which Approach to Take in the Age of AI?’ (*Faculteit Rechtsgeleerdheid En Criminologische Wetenschappen*2024) <https://www.law.kuleuven.be/ai-summer-school/blogpost/Blogposts/disinformation-as-a-cyber-threat-under-eu-law-which-approach-to-take-in-the-age-of-ai> accessed 21 July 2025

Vreese C de, ‘A Wave of Generative AI Disinformation? – EDMO’ (*Edmo.eu*2023) <https://edmo.eu/blog/a-wave-of-generative-ai-disinformation/?utm_source=chatgpt.com> accessed 8 January 2025

Vykopal I and others, ‘Disinformation Capabilities of Large Language Models’ (*arXiv.org*23 February 2024) <https://arxiv.org/abs/2311.08838> accessed 22 December 2024

Wagner G, ‘Liability Rules for the Digital Age’ (2022) 13 Journal of European Tort Law 191

Wandt AS, ‘Tort: Property’ in Lauren R Shapiro and Marie-Helen Maras (eds), *Encyclopedia of Security and Emergency Management* (Springer International Publishing 2020)

Wang J, ‘Platform Responsibility with Chinese Characteristics’ in Chakravorti B and Trachtman JP (eds), Defeating Disinformation: Digital Platform Responsibility, Regulation and Content Moderation on the Global Technological Commons (Cambridge University Press 2025)

Wang L, ‘Development and Prospect of False Information Detection on Social Media’ (2022) 53 Journal of Taiyuan University of Technology

Wang S and others, ‘Artificial Intelligence Policy Frameworks in China, the European Union and the United States: An Analysis Based on Structure Topic Model’ (2025) 212 Technological Forecasting and Social Change 123971

Wang X and Cao S, ‘Harnessing the Stream: Algorithmic Imaginary and Coping Strategies for Live-Streaming E-Commerce Entrepreneurs on Douyin’ (2024) 11 The Journal of Chinese Sociology

Wang Z and others, ‘CIEASR:Contextual Image-Enhanced Automatic Speech Recognition for Improved Homophone Discrimination’ (2024) 1 Proceedings of the 31st ACM International Conference on Multimedia 915

Wardle C and Derakhshan H, ‘Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making’ (*Council of Europe*2017) <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-

interdisciplinary-framework-for-research-and-policy-making.html> accessed 5 December 2024

Wayte L, 'Contributory and Vicarious Liability for Peer-To-Peer File Sharing Services: The Napster and Grokster Cases', *Pay for Play: How the Music Industry Works, Where the Money Goes, and Why* (Pressbooks 2023) <https://opentext.uoregon.edu/payforplay/chapter/chapter-38-contributory-and-vicarious-liability-for-peer-to-peer-file-sharing-services-the-napster-and-grokster-cases/> accessed 14 October 2025

Webber GCN, 'Proportionality, Balancing, and the Cult of Constitutional Rights Scholarship' (2010) 23 Canadian Journal of Law & Jurisprudence 179

Weber P, 'SmythOS - Symbolic AI in Natural Language Processing: A Comprehensive Guide' (*SmythOS*15 November 2024) <https://smythos.com/ai-agents/natural-language-processing/symbolic-ai-in-natural-language-processing/> accessed 9 February 2025

Weber-Wulff D and others, 'Testing of Detection Tools for AI-Generated Text' (2023) 19 ArXiv (Cornell University)

Wei Y-H, 'VAEs and GANs: Implicitly Approximating Complex Distributions with Simple Base Distributions and Deep Neural Networks -- Principles, Necessity, and Limitations' (*arXiv.org*2025) <https://arxiv.org/abs/2503.01898> accessed 28 September 2025

Weibo Corporation , 'Weibo Reports Second Quarter 2021 Unaudited Financial Results | Weibo Corporation' (*Weibo Corporation*2021) <http://ir.weibo.com/news-releases/news-release-details/weibo-reports-second-quarter-2021-unaudited-financial-results> accessed 3 September 2025

