Electoral Interference and AI-generated Images: the Misinformation Board of Shame

Authors:

Savannah Harlan[1][2]

savannah.harlan@mail.utoronto.ca

Andrew McFall[3]

mcfall96@yorku.ca

Abstract

Electoral misinformation is a serious problem that has become ever more urgent in the era of generative AI and widely accessible text-to-image models. Open-source models like Stable Diffusion provide key opportunities for bad actors and rogue states to leverage the emerging technology to: 1) increase political polarization, 2) reduce public trust in democratic institutions, and 3) scale up misinformation campaigns.

This paper proposes a novel policy approach to tackling photographic (graphic) misinformation during election periods that takes lessons from the western intellectual property (IP) rights regime, which has been greatly successful at censorship efforts relative to other areas of the law like hate speech legislation.

This involves creating a non-partisan, arm's length government agency to maintain an "AI Misinformation Board of Shame" on a publicly accessible site that highlights key cases of synthetic misinformation content to promote public literacy on recognizing misinformation; create an authoritative database of synthetic, graphic misinformation; and promote public trust in democratic institutions.

Crucially, this proposal leverages market forces to create a sustainable, self-enforcing system similar to digital copyright infringement protections through: 1) financial rewards for

---

[1] M.P.P. Candidate, Munk School of Global Affairs and Public Policy

[2] M.A. Candidate, Department of Politics, York University

[3] M.A. Candidate, Department of Politics, York University

misinformation ‘whistleblowers’ paid out from fines awarded to noncompliant social media platforms, 2) a federally funded ‘reward pool’ topped up by fines to (domestic) offenders, and 3) a competitive reward system that pays platforms that detect and remove misinformation content, but only the first platform to make the report. By creating a ‘market’ for misinformation removal, such competition creates rational incentives for compliance, while reducing the regulatory burden for lawmakers.

1 Asterisks and Constraints of the Paper

We begin the paper by addressing issues of terminology choices, which can significantly influence discourse and solution feasibility. While "misinformation" and "disinformation" are often used interchangeably, scholars differentiate them based on intent, with misinformation referring to inadvertently false information and disinformation to deliberately misleading content. However, due to the challenge of determining intent and the principle of innocence until proven guilty, the paper opts to broadly use "misinformation" to encompass both.

This decision aligns with the legal principle of requiring proof of malicious intent, rather than assuming it, and emphasizes the importance of good faith in democratic discourse. By choosing "misinformation," the paper aims to encourage responsible information dissemination and discourage assumptions of malicious intent without evidence.

1.1 Why Just Graphical Misinformation?

The proposal focuses specifically on visual or graphical misinformation, such as videos or images, despite the potential effectiveness of textual misinformation. Three main reasons support this decision. Firstly, detecting AI-generated images is presently more feasible than identifying synthetic text, as existing methods for text detection can be easily evaded through techniques like watermarking and recursive paraphrasing attacks. Secondly, there is a stronger precedent for removing images from online platforms and search engine results compared to textual content, making it a more practical target for intervention. Finally, visual information is inherently more influential and can indoctrinate viewers more effectively, especially those who are inattentive or politically disengaged, as it can be processed more rapidly and does not require high literacy levels to comprehend. These factors collectively justify the proposal's focus on visual misinformation as a key area for intervention.

2 Disinformation as a Tractable Problem

The possibility of effectively combating disinformation raises significant questions, likened to the overwhelming nature of spam on the internet. Critics argue that the endless supply of bad actors, coupled with the anonymity of the internet, makes fighting disinformation akin to a "whack-a-mole" game. However, dismissing the challenge due to its complexity undermines democracy's maintenance, which requires constant vigilance. Despite the lack of a permanent solution, active engagement is imperative for electoral integrity. The Pareto principle suggests that a majority of electoral disinformation may stem from a small number of actors, emphasizing the importance of data collection to understand the scope of the problem. Technological advancements and platform changes, such as tighter registration policies and browser tracking, offer potential solutions. While privacy concerns persist, leveraging existing policy frameworks and technological tools is necessary. Moreover, addressing AI-generated graphical misinformation is deemed even more manageable, as misinformation historically originates from human sources. Thus, while challenging, combating disinformation is both feasible and essential for safeguarding democratic processes.

