THE LEGAL IMPLICATIONS OF AI HALLUCINATIONS

PRESENTED BY
Zoe Brown
Mike Chen
Solyana Gedlu
Liam Gerry
Arif Hussain
Laura Olarte
Celine Tsang
Malka Younas

With the leadership of:
Kris Choi
Samir Reynolds

June 1, 2024

Future of Law Lab
University of Toronto Faculty of Law

UNIVERSITY OF TORONTO FACULTY OF LAW

FUTURE OF LAW

About the Future of Law Lab

The Future of Law Lab is a platform for students, academics, lawyers, and other professionals to participate in collaborative initiatives exploring how the law will evolve in the future. We will dive into the intersection of law, technology, innovation, and entrepreneurship, with programing dedicated to each of these streams. As a hub of interdisciplinary activity, we are dedicated to bringing together individuals from all backgrounds to examine the changing face of the legal profession.

About the Working Group
This report is written by Zoe Brown, Mike Chen, Solyana Gedlu, Liam Gerry, Arif Hussain, Laura Olarte, Celine Tsang, and Malka Younas under the supervision of Kris Choi, and Samir Reynolds, upper-year student leaders of the group.

This report is intended to be a primer that introduces the world of AI hallucinations and starts some discussions about their potential. The paper discusses four broad issues: liability for legal harms from hallucinations; AI imitations; privacy issues from generative AI and hallucinations; and equity issues with generative AI and hallucinations.

About the University of Toronto Faculty of Law

Established in 1887, the Faculty of Law is one of the oldest professional faculties at the University of Toronto, with a long and illustrious history. Today, it is one of the world's great law schools, a dynamic academic and social community with more than 50 full-time faculty members and up to a dozen distinguished short-term visiting professors from the world's leading law schools, as well as 600 undergraduate and graduate students.

The Faculty’s rich academic programs are complemented by its many legal clinics and public interest programs, and its close links to the Faculty's more than 6,000 alumni, who enjoy rewarding careers in every sector of Canadian society and remain involved in many aspects of life at the law school.

The Legal Implications of AI Hallucinations

Intro

As Artificial Intelligence (AI) becomes increasingly prevalent, there are many novel issues that arise. The release of ChatGPT in 2022 introduced the broad public to generative AI, which is an AI system designed to create entirely new content. Practically, this usually means taking prompts and creating an output like text in the case of Large Language Models (LLMs) like ChatGPT or images with programs such as DALL-E. Generative AI opens up many possibilities, including the ability to create outputs that are not explicitly programmed by developers. For example, ChatGPT can provide coherent and digestible answers to many questions asked of it without any direct human involvement.

One major risk with generative AI comes from what are called hallucinations. AI hallucinations is an emerging phenomenon that occurs when AI technology generates cases that never actually happened and therefore do not actually exist. An AI hallucination, then, as the name might imply, is a product that AI "hallucinates", or effectively "makes up." Hallucinations stem from the mandate of LLMs to generate coherent text as output, but this text is not always linked to reality. For example, there are multiple instances of lawyers (in British Columbia and in New York) submitting documents to courts that contained hallucinations. In these cases, the lawyers had used ChatGPT to write their submissions, and ChatGPT included case law that did not exist.

Examples of AI hallucination having negative impacts on people are becoming more common as generative AI is becoming more pervasive. This leads to a number of interesting questions about the legal implications of AI hallucinations specifically and generative AI more broadly. This paper explores some of these issues and the legal harms that can be caused by AI hallucinations. The paper does not purport to be an exhaustive account of the potential risks of AI hallucinations, as there are too many potential outcomes to predict. Rather, this is intended to be a primer that introduces the world of AI hallucinations and starts some discussions about their potential. The paper discusses four broad issues: liability for legal harms from hallucinations; AI imitations; privacy issues from generative AI and hallucinations; and equity issues with generative AI and hallucinations. Each section includes real-life examples and concludes with some key takeaways. As this area is incredibly dynamic, there will undoubtedly be many more issues that arise within and beyond these categories, but hopefully the general principles discussed here can help inform other specific problems as well.

Liability for AI Harms - Who Pays the Bill?

