Logan Bachman
CS 3043
Final Paper

In today's digital world, false information spreads faster than ever, driven by an interconnected population and severely aggravated by Generative AI which creates increasingly realistic content like deepfakes. This content is differentiated as misinformation, false content shared without intent to deceive, and disinformation, false content spread on purpose, with both now recognized as a leading global risk by the World Economic Forum. As the Harvard Law Forum notes, this convergence poses a significant risk for companies and investors, directly leading to dangers such as financial fraud, severe reputational damage, and rising compliance risk due to intensifying global regulation. Furthermore, corporate governance in this area is generally weak, as data indicates that most Interactive Media companies underperform in areas like Responsible Oversight of User-Generated Content and transparency, suggesting a significant gap in preparedness against these threats, especially as new frameworks like the EU AI Act, Digital Services Act (DSA), and the UK Online Safety Act impose greater accountability.

Generative AI, including tools like large language models and image creators, has made disinformation campaigns far more powerful and effective, allowing for the rapid creation of realistic content such as deepfakes, sophisticated fake media generated by Generative Adversarial Networks (GANs) that manipulate video, audio, and text. More importantly, a Cambridge study points out the biggest problem of AI is that it uses its algorithmic systems to quickly and precisely send this false content to specific, targeted audiences. This dangerous material is categorized as misinformation (false content shared by accident) and disinformation (false content spread on purpose), which the World Economic Forum now considers both a major global risk. As the Harvard Law Forum observes, this situation creates major risks for companies and investors, threatening them with financial fraud, severe damage to reputation, and increased legal compliance burdens due to growing global regulation. This risk is made worse by the poor oversight shown by companies over user-generated content and the difficulty of using AI to accurately tell the difference between malicious disinformation and accidental misinformation without limiting free speech. As a result, new laws like the EU AI Act, Digital Services Act (DSA), and the UK Online Safety Act are demanding more responsibility from platforms, forcing them to review the dangers of their algorithms and be more transparent, though many experts argue that truly fixing the problem requires changing the core web business model that relies on advertising revenue.

AI-driven misinformation and disinformation constitute a rapidly accelerating global risk, significantly impacting various sectors: for business and investors, AI dramatically lowers the barrier for creating deepfakes and false narratives that lead to financial fraud and severe reputational damage, while the majority of companies currently underperform in the necessary governance and oversight to manage these new compliance risks (Harvard Law Forum, 2025). In public health, AI-fueled falsehoods, particularly about vaccines and treatments, are amplified by platform algorithms that prioritize content designed for outrage, making such misinformation cheaper, easier to produce at scale, and highly personalized, which contributes to a high likelihood (78%) of encountering false content when seeking vaccine information and even enables online misogyny and climate denial (BU CEID, 2024). Furthermore, democracy is threatened as generative AI is utilized by governments and political actors to manipulate public opinion and automatically censor online speech, with the proliferation of sophisticated deepfakes leading to a "liar's dividend" effect where the public grows skeptical of all information, including verifiable facts, thus undermining political stability and trust (MIT Technology Review, 2023).

Different ethical frameworks offer essential guidance for addressing the pervasive issues of AI-driven misinformation across various sectors, helping to evaluate what companies, governments, and citizens should do in response to these complex challenges. A utilitarian approach, which prioritizes minimizing harm to the greatest number of people, directly supports the actions needed to counter the risks detailed in the public health sector; under this framework, interventions like content removal or the UN's positive communication campaigns are justified to prevent widespread societal harm, such as disease outbreaks and diminished trust that results from algorithms amplifying outrage and contributing to a high rate of vaccine misinformation (BU CEID, 2024). A deontological approach, centered on duties and rules regardless of outcome, argues that companies have an intrinsic obligation not to create or profit from misleading information; this perspective underscores the failure highlighted in the business sector, where analysis shows a lack of corporate responsibility, with only about 10% of companies achieving a good rating on responsible content oversight, suggesting a widespread failure of duty that regulators like the EU are attempting to correct through acts like the DSA and EU AI Act, which impose strict new duties on businesses (Harvard Law Forum, 2025). Finally, virtue ethics encourages the creation of an organizational culture of honesty and responsibility in AI development, which is critical in combating the threats to democracy, as the widespread accessibility of generative AI is being exploited by political actors to create propaganda and implement automated censorship, eroding the public's ability to discern truth and creating a "liar's dividend" effect that fundamentally undermines trust in democratic institutions (MIT Technology Review, 2023). These ethical perspectives therefore provide the necessary tools for evaluating and determining the required responses from companies, governments, and citizens.

