Rethinking Section 230: The Black Box Problem in Generative AI

Veronica Arias

LGST 2220

The internet's landscape has been reshaped by the arrival of Generative Artificial Intelligence (GAI), a new class of innovative technology that operates with unprecedented autonomy. These complex algorithms can craft everything from novels to convincing news articles, blurring the lines between humans and machines in ways that were unimaginable just a few years ago. While this unprecedented potential promises to revolutionize numerous industries, it also casts a long shadow of concern: how do we address the challenges of misinformation and bias within this rapidly evolving terrain?

This paper argues for a nuanced approach that prioritizes both innovation and user protection. While concerns about AI-generated content are valid, applying Section 230's publisher immunity indiscriminately to GAI outputs would stifle the very innovation that fuels this nascent field. Unlike traditional publishers or editors who exercise direct control over content, GAI developers operate through a "black box" phenomenon. They design and train these complex systems, but they do not directly compose the content itself, and therefore can not be said to be the speakers of said content. Applying Section 230 protections to GAI developers, therefore, safeguards this vital engine of technological progress, preserving the economic and social benefits it promises.

Section 230

The burgeoning online world of the 1990s, with platforms like Prodigy and CompuServe, presented a double-edged sword. Exciting possibilities of user-generated content were coupled with concerns about potentially harmful materials like obscenity and indecency[1]. Two landmark court cases in the decade highlighted the inadequacy of existing laws in addressing online content liability, exposing the need for clear rules to govern platform responsibility.

In 1991, the Cubby v. CompuServe case[2] highlighted the need for content moderation in the untamed landscape of the internet. CompuServe, a popular online service, faced a lawsuit from a news source called

[1] Mehta, M. D. (2002). Censoring Cyberspace. *Asian Journal of Social Science*, *30*(2), 319–338. http://www.jstor.org/stable/23654709
[2] Cubby, Incv. CompuServe, Inc., 776 F. Supp. 135 (S.D.N.Y. 1991)

"Skuttlebut", which claimed that another user's newsletter "Rumorville", which was posted on CompuServe's platform, was posting defamatory statements on CompuServe's platform. The court absolved CompuServe, reasoning that they were distributors, not publishers, lacking knowledge and control over user-generated content. This precedent established online platforms as distributors with limited liability, only responsible for content they knowingly control.

Four years later, the Stratton Oakmont v. Prodigy case[3] further complicated the landscape. Prodigy was an early online platform with bulletin boards for user-generated content. An anonymous user on Prodigy posted defamatory claims about Stratton Oakmont and its president. In 1995, Stratton Oakmont sued Prodigy for defamation. The court, differentiating Prodigy from CompuServe due to its active content moderation, found it liable as a "publisher." This conflicted with the Cubby precedent and ignited controversy, prompting calls for clearer legal frameworks.

Following the lingering uncertainty regarding the liability of internet service providers with respect to user-generated content, as well as concerns about online obscenity and child safety, the Communications Decency Act (CDA) was proposed by Congress in 1995.[4] The CDA aimed to "modernize the existing protections against obscene, lewd, indecent or harassing uses of a telephone."[5] The CDA's controversial indecency provisions were struck down due to First Amendment concerns, but Section 230, an amendment championed by Representatives Cox and Wyden, survived. This landmark provision established broad immunity for online platforms regarding user-generated content, protecting them from being treated as publishers and shielding them from liability for most user-posted content.

So what exactly is Section 230? As set in precedent by the CompuServe case, Section 230 protects online platforms, like social media sites such as Twitter and Facebook, from being sued for content posted by their users. Online platforms walk a tightrope between protecting free speech and shielding users from harmful content. Section 230 strikes a delicate balance by granting platforms two key immunities: Section 230(c)(1) says platforms can't be sued for content posted by their users, even if it's defamatory or offensive.[6] Section 230(c)(2)

---

[3] Stratton Oakmont, Inc. v. Prodigy Services Co., 23 Media L. Rep. 1794 (N.Y. Sup. Ct. 1995)
[4] *Section 230: Legislative History*. (2012, September 18). Electronic Frontier Foundation.
[5] S. REP. NO. 104-23, at 59 (1995)
[6] Legal Information Institute. (2018). *47 U.S. Code § 230 - Protection for private blocking and screening of offensive material*. Legal Information Institute; Cornell Law School.

allows platforms to take down content they think is “obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable”[7], without being sued for censorship.

Despite its significant contribution to online growth and free expression, Section 230 faces ongoing criticism. Some argue it gives platforms excessive power and immunity, while others fear its alteration could stifle online speech[8]. Finding the right balance between fostering innovation and protecting users from harm remains a complex challenge for policymakers and internet stakeholders as new technologies emerge on a daily basis.