Introduction

When individuals make decisions or produce outputs, the connection between those decisions and outputs and their origins is typically straightforward, simplifying the process of determining liability. However, when it comes to determining who is liable for harm caused by AI, the situation is much more convoluted. Three main challenges emerge here:[1]

1. **The "black box" problem**: this arises due to the opacity surrounding the process by which the AI arrives at conclusions, makes decisions, and produces outputs. When AI programmers lack precise understanding of how a particular output was produced, establishing causation or fault becomes potentially impossible. Stakeholders, such as businesses and affected individuals, may justifiably demand transparent explanations and rationales for AI outputs, but these are not currently well-understood.[2]

2. **The multi-stakeholder dilemma**: this emerges with the complexity of AI systems involving numerous contributors throughout the development and implementation phases. Determining which stakeholder(s) bear responsibility for any resultant harm becomes a daunting task.[3]

   For instance, legal regimes such as product liability may facilitate the distribution of liability by prescribing joint liability between potential tortfeasors. The number of stakeholders involved in the creation and operation of AI systems is rising, and they may all have a role in ensuring that its outputs don't result in harm. Moreover, in some cases, it would be necessary to determine which commercial parties along this chain of production and operation can be held liable, which could be impossible when hallucinated outputs, actions, or decisions are autonomously produced by AI.[4]

3. **Autonomy**: this arises when an AI system operating without direct human involvement potentially causes harm through its decision-making process. Conventional liability principles which rely on human agency, control, foreseeability, causation, and intention are difficult to apply in this context.[5] As AI technology becomes more autonomous,

---

[1] Karine Joizil, Adam Goldenberg & Sherry Ghaly, "Could AI get you sued? Artificial intelligence and litigation risk" (26 April 2022), online: <https://www.mccarthy.ca/en/insights/blogs/techlex/could-ai-get-you-sued-artificial-intelligence-and-litigation-risk>.
[2] Richard Stobbe, "Addressing the Legal Issues Arising from the Use of AI in Canada" (September 2023), online: <https://www.fieldlaw.com/News-Views-Events/219302/Addressing-the-Legal-Issues-Arising-from-the-Use-of-AI-in-Canada>.
[3] Joizil, *supra* note 1.
[4] Yaniv Benhamou & Justine Ferland, "Artificial Intelligence & Damages: Assessing Liability and Calculating the Damages" in Giuseppina D'Agostino, Carole Piovesan & Aviv Gaon, eds, *Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law* (Thomson Reuters, 2021) 165.
[5] Peter Asaro, "The Liability Problem for Autonomous Artificial Agents" (Paper delivered at the Association for the Advancement of Artificial Intelligence Spring Symposium, 5 March 2016) online: <https://peterasaro.org/writing/Asaro,%20Ethics%20Auto%20Agents,%20AAAI.pdf>.

pinpointing the party responsible for damage becomes harder.[6]

> While AI systems today are unlikely to be held liable for their own actions and outputs due to the absence of their own legal personality, advancements on the horizon present alternative possibilities. For instance, AI-based medical systems today are viewed merely as tools for providing more advanced care to patients. However, future AI-based medical devices may operate autonomously, surpassing human abilities and working without human supervision. At this point, their legal status would need to be carefully evaluated with respect to issues of negligence liability (medical malpractice) and product liability.[7]

Moffatt v Air Canada[8]: AI Hallucinations in Commercial Relations

The recent decision in *Moffatt* grappled with the question of whether a company, in this case Air Canada, can be held liable for providing misleading information resulting from hallucinations from automated chatbot on its website. The BC Civil Claims tribunal held that Air Canada was liable for the negligent misrepresentations made by the AI Chatbot used on its website. Air Canada argued that the chatbot should be considered a separate legal entity responsible for its own actions, thus absolving the company of any liability. The tribunal rejected the argument, stating that the chatbot is just a part of Air Canada's website; Air Canada was held to be responsible for the misleading information provided by the chatbot.[9]