Weidinger L and others, 'Taxonomy of Risks Posed by Language Models' [2022] 2022 ACM Conference on Fairness, Accountability, and Transparency 214

Weizenbaum J, 'ELIZA - a Computer Program for the Study of Natural Language Communication between Man and Machine' (1966) 9 Communications of the ACM 36

Werner J, 'New Mexico Enacts Law Requiring Disclosure of AI-Generated Content in Political Campaign Ads' (*BABL AI*27 August 2024) <https://babl.ai/new-mexico-enacts-law-requiring-disclosure-of-ai-generated-content-in-political-campaign-ads/?utm_.com> accessed 30 April 2025

Westerlund M, 'The Emergence of Deepfake Technology: A Review' (2019) 9 Technology Innovation Management Review 39 <https://timreview.ca/article/1282>

Wheeler T, ‘The Supreme Court Takes up Section 230’ (*Brookings*31 January 2023) <https://www.brookings.edu/articles/the-supreme-court-takes-up-section-230/>

White DM, ‘The “Gate Keeper”: A Case Study in the Selection of News’ (1950) 27 Journalism Quarterly 383

Whitehouse T, ‘How AI Governance Can Adapt to a Fragmented Regulatory Landscape’ (*WSJ*2 December 2024) <https://deloitte.wsj.com/cfo/how-ai-governance-can-adapt-to-a-fragmented-regulatory-landscape-23e47f94>

Whittle J, ‘AI Can Now Learn to Manipulate Human Behaviour’ (*The Conversation*11 February 2021) <https://theconversation.com/ai-can-now-learn-to-manipulate-human-behaviour-155031> accessed 18 January 2025

Wiener A, ‘Trump, Twitter, Facebook, and the Future of Online Speech’ (*The New Yorker*6 July 2020) <https://www.newyorker.com/news/letter-from-silicon-valley/trump-twitter-facebook-and-the-future-of-online-speech> accessed 28 May 2025

Wilson M, ‘Study: Facebook’s Fake News Labels Have a Fatal Flaw’ (*Fast Company*4 March 2020) <https://www.fastcompany.com/90471349/study-facebooks-fake-news-labels-have-a-fatal-flaw> accessed 22 April 2025

Windwehr S, ‘Systemic Risk Reporting: A System in Crisis?’ (*Electronic Frontier Foundation*16 January 2025) <https://www.eff.org/deeplinks/2025/01/systemic-risk-reporting-system-crisis> accessed 15 July 2025

Wingfield R, ‘A Human Rights-Based Approach to Disinformation | Global Partners Digital’ (*Global Partners Digital*15 October 2019) <https://www.gp-digital.org/a-human-rights-based-approach-to-disinformation/>

Wodecki B, ‘Generative AI Funding Hits $25.2 Billion in 2023, Report Reveals’ (*AI Business*2023) <https://aibusiness.com/verticals/generative-ai-funding-hits-25-2-billion-in-2023-report-reveals> accessed 5 October 2025

Wolff J, ‘Policy Approaches to Defining and Enforcing Responsibilities for Online Platforms’, *Defeating Disinformation*(Cambridge University Press 2025) <https://www.cambridge.org/core/books/defeating-disinformation/policy-approaches-to-defining-and-enforcing-responsibilities-for-online-platforms/EFF7B8FAC2D22BD36CA860B97755679E> accessed 4 September 2025

Wolters P and Borgesius FZ, ‘The EU Digital Services Act: What Does It Mean for Online Advertising and Adtech?’ (2025) 33 International Journal of Law and Information Technology

Woods L and Perrin W, ‘Obliging Platforms to Accept a Duty of Care’, *Regulating Big Tech: Policy Responses to Digital Dominance* (Oxford University Press 2021) <https://academic.oup.com/book/39213/chapter/338717347>