2.1 Disinformation and Security Policy

The security implications of disinformation, highlighted by instances of foreign interference in Canadian elections, underscore the urgency of addressing the issue. The diverse motivations of bad actors make it challenging to ascertain the origin and intent behind misinformation, emphasizing the need for proactive measures. While prevention of foreign interference may be difficult due to the sophistication of actors and the inherent openness of the internet, cooperation between nations presents an opportunity for mitigating the impact. Western dominance in internet traffic, particularly through US-owned platforms, suggests potential for bilateral cooperation akin to efforts in enforcing intellectual property rights. With the United States and Canada's strong diplomatic relationship, collaboration in addressing disinformation can be a pragmatic and effective approach to safeguarding democratic processes.

3 The Freedom of Expression and the Marketplace of Ideas

The concept of the marketplace of ideas, popularized by John Milton in his 1644 work *Aeropagitica* forms the basis of arguments in favor of free expression and against censorship. It suggests that in an open exchange of ideas, truth will prevail as competing ideas vie for

acceptance, akin to competition in a free market economy where the best products succeed. This notion is rooted in the belief that democracies thrive on the free flow of ideas and minimal state intervention, allowing reasoned debate to flourish.

However, some critique this as a dogmatic belief in the inherent triumph of truth, exemplified by Justice Oliver Wendell Holmes's assertion that "the best test of truth is the power of the thought to get itself accepted in the competition of the market." Despite criticisms, the marketplace of ideas remains a cornerstone of Western liberal thought, emphasizing the value of competition and fairness in the pursuit of truth.

3.1 Free Markets Do Not Work

The section critiques the analogy that equates the free market with a self-regulating 'marketplace of ideas' by questioning the effectiveness of unregulated markets. We argue that the belief in markets naturally achieving a beneficial equilibrium is empirically unfounded, as demonstrated by the economic collapse during the Great Depression, which led to the rise of Keynesian economics. Keynesian principles challenge the neoclassical notion that markets are self-correcting, especially in the short-term scales that impact people directly, advocating instead for the necessity of state intervention in monetary policy to mitigate economic crises.

The section further explores how laissez-faire capitalism often results in reduced competition and growth, as market forces inherently produce winners and losers, leading to a concentration of market power in few hands. This concentration diminishes competition and increases the influence of dominant firms on politics and policy, making it difficult to reintroduce true competition. This discussion suggests that relying solely on neoclassical economic theories for immediate fiscal policy is misguided, as these theories typically assume conditions that only hold over the long term.

3.2 Free Markets of Ideas Do Not Work Either

We question the belief in the self-correcting nature of the marketplace of ideas, paralleling critiques of unregulated free markets. This notion presupposes that individuals are exposed to diverse viewpoints, yet personalized content algorithms and online echo chambers challenge this assumption. Research suggests that internet users often inhabit tailored online environments, limiting exposure to opposing views. Additionally, psychological heuristics can distort perceptions and lead to suboptimal decision-making when evaluating ideas. For example, the availability heuristic influences individuals' judgments by prioritizing easily

accessible information. Thus, the efficacy of the marketplace of ideas in fostering informed debate and truth-seeking is called into question amidst contemporary digital landscapes and psychological biases.

4 Platforms, Social Media, and Intellectual Property Law

The intersection between platforms, social media, and intellectual property (IP) law presents opportunities for combating misinformation through responsible content moderation. Rather than focusing on individual users - an approach that requires government intervention and vast resources to targets users on platforms - a platform-focused process places the onus on social media platforms to moderate themselves through IP law.