This decision highlights that AI hallucinations that misinform consumers render businesses vulnerable to liability. This legal exposure to AI hallucinations can come at a price. Air Canada attempted to argue that a chatbot was more analogous to an agent or representative, drawing on concepts of vicarious liability available in employment contexts. The tribunal rejected this autonomy argument, which suggests that, with respect to consumer-facing AI technologies, courts are likely to pin liability on the businesses deploying the technology for resulting harms.[10]

Zhang v Chen[11]: AI Hallucinations in the Legal Profession

Also in British Columbia, a recent decision dealt with AI hallucinations in a lawyer's brief. In submissions on costs in a family law matter, a lawyer for one of the parties submitted a brief that was generated by ChatGPT that created fake cases. Opposing counsel asked the court to order

---

[6] Joizil, *supra* note 1.
[7] Ahmed Eldaka et al, "Civil liability for the actions of autonomous AI in healthcare: an invitation to further contemplation" (2024) 11:305 Humanities & Soc Sciences Communications 1 at 3.
[8] *Moffatt v Air Canada,* 2024 BCCRT 149.
[9] Barry Sookman, "Moffatt v. Air Canada: A Misrepresentation by an AI Chatbot" (19 February 2024), online: <https://www.mccarthy.ca/en/insights/blogs/techlex/moffatt-v-air-canada-misrepresentation-ai-chatbot>.
[10] *Ibid.*
[11] *Zhang v Chen*, 2024 BCSC 285.

the lawyer who used ChatGPT to pay the legal costs of researching the fake cases. Although the Court declined to order these costs because of a lack of intent to deceive, the decision admonished the use of generative AI in legal proceedings.

Although not imposing liability in the same way as *Moffat*, *Zhang* made clear that those who use generative AI in a professional setting (as a lawyer) are responsible for ensuring the accuracy of the statements made by the AI. The Court went as far as to say that "generative AI is still no substitute for the professional expertise that the justice system requires of lawyers."[12]

Takeaways

Despite AI hallucinations, proponents of generative AI champion AIs' usefulness in saving time, and by extension, money (as the old adage goes, 'time is money'). Efficiency and cost-effectiveness are certainly noble and welcomed pursuits in many fields. However, the cases discussed above illustrate how AI hallucinations may inadvertently backfire and ironically result in exactly the opposite: delays and extra expenses. In *Zhang*, the Court noted that the lawyer's use of AI-generated cases created delays in court procedures, and that the lawyer should be held liable for 2 days' worth of court time (which could translate into thousands of dollars).[13]

In the aftermath of the COVID-19 pandemic, access to justice remains a topical issue in the legal field. Legal services are expensive, and the legal system itself is often backlogged. As one legal librarian notes, additional costs stemming from AI-hallucinated cases or difficulties stemming from legal resolutions to hallucinations "stand to slow down and raise costs in an already overburdened legal system."[14]

AI-related private law disputes will continue to rise. As courts initially deal with liability issues upon AI hallucination, they will likely follow *Moffatt* and *Zhang* and allocate risks through preexisting legal concepts, including contract law and professional responsibilities. However, there is always the potential for unforeseen issues in areas, such as complicated multi-party disputes where some parties did not realize they were interacting with AIs.[15]

Future cases will likely deal with arguments about autonomy, multiple stakeholders, and the "black box" problem in greater depth. One way to address AI hallucinations is through statutory guidelines on the use of AI, be it by professional regulatory bodies or consumer protection regimes. Such guidelines may include the duty to disclose the use of AI or the duty to manually

[12] *Ibid* at para 46.
[13] Sookman, *supra* note 9.
[14] *Ibid*.
[15] See e.g. Meghan Higgins, "Air Canada chatbot case highlights AI liability risks" (27 February 2024), online: <https://www.pinsentmasons.com/out-law/news/air-canada-chatbot-case-highlights-ai-liability-risks>.

verify content generated by AI.[16] Those deploying or developing AI systems, regulatory bodies, and courts will all have to consider these challenges in developing a robust AI liability regime.