To stop the spread of false information made by AI, we need to try different solutions all at once. One way is using new technology: companies are making AI tools that can automatically find and flag false stories and propaganda (Columbia, 2025; Cambridge, 2024). Experts at Columbia and Cambridge also say that we should put a digital stamp, or watermark, on all AI-made pictures and words so people know right away they came from a machine, not a real person.

We also need better rules and laws for companies and governments. Companies need to create strong internal policies, and governments need new laws to deal with the money and social risks that come from all this fake content. This focus on governance is key because, as the Harvard Law Forum points out, without strong rules, businesses face huge money problems, damage to their public image, and legal risks due to deepfakes and false stories (Harvard Law Forum, 2025). New laws are also needed to fight the way AI makes political propaganda and deepfakes so easily, which hurts our democracy and creates the "liar's dividend" where people stop believing *any* news (MIT Technology Review, 2023). Lastly, we need to teach people how to protect themselves. Education and programs that teach digital skills are a vital part of the solution, helping individuals build up a defense against misleading content. The BU CEID stresses that teaching people how to spot fake news and getting different experts to work together is the best way to help rebuild public trust, especially around important issues like health and vaccines, where AI-fueled algorithms cause the most harm (BU CEID, 2024).

A clear and dangerous example of this issue appears in public health, where the stakes are literally life and death. During the COVID-19 pandemic, AI tools didn't just spread anti-vaccine messages; they amplified the content by prioritizing posts designed to generate outrage, while intentionally pushing down accurate, factual information from trusted institutions (BU CEID, 2024). This use of AI makes health misinformation much cheaper and easier to produce at scale and allows it to be highly personalized and targeted to vulnerable users. The result is a total collapse of public trust, leading to a high likelihood that people searching online for health information, such as about vaccines, will encounter false claims, sometimes up to a 78% chance of seeing misinformation. This environment brings up urgent ethical questions: Do tech companies have a responsibility to remove clearly harmful health misinformation, even if it limits some speech? From a utilitarian view, the answer is a clear "yes," because intervening to prevent the spread of medical lies minimizes harm to the greatest number of people, protecting them from vaccine hesitancy, disease outbreaks, and the exploitation by "merchants of outrage" who profit from polarizing content. Conversely, a deontological perspective argues that allowing harmful health lies to flourish on a platform is morally wrong regardless of the profit or specific outcome, imposing a strict duty on companies not to enable or profit from content that deliberately deceives the public. This situation highlights the urgent need for companies and governments to act against AI-driven misinformation, as allowing the current algorithmic system to persist is fundamentally irresponsible to public well-being.

The argument presented in this paper is that the danger of AI misinformation is systemic, driven by weak corporate oversight and algorithmic incentives, which is correct from a scholarly standpoint. However, this scholarly view contrasts sharply with popular perception. In popular media, the problem is often framed simply as "fake news" or individual bad actors creating deepfakes, such as the AI-manipulated videos of political leaders shown on social media. This popular view focuses only on the content itself and the malicious creator. The paper's scholarly argument goes deeper by identifying the root problem as the algorithmic amplification of content designed to generate outrage, the financial incentive of the advertising-driven business model, and the subsequent lack of corporate duty to protect public well-being, which are all factors largely ignored by the general public but are necessary to truly fix the crisis.