4.1 Target Gatekeepers and Platforms, Not Users

Addressing the issue of misinformation on social media platforms requires the monitoring and regulating of online content. Misinformation often proliferates through social media platforms where the platforms themselves are the arbiter and gatekeeper of online content. In other words, platforms are entities with the capability to control and regulate content. A single platform can host millions of users at any given time, while there are only a handful of platforms. Because of this, it is desirable to direct efforts towards online gatekeepers, as it promises a more efficient and scalable approach to regulation.

4.2 Learn from IP Law, Not Hate Speech

Taking lessons from strong and enforceable laws, such as IP law, can strengthen and support possible policy futures for fighting against misinformation. This is in contrast to more obfuscated and nebulous laws such as hate speech. Hate speech falters due to legislation, the judicial system and prosecution. There is a high barrier to reporting (police report), the court systems are slow, costly, and requires a high burden of proof, and it is almost impossible to prosecute individuals for hate speech.

Intellectual Property is very powerful in the West and has spread across the world, in part, due to the World Trade Organization. IP is enforced by private companies at the level of service providers and not users. A few examples include: Googleâ€™s search and DMCA requests, Youtube copyright takedowns, and Metaâ€™s copyright reports. Private social media platforms work within market forces and are costless to the government. IP law allows for market forces, guided by state policy, to create a very responsive, low burden of proof (or sometimes none at all), and scalable systems. IP law enforcement has advanced so much that

some have automatic detection systems to immediately take down copywritten content. This technology can be utilized to greatly reduce the power and influence of misinformation by blocking it before any eyes see it. Moreover, once a given image has been â€œcopystrikedâ€ it becomes easily detectable and all other instances of it can be targeted. This helps avert the spread of misinformation. By leveraging insights from IP law, platforms can play a crucial role in mitigating the spread of false information and promoting a more trustworthy online environment.

5 Policy Proposal

We propose a novel policy approach to tackling misinformation that leverages a combination of rewards for participants (reporters) and fines for non-compliance and bad actors to create a 'market' for fighting disinformation. The goal is to create a self-enforcing system that leverages market forces and competition to shift the burden of enforcement from the state to firms, platforms, and private individuals, similar to the way carbon pricing schemes with 'cap and trade' policies automate the accounting of negative externalities.

5.1 The AI 'Misinformation Board of Shame'

The first component of this approach involves creating a publicly accessible site to allow participants to highlight and report AI-generated online misinformation. Crucially, this site should have 1) minimal barriers to entry, meaning no requirements to sign-up or create an account before a report can be made; 2) information presented in an easily readable and publicly accessible format, meaning the identifying elements of the misinformation should be well explained and easily understood; and 3) measures against spam and abuse of the reporting system, such as a lone individual making an excessive number of reports to slow down the system.

5.1.1 Goals

The goal of this proposal is twofold. First, the 'Board' should educate users and the public by making it clear in each instance how one can determine that the implicated content is synthetic (AI-generated) and ways to spot future, similar cases of misinformation. For instance, in the following AI-generated image of Trump posing with African-American voters, the board

should describe the overly smooth skin tone of the subjects in the photo (a hallmark of current AI-generated images), the abnormal and mutated text on the hat of the second subject from the left, and the strange distortion and number of fingers of the subject on the far left. The point is to inform voters so as to educate the public and allow them to independently judge and discern future synthetic content they encounter.

Second, the Board should give individuals an easy and reliable source to say “see, I’m right!” when engaging in informal discussions and political ‘thanksgiving dinner’ debates. This is especially important because of the role of personal relationships and interactions in fighting misinformation. Beyond political science students, most arguments over politics occur in casual settings between friends and family members. In such scenarios, the main incentive for fact-finding and adherence to the truth depends on the personal pride of the individuals arguing and the desire to demonstrate superiority of knowledge and information.

We contend that this is highly significant as a realistic model of fighting misinformation as it implies that lofty goals such as ‘educating the public’ take a backseat to smug assertions of being correct. Furthermore, it can be argued that settling a score on truth is easier between individuals that already have some personal relationship in the sense that people are usually more likely to be convinced when arguing with friends or family, as opposed to pointless internet debates that often end in contrarianism or recourse to logical fallacies.