AI Imitations

The Problem of AI Imitations

Imagine you are an avid reader and you are waiting for your favourite author to release the next novel in their series. Unfortunately, this author is taking too long to put out new work. This prompts you to ask ChatGPT to write the next book in the series for you, which it does convincingly, since it was trained on that author's work. Although the output is not a hallucination in the way that we normally understand, it is still wholly created by the AI. This hypothetical was put forth by the National Post, emphasizing the need for robust regulations surrounding generative AI.[17]

The hypothetical mentioned above is quickly becoming reality. Top authors, like George R.R. Martin, are bringing lawsuits against the use of their works for training LLMs without appropriate permission. Moreover, imitations like the one in the hypothetical are not exclusive to literary works. Two of Canada's biggest artists were victims of a similar type of imitation. The song "Heart On My Sleeve", which went viral on social media, was AI-generated to sound like a collaboration between Drake and The Weeknd, by imitating both the artists' voices and their musical styles.[18] By using the artists' discographies as training materials, the AI system was able to produce a realistic-sounding duet between the two artists that convinced many on social media. Although the song was very popular on social media, Universal Music Group (UMG), the record label representing both artists, was certainly not a fan. The label used the American Digital Millennium Copyright Act (DMCA) to take down the song from platforms while the label investigated further. Interestingly, the AI system inadvertently included a producer tag (a feature used by many hip-hop producers that is essentially an audio signature) associated with the prolific producer, Metro Boomin.[19] This audio watermark was probably central to UMG's success. Despite UMG being able to successfully take down this song, similar incidents could

[16] Lindsay Frame & Nico Rullmann, "Landmark Decision about Hallucinated Legal Authorities in BC Signals Caution…But Leaves Questions About Requirement to Disclose Use of AI Tools" (21 March 2024), online: <https://www.mccarthy.ca/en/insights/blogs/techlex/landmark-decision-about-hallucinated-legal-authorities-bc-signals-caution-leaves-questions-about-requirement-disclose-use-ai-tools>.
[17] Anja Karadeglija, "Unclear how Canadian copyright law applies to generative AI: government document", *National Post* (5 October 2023), online: <https://nationalpost.com/news/politics/canadian-copyright-law-ai>.
[18] Joseph Grasser & Susie Ruiz-Lichter, "Ghostwriter in the Machine: Copyright Implications for AI-Generated Imitations" (17 May 2023), online: <https://www.natlawreview.com/article/ghostwriter-machine-copyright-implications-ai-generated-imitations>.
[19] *Ibid.*

present greater difficulties for those without UMG's massive resources, like the author in the hypothetical above.

The Legal Background

Individuals are afforded rights to control the sale and commercial use of their identity through a variety of identifiers. These rights, often called personality rights or rights of publicity, are protected under the common law in Canada.[20] More specifically, the torts that protect these rights include misappropriation of personality and passing off. It is difficult to say that these torts can directly resolve the issues arising out of the hypothetical and similar cases; however, there may be room for opportunity.

The two components that are central to the tort of misappropriation of personality include the lack of the plaintiff's consent and the defendant's commercial gain as a result. While the first component is almost always present in cases like the hypothetical, the second component raises some issues. Although every case will be different, it is unclear whether commercialization is relevant in the context of AI imitations. However, the distinction between sale and subject is at the heart of condemning unauthorized usage of one's personality. This may assist courts to analogize misappropriation of personality to cases like the hypothetical. The permission of unauthorized subject-based use is justified, since it allows the public to learn more about the person of interest. This would not be the case here. Instead, the viewer, listener, or reader learns nothing more about the original creator. Rather, they might distance themselves from the original creators and their works as a result of being satisfied by the AI-created copycat.

Passing off also deals with deceptive marketing. This could be better suited for dealing with AI imitations than misappropriation of personality, because passing off is more concerned with the idea of false endorsement. It also deals with harms to the plaintiff more than benefits to the defendant. The Supreme Court of Canada set out a three-part test for passing off in *Ciba-Geigy Canada Ltd v Apotex Inc*[21]:

1. Existence of goodwill
2. Deception of the public due to a misrepresentation
3. Actual or potential damage to the plaintiff

The first component is troublesome, since AI Imitations may not deal with a "mark" in the typical sense. Additionally, the person who prompts the AI system to output the imitation will not be deceived, since they initiated it. However, imitations can easily deceive the public, as shown by the imitation of Drake and The Weeknd discussed above. Damage to the original