My individual contribution to this project focused on synthesizing current scholarly research to create a robust ethical and risk framework for AI misinformation. I conducted the literature review that established the three major risk categories, business/investment, public health, and democracy, which organize the main body of this paper. Furthermore, I developed the ethical analysis, applying Utilitarian and Deontological frameworks to the public health and business risks, which forms the core of the ethical argument. This research directly informed the content in the Risk Analysis and Ethical Frameworks sections. This work was intended to complement a partner's focus, such as one on developing technical defenses like watermarking AI, by clearly defining the moral necessity and legal context for why such a tool is urgently needed.

The findings and proposed solutions in this paper have significant broader impacts. They directly inform legal regulators by defining the scope of necessary accountability measures, such as those imposed by the EU AI Act and DSA, which manage legal risk for businesses. They also impact corporate boards by defining the failure in governance (only 10% good oversight) and the need for stronger internal policies to prevent financial fraud. Furthermore, they impact public health agencies by explaining the mechanics of vaccine misinformation spread and advocating for education programs to restore public trust. However, the paper faces limitations in scope and validity. The scope is heavily reliant on the success of existing or proposed policy frameworks (DSA, UK Online Safety Act). Validity is threatened by the fundamental conflict of interest: the current web business model relies on ad revenue generated by "outrage" content, making companies resistant to changes that harm their bottom line. The paper is also limited by the lack of direct primary research on algorithmic design.

In conclusion, AI has undeniably made the spread of misinformation and disinformation significantly stronger and more dangerous, but this same technology offers powerful tools to help us fight back. The danger comes from how easy and cheap AI makes it to create lies at a massive scale, leading to real-world harm like financial fraud and major reputational damage for companies, while algorithms profit from promoting outrage and causing the collapse of public trust when people seek health information (Harvard Law Forum, 2025; BU CEID, 2024). To find a necessary balance between innovation and responsibility, businesses, governments, and citizens must work together on a few key fronts. On the technical side, we can use AI against itself by developing new systems that can automatically find and flag false content and by requiring digital stamps or watermarks on all AI-made content so people know its origin

(Columbia, 2025; Cambridge, 2024). On the rules side, governments must put in place strong laws and corporate policies, like those in Europe, to enforce a moral duty on tech companies and prevent the creation of political propaganda that hurts democracy and creates the 'liar’s dividend' effect (MIT Technology Review, 2023). Finally, education and teaching people digital skills will empower citizens, making society more resistant to false information while still allowing everyone to benefit from the great potential of AI (BU CEID, 2024).

Sources:

Boston University Center on Emerging Infectious Diseases. (2024). *How Can We Tackle AI-Fueled Misinformation and Disinformation in Public Health?* https://www.bu.edu/ceid/2024/04/25/how-can-we-tackle-ai-fueled-misinformation-and-disinformation-in-public-health/

Cambridge University Press (2021, November 25) (Data & Policy). (n.d.). *The Role of Artificial Intelligence in Disinformation*. https://www.cambridge.org/core/journals/data-and-policy/article/role-of-artificial-intelligence-in-disinformation/7C4BF6CA35184F149143DE968FC4C3B6

Columbia Business School. (2025, December 10). *AI and Misinformation: How to Combat False Content. Insights Magazine.* https://business.columbia.edu/insights/magazine/ai-and-misinformation-how-combat-false-content-2025

Harvard Law School Forum on Corporate Governance. (2025, May 12). *Misinformation and Disinformation in the Digital Age: A Rising Risk for Business and Investors.* https://corpgov.law.harvard.edu/2025/05/12/misinformation-and-disinformation-in-the-digital-age-a-rising-risk-for-business-and-investors/

MIT Technology Review. (2023, October 4). *How generative AI is boosting the spread of disinformation and propaganda*. https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/