Generative Artificial Intelligence

Amidst this dynamic technological landscape, the buzz surrounding artificial intelligence is now dominating headlines and promising revolution or ruin, or perhaps both. The tidal wave of AI is poised to reshape industries, and its impact on fields like law, which we recently examined through the lens of Section 230, demands careful consideration. Before algorithms rewrite legal codes, we must decipher their DNA and understand their future influence on content moderation and free speech.

Artificial intelligence (AI) is a field of computer science that is focused on the development of machines that are able to perform tasks typically done by humans. One of the earliest pioneers of AI was Alan Turing, a mathematician who proposed the Turing Test in 1950. The Turing Test[9] is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. While machine learning, which uses statistical models to analyze and predict data, marked the initial chapter of AI, the field has rapidly evolved[10]. Enter Generative AI, a groundbreaking subfield powered by “deep learning” abilities that mimic the human brain's learning process to create entirely new content.[11] This technology opened doors to unprecedented possibilities, from art and music generation to personalized healthcare and language translation. Unlike the "Turing Test" which focused on mimicking human output, deep learning mimicks how humans learn. This allows computers to understand data on a deeper level, extracting complex patterns from images, text, and

---

[7] Legal Information Institute. (2018). *47 U.S. Code § 230 - Protection for private blocking and screening of offensive material*. Legal Information Institute; Cornell Law School.
[8] Section 230: An Overview. (2021). https://crsreports.congress.gov/product/pdf/R/R46751
[9] *What is the Turing Test?* SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/Turing-test
[10] *History of generative AI*. History of Generative AI. https://toloka.ai/blog/history-of-generative-ai/
[11] *History of generative AI*. History of Generative AI. https://toloka.ai/blog/history-of-generative-ai/

audio. Unlike dated pre-programmed machines, Generative AI uses "neural networks" that learn through processing vast datasets to unlock hidden patterns and relationships.[12] Just like how humans build a mental map of the world through experiences, these networks build their own maps of information through data, allowing them to recognize patterns, make predictions, and even create new things.

The pinnacle of this evolution sits in the rise of Generative Pre-trained Transformer (GPT) models, particularly GPT-4. In 2018, OpenAI introduced the first GPT, a neural network utilizing deep learning architectures. This innovation enabled machines to not only generate text and converse with users, but also perform various language tasks with unprecedented capability[13]. GPT models can adapt their writing style to match different tones and genres, even exhibiting a level of creativity traditionally reserved for human minds.

Now, imagine scrolling through ones social media feed, only to encounter a news article written by an AI. At first glance, it appears to be a grammatically perfect and engaging article. But a nagging question inevitably arises: is this content accurate, and who is responsible for it? The author, the platform, or the algorithm that crafted it? This conundrum lies at the heart of applying Section 230 to generative AI like GPT-4. While these platforms boast remarkable creative capabilities, concerns about misinformation, bias, and even legal ramifications cast a shadow on their seemingly limitless potential. But is removing Section 230 protections from AI platforms the best solution to these limitations? All evidence points towards no.

Applying Section 230 to Generative AI:

The rise of generative AI like GPT-4 presents a novel question: should we regulate the content it creates, and what are the consequences of both action and inaction? We've all witnessed examples of its limitations firsthand. A simple query can elicit demonstrably false information, easily debunked by even a young student. I recall asking a generative AI platform for the song playing in a film scene for a paper, only to receive a seemingly credible answer with fabricated artist and timestamp, only entirely the wrong song. Luckily, my extensive film knowledge saved me from embarrassment, but such misinformation can lead to far more than a

[12] Marr, B. (2023, June 16). *A Simple Guide To The History Of Generative AI*. Bernard Marr.
[13] Marr, B. (2023, June 16). *A Simple Guide To The History Of Generative AI*. Bernard Marr.

bad grade. In the hands of professionals, it's already impacting real-world scenarios with potentially drastic consequences.

In a case that put the reliability of AI-powered research into question, a New York lawyer admitted to using ChatGPT to fabricate legal citations in a client's brief.[14] The scheme unraveled when Avianca, the opposing party, couldn't track down any of the false "precedents" cited. In fields where accuracy and integrity are paramount, widespread use of generative AI services could genuinely threaten the fabrics of a well ordered society. It is cases like these that cause legal professionals to beg for the exception of generative AI sources from the immunities provided by section 230.

This concern was reignited during the recent Supreme Court case Gonzalez v. Google[15], in which the family of a victim of the 2015 Paris attacks sued the social media platform YouTube for its recommendation algorithm playing a role in radicalizing the attackers through tailored content. Google claimed Section 230's immunity for user-generated content and argued its algorithm simply personalized content based on user preferences, not actively promoted terrorism. This case prompted lawmakers Wyden and Cox, Section 230's architects, to clarify that AI tools like ChatGPT won't be protected by Section 230's immunities.[16] Wyden emphasized that Section 230 shields platforms hosting user content, not companies' own AI-generated outputs. Similarly, Cox pointed out the law's immunity hinges on not contributing to content creation. The case resulted in the court sending the case back for reconsideration.