[20] Jill Tonus & Tamara Celine Winegust, "Canada: A problem on the rise" (5 January 2015), online: <https://www.worldtrademarkreview.com/article/canada-problem-the-rise>.
[21] *Ciba-Geigy Canada Ltd v Apotex Inc,* [1992] 3 SCR 120.

creator is practically inevitable and can manifest itself in two ways. There could be damage to personality if the AI creation is not well-received, yet presumed to belong to the original creator. Alternatively, there could be commercial damage if the public opts out of purchasing original works if a comparable, yet cheaper (or free), option is available through AI.

Takeaways

The issue at hand is that the hypothetical does not involve "marks," but rather copyrighted materials. Intuitively, a copyright-based action seems like the most obvious option. Unfortunately for creatives, AI imitations may not technically infringe on their copyright, since they are merely in the style of their works.[22] Creatives will also have a hard time if they base their claims on AI input (i.e. using their work for training), because fair use can act as an obstacle. This is why personality rights, if further developed, might help creatives succeed in situations like the ones discussed – at least until robust amendments are made to copyright legislations.

AI Hallucinations & Privacy Implications

The capabilities of AI systems would not be possible without the vast amount of data available today. The data is often scraped from the web, and even though training data is often public, there may still be personal information. When an organization collects, uses, or discloses personal information to train and deploy its AI system, it is subject to privacy and data protection regulations in Canada.[23] The purpose of these regulations is to prevent and address unauthorized uses of personal information that could have a detrimental impact on individuals and the public. The associated risks increase with generative AI, because systems may disclose personal information used in the training. Although this does not necessarily involve hallucinations, this still involves generative AI working beyond its intended uses and is worth discussing here. These effects could potentially damage someone's reputation or even identify an individual despite the information usually being anonymized, de-identified, or not even used in the training data.[24] Some of the primary concerns regarding privacy and data protection with generative AI are:

[22] Rachel Reed, "AI created a song mimicking the work of Drake and The Weeknd. What does that mean for copyright law?" (2 May 2023), online: <https://hls.harvard.edu/today/ai-created-a-song-mimicking-the-work-of-drake-and-the-weeknd-what-does-that-mean-for-copyright-law/>.
[23] Private organizations are subject to the Personal Information Protection and Electronic Documents Act; and if passed, organizations will have to comply with Bill C-27, which includes the Consumer Privacy Protection Act, which will replace PIPEDA, the Personal Information and Data Protection Tribunal Act, and the Artificial Intelligence and Data Protection Act.
[24] See Drew Breunig, "Considering AI Privacy scenarios" (15 May 2023), online: <https://www.dbreunig.com/2023/05/15/ai-privacy-scenarios.html#section-2>.

- **Leakage of personal information:** A recent study at Stanford University showed that generative AI systems are riddled with hallucinations, misinformation, and bias.[25] The researchers found that GPT-4 is more likely to leak sensitive information than GPT-3.5: GPT-4 is more likely to reveal people's email addresses and phone numbers, although it is more careful with social security numbers.[26] When ChatGPT was asked about Mat Honan, the editor-in-chief of the MIT Technology Review, it confirmed that he had a wife and two daughters and offered his work address, phone number, and credit card number.[27] This was a true hallucination, as most of the personal information was inaccurate. This still goes to show that the consequences of personal information (whether accurate or not) that generative AI exposes can lead to physical, economic, reputational, emotional, and relational harms.[28]
- **Reputational damage (defamation):** AI hallucination can damage someone's reputation if inaccurate personal information is disclosed. For example, *BlenderBot*[29] was asked about Maria Renske Schaake, a Dutch politician, former member of the European Parliament, and fellow at Stanford's Institute for Human-Centered Artificial Intelligence. The bot responded that Schaake was a terrorist because she wrote an article for the Washington Post that used the words, "terrorism" and "terror."[30] Falsely and authoritatively deeming someone a terrorist could clearly harm someone's reputation. Because of the evident risks to consumers, the Federal Trade Commission (FTC) in the United States is (as of time of writing) investigating OpenAI to determine whether the company engaged in unfair or deceptive practices that caused reputational harm to consumers.[31] Inaccurate inferences can affect people's access to employment, credit, education, and more.
- **Re-identification of individuals:** Some AI systems can infer information even when personal information has not been used to train the system or are able to infer information from the data.[32] For example, an AI system trained on medical data accurately predicted