5.1.2 Elements of the Board

The Board consists of four elements. First, it should name and shame the offender, meaning the full name of the user that published the misinformation or their internet handle (“username”). Second, it should name and shame the platform or site responsible for hosting the misinformation, for example “Facebook” or “Reddit (/r/politics)”. Third, it should name and praise the whistleblower or reporter optionally, as a means of assigning credit or boosting the reporter’s ego, since personal pride and reputation may be the main incentive for some individuals to participate in the system. Fourth, the Board should provide key information on how to detect the misinformation in the content and advice on how to recognize similar patterns and telltale signs in the future.

5.1.3 Considerations

There are two main considerations that we raise in the context of our proposal. First, for obvious reasons, the Board should be operated by an arm’s length agency outside of political

influence like Elections Canada. This is to discourage political interference in the system as well as ensure that all parties and voters perceive the Board as legitimate and politically neutral. In addition however, we consider that not all government agencies have the same (or a positive) public perception. For instance, public opinion research routinely shows that agencies such as the Canada Revenue Agency are widely viewed in a negative way due to the slow response time of officials, a long backlog of cases, and poor communication interfaces (such as understaffed phone lines with excessive wait/hold times) compared to agencies like Elections Canada.

Second, to prevent spam and bad actors from gumming up the system by making excessive amounts of false reports, the Board should require participants to submit their social insurance number (SIN) when making reports. Since SINs are available to all Canadian residents over the age of 15—anyone eligible to work in the country—this reduces the barriers to accessibility by allowing immigrants and permanent residents to make reports. At the same time, requiring a SIN prevents foreign actors from creating havoc since it is unlikely that they will have access to a SIN. It furthermore discourages false reporting by creating an air of 'official legitimacy' in the system since SINs are easily traceable to an individual's personal information and allows the state to identify who is responsible when abuse of the system occurs.

This can be further augmented by introducing a 'graduated trust' system that limits the number of initial reports a participant can make, for instance limiting new reporters to just 3 reports per SIN, until their initial reports have been verified and approved, and gradually increasing the number of reports they can make.

5.2 Creating a Market for Fighting Misinformation (Leveraging Market Forces)

A crucial aspect of our proposal concerns the use of market forces to counter AI-generated misinformation online. Our analysis of existing approaches to tackling misinformation finds that most proposals fall flat on the basis of either excessive ongoing costs, or the difficulty of state agencies with limited resources to prosecute non-compliance. For example, a key limitation of provincial labour relations boards for enforcing workers' rights under contracts and statutory provisions lies in the lack of manpower for on-site inspectors and funding for auditors. Creating a market for fighting misinformation allows the state to fill these gaps by shifting the incentives and penalties for (non)compliance to private actors like

social media platforms which have more resources and better tools and technology to fight misinformation.

5.2.1 Rewards

The first element of this market involves creating a system for the dissemination of rewards through incentives for users and platforms to report misinformation without delay. We propose developing a financial pool for paying out rewards to participants that is supported primarily from fines awarded to noncompliant platforms and sites. Thus for instance, if a user reports misinformation from an image on a Facebook post online, the reward they receive comes directly from the corresponding fine given to Meta, the parent company of Facebook. This effectively reduces the nominal cost of the rewards system to zero, since the agency assumes only an administrative role in the system, while delegating the responsibility of financial renumeration to firms and individuals.

Additionally, if the actor responsible for a piece of misinformation is deemed to be a user from Canada—and thus under Canadian law—a fine should also be awarded to them, further contributing to the available pool of financial rewards.

Second, the amount of the fine should also be tied to aspects like the amount of time that the misinformation-containing content has been on the platform, for instance (*x* amount multiplied by the *y* number of days the content has been posted). This creates an incentive for firms to act fast and proactively, as opposed to waiting until a report has been made, or public attention has been generated.