"Treatment of Publisher or Speaker" Clause of Section 230 and the "Black Box Problem"

In the wake of this complex case, representatives Wyden and Cox argue that Section 230 protects user speech, not companies' own actions or products. But can companies really say that they have control over their AI products? As I understand it, unlike human creators who consciously choose content, AI models operate on complex algorithms and vast datasets, often leading to unforeseen outputs. Even with careful design and mechanisms, predicting every possible output is almost impossible.

---

[14] Weiser, B. (2023, May 27). Here's What Happens When Your Lawyer Uses ChatGPT. *The New York Times*.
[15] *SUPREME COURT OF THE UNITED STATES*. (2023). https://www.supremecourt.gov/opinions/22pdf/21-1333_6j7a.pdf
[16] Analysis | AI chatbots won't enjoy tech's legal shield, Section 230 authors say. (n.d.). *Washington Post*. https://www.washingtonpost.com/politics/2023/03/17/ai-chatbots-wont-enjoy-techs-legal-shield-section-230-authors-say/

This particular dilemma has been labeled as the "Black Box Problem" by industry experts. Associate Professor of Electrical and Computer Engineering at the University of Michigan, Samir Rawashdeh, explains that humans and AI both learn by analyzing patterns in examples. In this deep learning princess, algorithms "see" examples, identify patterns, and make predictions about new things. It's like training a child to recognize cats. The Black Box Problem lies in that we can't fully understand how deep learning systems make decisions.[17] It's like trying to pinpoint why a child suddenly recognizes dogs after seeing enough examples.

While proponents of holding AI developers liable for content generated by their platforms often invoke Section 230's immunity for publishers and speakers, this invocation stumbles upon the very nature of Generative Artificial Intelligence. Unlike human users who consciously choose and express content, AI models operate on complex algorithms and vast datasets, generating outputs through opaque internal processes (hence the black box label). Tracing the reasoning behind an AI's decision is like untangling an infinite knot. For example, imagine an AI denying a loan application based on race. Training data often reflects societal biases embedded in historical records, demographics, and economic trends, and loan databases could contain past discriminatory practices that unconsciously skew the AI's understanding of creditworthiness. Even if the data bias originates from the training set, attributing responsibility directly to the developers becomes challenging when they lack complete control over the decision-making process. All evidence points toward the real speaker being the AI system itself. Holding machines legally accountable, however, necessitates a complete paradigm shift in legal frameworks and ethical considerations, moving far beyond the current Section 230 debate.

Now consider the vibrant US digital sector, valued at an astounding $3.7 trillion[18], which owes much of its success to the innovation fostering effect of Section 230. This legal shield immunity incentivizes platform development and attracts talent and investment, thereby fueling the very engine of innovation that propels American technological leadership. Section 230 is critical in fostering an environment conducive to experimentation and risk-taking, leading to groundbreaking advancements. Therefore, any potential weakening of Section 230, particularly through carving out exceptions for generative AI (GAI), necessitates careful

---

[17] Blouin, L. (2023, March 6). *AI's mysterious "black box" problem, explained | University of Michigan-Dearborn*. Umdearborn.edu.
[18] Highfill, T., & Surfield, C. (2022). *New and Revised Statistics of the U.S. Digital Economy, 2005–2021*. https://www.bea.gov/system/files/2022-11/new-and-revised-statistics-of-the-us-digital-economy-2005-2021.pdf

consideration due to the risk of stifling innovation and jeopardizing national competitiveness in this critical field.

In conclusion, attributing legal responsibility for GAI-generated content directly to developers contradicts the fundamental principles of Section 230. AI companies, akin to Twitter, function primarily as content hosts, facilitating platform interactions without exercising editorial control over the opaque outputs of their GAI models. This "black box" phenomenon, where internal algorithms churn out unpredictable results, makes direct pinpointing of the source of harmful content a complex and pointless task. Holding developers liable for user misuse not only misplaces culpability but also stifles the innovation that thrives on experimentation and risk-taking.

Therefore, the path forward lies in embracing a nuanced approach that recognizes the unique challenges posed by GAI. Instead of wielding the blunt instrument of Section 230 exemptions, policymakers should prioritize fostering a responsible AI ecosystem through targeted solutions. This includes holding users accountable for their actions, incentivizing robust content moderation tools, and promoting industry-wide adoption of ethical frameworks for GAI development. Only by striking this balance can we harness the true potential of GAI while ensuring a future where technological progress thrives alongside responsible, ethical considerations.