Xiao B, ‘Making the Private Public: Regulating Content Moderation under Chinese Law - Repository of the Academy’s Library’ (2023) 51 Computer Law & Security Review <https://real.mtak.hu/180470/> accessed 17 October 2025

Xiao F, ‘Improvement of E-Commerce Platform Responsibility Legislation for Consumer Protection from the Perspective of Informational Interests’ (2022) 24 Journal of Shanghai University of Finance and Economics

Xiaoming H, Zhang K and Yu H, ‘The Internet and Information Control: The Case of China’ (1996) 3 Javnost - the Public 117

Xu J, ‘Opening the “Black Box” of Algorithms: Regulation of Algorithms in China’ (2024) 10 Communication Research and Practice 288

Xue J and others, ‘Detecting Fake News by Exploring the Consistency of Multimodal Data’ (2021) 58 Information Processing & Management 102610

Yang C and Luo X, ‘A Preliminary Study on the Comprehensive Management of Algorithmic Discrimination’ (2018) 8 Science and Society

Yang F and Yao Y, ‘A New Regulatory Framework for Algorithm-Powered Recommendation Services in China’ (2022) 4 Nature Machine Intelligence 802 <https://www.nature.com/articles/s42256-022-00546-9>

Yang K-C and Menczer F, ‘Anatomy of an AI-Powered Malicious Social Botnet’ (*arXiv.org*30 July 2023) <https://arxiv.org/abs/2307.16336> accessed 14 April 2025

Yang Y, ‘Attribution of Liability for Copyright Infringement by Artificial Intelligence Generated Content’ (2023) 29 Lecture Notes in Education Psychology and Public Media 115

Yankoski M, Scheirer W and Weninger T, ‘Meme Warfare: AI Countermeasures to Disinformation Should Focus on Popular, Not Perfect, Fakes’ (2021) 77 Bulletin of the Atomic Scientists 119

Yeung K, ‘A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework’ (*papers.ssrn.com*9 November 2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3286027> accessed 4 August 2025

Ying Cheng Wu and Wang X, ‘Balancing Innovation and Regulation in the Age of Generative Artificial Intelligence’ (2024) 14 Journal of Information Policy

Yli-Huumo J and others, ‘Where Is Current Research on Blockchain Technology?—a Systematic Review’ (2016) 11 PLOS ONE <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163477>

York JK and JC, ‘Seven Times Journalists Were Censored: 2017 in Review’ (*Electronic Frontier Foundation*30 December 2017) <https://www.eff.org/deeplinks/2017/12/seven-times-2017-journalists-were-censored> accessed 11 March 2025

You C, ‘Law and Policy of Platform Economy in China’ (2020) 39 Computer Law & Security Review 105493

Young A, ‘History of Photo Editing and Photo Manipulation’ (FixThePhoto.com2019) <https://fixthephoto.com/blog/retouch-tips/history-of-photo-retouching.html> accessed 20 December 2024

Yu G and others, ‘Tamperproof IoT with Blockchain’ (*arXiv.org*2022) <https://arxiv.org/abs/2208.05109> accessed 10 March 2025

Yu G and others, ‘The Manifestation, Governance and Effect Evaluation of False Information on the Internet’ (2024) 02 Youth Journalist

Zanotto SE and Aroyehun S, ‘Human Variability vs. Machine Consistency: A Linguistic Analysis of Texts Generated by Humans and Large Language Models’ (*arXiv.org*2024) <https://arxiv.org/abs/2412.03025> accessed 14 February 2025

Zellers R and others, ‘Defending against Neural Fake News’ (*arXiv.org*29 May 2019) <https://arxiv.org/abs/1905.12616> accessed 14 April 2025

Zeng M and Kim Y, ‘Institutional Reforms and Regulatory Shifts in China’s Digital Platform Sector: How Domain-Specific Centralization Shaped the 2020–2022 Transition’ [2025] Business and Politics 1