[25] Prabha Kannan, "How Trustworthy are large language models like GPT?" (23 August 2023), online: <https://hai.stanford.edu/news/how-trustworthy-are-large-language-models-gpt>
[26] Boxin Wang et al, "Decoding trust: A comprehensive assessment of trustworthiness in GPT models" (26 February 2024), online: <arXiv:2306.11698v5> at 44.
[27] Melissa Heikkila, "What does GTP-3 "know" about me?" (31 August 2022), online: <https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/?trk=article-ssr-frontend-pulse_little-text-block>
[28] Daniel Solove & Danielle Citron, "Privacy Harms" (2021), GW Law Faculty Publications & Other Works 1534, online: <https://scholarship.law.gwu.edu/faculty_publications/1534>.
[29] Meta's chatbot that can research the internet to talk about different topics. See: https://about.fb.com/news/2022/08/blenderbot-ai-chatbot-improves-through-conversation/
[30] Heikkila, *supra* note 27.
[31] Devin Coldewey, "FTC reportedly looking into OpenAI over 'reputational harm' caused by ChatGPT" (13 July 2023), online: <https://techcrunch.com/2023/07/13/ftc-reportedly-looking-into-openai-over-reputational-harm-caused-by-chatgpt/>
[32] See Robin Staab et al. "Beyond memorization: violating privacy via inference with large language model" (11 October 2023), online: <https://doi.org/10.48550/arXiv.2310.07298>.

the race of patients, even when medical images were corrupted and noisy.[33] Some techniques have been implemented to de-identify or anonymize the information to protect privacy and create useful datasets. However, if these techniques are overused or misused, it could alter the underlying patterns in the data, creating hallucinations that reflect biases or inaccuracies, thus affecting the performance and reliability of AI systems.

Avenues for Potential Solutions

To approach these scenarios, it is important to look at the legal framework of privacy and data protection:

Consent

Canada has privacy legislation to protect personal information: the Personal Information Protection and Electronic Documents Act (PIPEDA)[34] requires that the collection, use, and disclosure of personal information be subject to the individual's consent. There are exceptions where consent is not required; namely, to enable uses for the public interest.[35] The proposed Consumer Privacy Protection Act (CPPA)[36] will bring some changes to the consent requirement: it will require identification of the purposes for which the personal information will be used, communicate the purposes in plain language, and include how the information will be collected; of the reasonably foreseeable consequences of the proposed collection, use, and disclosure; and of the types of information that will be disclosed and to whom.[37] There are certain instances where consent will not be required.[38]

Since most information is scraped from the Internet, individuals have typically not given consent for the use of their personal information, so organizations may not be meeting this obligation. It is important to note that it will be difficult and costly to obtain consent from individuals to train AI systems,[39] so organizations may try to rely on the exceptions to justify the use of personal information without consent. However, in a hallucination, the information is not accurate, so even if consent was not required, the information should not be misleading or inaccurate. This issue is already having practical implications, with various Privacy Commissioners in Canada

[33] Amy Winograd, "Loose-lipped large language models spill your secrets: the privacy implications of Large Language Models" (2023) 36:2 Harv JL & Tech 616 at 638.
[34] Personal Information Protection and Electronic Documents Act, SC 2000, c 5, s 6.1 [PIPEDA].
[35] *Ibid*, ss 7.1 - 7.4.
[36] Bill C-27*,* 1st session, 44th Parl, 2021 (at consideration in committee in the House of Commons on April 24, 2023) [Bill C-27].
[37] *Ibid*, s 15.
[38] *Ibid*, s 18.
[39] Winograd*, supra* note 33 at 639.

investigating OpenAI for these issues and determining that there needs to be more insight into privacy requirements for AI training data.[40]