Conversely, if platforms identify some misinformation and are the first ones to report and remove it, they are given the financial reward paid out from a combination of previous fines and some investment into the pool of rewards. Importantly, platforms should be allowed to report content not just on their own platforms, but on other sites and platforms as well. While this is unlikely to be a significant source of reports, it nonetheless encourages the natural incentives for competition necessary for implementing an effective ‘market’.

We might also consider allowing private individuals to make tax-deductible donations to this ‘misinformation fund’ solely as a means of bridging the gap between rewards and fines, while also promoting some semblance of civic responsibility or pride in maintaining the legitimacy of the democratic system.

References

1. Boda, Károly, Ádám Máté Földes, Gábor György Gulyás, and Sándor Imre. "User tracking on the web via cross-browser fingerprinting." In *Information Security Technology for Applications: 16th Nordic Conference on Secure IT Systems*, NordSec 2011, Tallinn, Estonia, October 26-28, 2011, *Revised Selected Papers* 16, pp. 31-46. Springer Berlin Heidelberg, 2012.
2. Crouch, Colin. "Privatised Keynesianism: An unacknowledged policy regime." *The British journal of politics and international relations* 11, no. 3 (2009): 382-399.
3. Gill, Stephen. "Market civilization, new constitutionalism and world order." In *New constitutionalism and world order* 29 (2014): 30.
4. Intumwayase, Jean Luc, Imane Fouad, Pierre Laperdrix, and Romain Rouvoy. "UA-Radar: Exploring the Impact of User Agents on the Web." In *Proceedings of the 22nd Workshop on Privacy in the Electronic Society*, pp. 31-43. 2023.
5. Jamali, Lily, and Daniel Shin. "FCC cracks down on AI robocall scams, Meta tightens oversight of AI content and Sen. Klobuchar discusses Section 230 reform." *Marketplace*. February 9, 2024.
6. Kaczynski, Andrew, and Em Steck. “Fake Joe Biden Robocall Urges New Hampshire Voters Not to Vote in Tuesday’s Democratic Primary.” *CNN*, January 22, 2024.
7. Lerche, Charles O. "Jefferson and the election of 1800: A case study in the political smear." *The William and Mary Quarterly: A Magazine of Early American History* (1948): 467-491.
8. Lombardi, Claudio. "The Illusion of a “Marketplace of Ideas” and the Right to Truth." *American Affairs* 3, no. 1 (2019): 198-209.
9. Milton, John. *Areopagitica* (1644; Cambridge: Cambridge University Press, 1918) in Jill Gordon, “John Stuart Mill and the ‘Marketplace of Ideas,’” *Social Theory and Practice* 23, no. 2 (Summer 1997): 235–49.
10. Mitchell, Eric, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. "Detectgpt: Zero-shot machine-generated text detection using probability curvature." In *International Conference on Machine Learning*, pp. 24950-24962. PMLR, 2023.
11. Osman, Laura and Robertson, Dylan. “What We Learned from the Inquiry into Foreign Meddling in Canada’s Elections,” *CTVNews*, April 13, 2024.

12. Rooney, Paula. “Microsoft's CEO: 80-20 Rule Applies To Bugs, Not Just Features.” *CRN*, 03 October 2002
13. Sadasivan, Vinu Sankar, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. "Can AI-generated text be reliably detected?." *arXiv preprint* arXiv:2303.11156 (2023).
14. Sanders, Nathan E., and Bruce Schneier, “How Chat-GPT Hijacks Democracy,” *NYtimes*, 15 Jan 2023.
15. Silverman, Craig and Kao, Jeff. “Infamous Russian Troll Farm Appears to Be Source of Anti-Ukraine Propaganda.” *ProPublica*, March 11, 2022.
16. Skelley, Geoffrey. “Why Biden Probably Won’t Get A Serious Primary Challenger.” *FiveThirtyEight*, 5 July 2023.
17. Wherry, Aaron. “It’s a Shame We Didn’t Have Trudeau’s Testimony on Foreign Interference Earlier. Much Earlier.” *CBC News*, April 11, 2024.