Zhang AH, ‘Agility over Stability: China’s Great Reversal in Regulating the Platform Economy’ (*papers.ssrn.com*28 July 2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3892642> accessed 4 August 2025

zhang B, ‘Limitation of Safety-Guard Responsibility of Platform Operator’ (2019) 22 Economic Law Review

Zhang H and others, ‘R-Tuning: Instructing Large Language Models to Say `I Don’t Know’’ (*arXiv.org*2023) <https://arxiv.org/abs/2311.09677> accessed 14 February 2025

Zhang J, ‘A Comparative Study of Fake News Governance Measures in China and Abroad’ (*Fx361.com*2020) <https://m.fx361.com/news/2020/0718/6879933.html> accessed 12 July 2025

Zhang S, ‘Research on the Security Guarantee Obligation of E-Commerce Platform Operators’ (2025) 13 E-commerce Reviews <https://www.hanspub.org/journal/paperinformation?paperid=100698&> accessed 27 August 2025

Zhang T, ‘Deepfake Generation and Detection, a Survey’ (2022) 81 Multimedia Tools and Applications 6259

Zhang Y and Chen C, ‘Which Province or City Responded More Decisively and Effectively to the Epidemic? We Did Some Data Analysis’ (*Baidu.com*2020) <https://baijiahao.baidu.com/s?id=1662292560776616747&wfr=spider&for=pc> accessed 14 July 2025

Zhang Y and others, ‘Shifting Trust: Examining How Trust and Distrust Emerge, Transform, and Collapse in COVID-19 Information Seeking’ [2022] CHI Conference on Human Factors in Computing Systems

Zhou J and others, ‘Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions’ (2023) 1 Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems

Zhou X, ‘E-Commerce Platforms’ Security Obligations and Legal Responsibilities’ (*Southcn.com*2022) <https://theory.southcn.com/node_203ed94b00/5fbbe625d3.shtml> accessed 27 August 2025

Zhuk A, ‘Beyond the Blockchain Hype: Addressing Legal and Regulatory Challenges’ (2025) 5 SN Social Sciences

Ziedler S, ‘Navigating Platform Power: From European Elections to the Regulatory Future’ (*HIIG*18 July 2024) <https://www.hiig.de/en/dsa-navigating-platform-power/> accessed 18 August 2025

Zilberman E, ‘Platform Liability Regimes around the World | Heinrich-Böll-Stiftung | Tel Aviv - Israel’ (*Heinrich-Böll-Stiftung | Tel Aviv - Israel*2022) <https://il.boell.org/en/2023/03/30/platform-liability-regimes-around-world> accessed 31 May 2025

Zimmermann R, ‘The Civil Law in European Codes’, *Regional Private Laws and Codification in Europe* (Cambridge University Press 2003) <https://www.cambridge.org/core/books/abs/regional-private-laws-and-codification-in-europe/civil-law-in-european-codes/5D6AAF67CB7A86C1380FB853DEF9803B> accessed 7 October 2025

Zingales L, ‘Preventing Economists’ Capture’ [2013] Preventing Regulatory Capture 124

Zittrain J, ‘A History of Online Gatekeeping’ (2006) 19 Harvard Journal of Law & Technology 253

Zittrain JL, ‘Three Eras of Digital Governance’ (*papers.ssrn.com*23 September 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458435> accessed 4 August 2025

Zmud RW, ‘Opportunities for Strategic Information Manipulation through New Information Technology’, Organizations and Communication Technology (SAGE Publications 1990)

Ződi Z, ‘Characteristics of the European Platform Regulation’ (2022) 7 Public Governance, Administration and Finances Law Review 91

Zou M and Zhang L, ‘Navigating China’s Regulatory Approach to Generative Artificial Intelligence and Large Language Models’ (2025) 15 European Journal of Risk Regulation

Zyskind G, Nathan O and Pentland A ’Sandy’, ‘Decentralizing Privacy: Using Blockchain to Protect Personal Data’ [2015] 2015 IEEE Security and Privacy Workshops <https://ieeexplore.ieee.org/abstract/document/7163223>