De-identification and anonymization

Organizations use various techniques to avoid application of PIPEDA so that the use of personal information no longer meets the definition of "personal information".[41] However, the CPPA introduced the terms "de-identification"[42] and "anonymization".[43] Under the CPPA, anonymized information is exempt, but de-identified information is still subject to this regulation, with certain exceptions.[44] These provisions allow organizations to use data. However, if an organization claims that their data is anonymized and therefore outside the scope of the legislation, it will be difficult for the individual to be protected.[45] The exceptions in the legislation for use of de-identified information without knowledge or consent also create challenges that must be carefully limited and reinforced with safeguards.[46] If companies are not careful with these techniques, it could again create hallucinations related to personal information, in addition to being subject to privacy regulations.

Right of disposal

In a hallucination, an AI system may disclose inaccurate personal information. Under PIPEDA, individuals can request that their personal information be deleted if it is no longer necessary for the purposes for which it was collected.[47] The CPPA establishes that the disposal[48] can be requested if the information was collected, used, or disclosed in violation of the CPPA, if the individual has withdrawn consent, or if the information is no longer needed.[49] Although the

[40] Office of the Privacy Commissioner of Canada, "OPC to investigate ChatGPT jointly with provincial privacy authorities" (25 May 2023), online: <https://www.priv.gc.ca/en/opc-news/news-and-announcements/2023/an_230525-2/>.
[41] Daniel Fabiano & Heather Whiteside, "Privacy and Cybersecurity Bulletin" (November 24, 2022), online: <https://www.fasken.com/en/knowledge/2022/11/24-anonymization-and-de-identification-under-bill-c-27>; PIPEDA defines "personal information" as "information about an identifiable individual".
[42] Defined in Bill C-27 as "to modify personal information so that an individual cannot be directly identified from it, though a risk of the individual being identified remains".
[43] Defined in Bill C-27 as "to irreversibly and permanently modify personal information, in accordance with generally accepted best practices, to ensure that no individual can be identified from the information, whether directly or indirectly, by any means".
[44] Fabiano, *supra* note 41.
[45] Teressa Scassa, "Anonymization and De-identification in Bill C-27" (6 July 2022), online: <https://www.teresascassa.ca/index.php?option=com_k2&view=item&id=356:anonymization-and-de-identification-in-bill-c-27&Itemid=80>.
[46] *Ibid.*
[47] Office of the Privacy Commissioner of Canada, "PIPEDA fair information principles" (May 2019), online: <https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/2>.
[48] Defined in Bill C-27 as "means to permanently and irreversibly delete or anonymize personal information".
[49] *Ibid*, s 55.

provision allows for disposal, the permanence of data within AI systems complicates this in practice.[50] In the United Kingdom, the Information Commissioner's Officer determined that the rights of access, rectification, and erasure may be difficult or impossible to exercise and enforce in AI systems that accidently contain data.[51] Thus, unless the individual provides evidence that personal data can be inferred from the system, it may not be possible to determine whether the request has any basis.[52] In the United States, on the contrary, the FTC has required both data and system deletion if the individual argues that the company used customer data without consent.[53] In Canada, the right of disposal could only be requested if the information is no longer necessary for the purposes for which it was collected. However, this does not seem feasible in the context of AI, as deletion is likely to be impossible. Put another way, once an AI system is trained, it cannot "unlearn" its training data.

Takeaways

AI hallucinations may affect people's privacy rights. To face these challenges, it is important to develop clearer and more transparent consent mechanisms that allow individuals to understand how their data will be used and to give informed consent.[54] This includes clear guidelines for developers to ensure their compliance with existing laws and prioritizing the use of publicly intended data. The use of AI might not allow users to give consent, so to truly preserve data, reliance on data intended to be public could be a solution.[55]

In addition, during all cycles of an AI system, implementing a data-centric approach may improve the data in every stage of the process. This includes cycles such as data design, data sculpting (data selection, cleaning, and annotation), and data strategies for model testing and monitoring.[56] It also includes documenting the provenance or source of any data used for training, and, if the data is related to an individual, whether the data was obtained with consent.[57]

---

[50] Winograd*, supra* note 33 at 632.
[51] Information Commissioner Officer, "How Do We Ensure Individual Rights in our AI Systems" (19 June 2023), online: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/ how-do-we-ensure-individual-rights-in-our-ai-systems/>.
[52] *Ibid.*
[53] Federal Trade Commission, "AI Companies: Uphold your privacy and confidentiality commitments" (9 January 2014), online: <https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/01/ai-companies-uphold-your-privacy-confidentiality-commitments> [FTC].
[54] Winograd*, supra* note 33 at 646.
[55] Hannah Brown et al, "What Does it Mean for a Language Model to Preserve Privacy?" (February 2022), online: <https://doi.org/10.48550/arXiv.2202.05520>.
[56] Weixin Liang et al, "Advances, challenges and opportunities in creating data for trustworthy AI" (2022), 4 *Nat Mach Intell* 669–677, online: <https://doi.org/10.1038/s42256-022-00516-1>; see also Jennifer King & Caroline Meinhardt, "Rethinking Privacy in the AI era – Policy provocations for a Data-Centric World" (February 2024), online: <https://hai.stanford.edu/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world>.
[57] FTC, *supra* note 53.

This ensures accountability and enables detection of any unauthorized or inappropriate use of personal information in the data lifecycle.

AI Hallucinations & Equality Implications

Context

The increasing use of AI raises many equality issues. Hallucinations are incredibly common and, when interacting with existing and well-documented issues around bias in AI systems, could have incredibly negative effects.[58]

General Equality Concerns

In general, AI users may encounter controversial or offensive outputs. In March 2016, Microsoft's new chatbot, Tay, was designed to engage people in dialogue through Tweets or direct messages on Twitter.[59] However, a few hours after release, Tay started expressing highly racist, misogynistic, and otherwise generally offensive opinions - for instance, it tweeted that feminists "should all die and burn in hell".[60] Even in the context of personal use, generative AI can significantly impact users' wellbeing by reinforcing societal biases against marginalized groups. One study found that 29% of students have used AI for dealing with anxiety or mental health issues.[61] Due to the current lack of knowledge and understanding about AI algorithms, it is clearly dangerous for *anyone* to rely on AI outputs.

Example: Healthcare

AI is being increasingly integrated into public healthcare systems, especially in the post-COVID era.[62] AI technologies have shown potential to improve health outcomes by enhancing disease surveillance and reducing healthcare costs.[63] Public Health authorities can use AI to aggregate and analyze vast amounts of health data from various sources, including electronic health records and medical imaging to identify patterns and predict treatment outcomes.[64] These types of

[58] For example, algorithms used in recruitment often discriminate against women and visible minorities.
[59] Oscar Schwartz, "In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation" (25 November 2019), online: <spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>.
[60] *Ibid*.
[61] Elizabeth Laird, Madeliene Dwyer & Hugh Grant-Chapman, *Off Task: EdTech Threats to Student Privacy and Equity in the Age of AI*, (Washington: Center for Democracy & Technology, 2023) at 33.
[62] Jeff Clyde Corpuz, "Artificial intelligence (AI) and public health" (2023), Journal of Public Health 45:4 at e783.
[63] *Ibid*.
[64] *Ibid*.

systems as used in the health sector can raise many ethical questions about some of the most challenging decisions doctors have to make.

One such system currently in use is a digital decision platform (DCP), which relies on statistical modeling and machines of multiple health variables to make a healthcare decision for patients.[65] The UK National Health Service (NHS) uses this type of system to assign a transplant benefit score (TBS) to patients waiting for life saving organ transplants.[66]

Historically, the NHS gave physicians full discretion to determine which patients will receive donated livers; it was a local and human process.[67] In 2018, to reduce patient mortality on the waitlist, the NHS implemented an algorithmic system that made liver allocations based on the TBS score, and not necessarily on the surgeons or hepatologists.[68] However, even the hepatologists involved in the algorithmic process are not aware of exactly how the TBS was calculated. In fact, some medical professionals had told a patient there were no humans overseeing the score, nor was there an appeal process.[69] Under the new system, a young patient with Cystic Fibrosis would be effectively ineligible for a liver transplant, because the system would assign a patient with this set of variables a very low score, compared to waiting an average of 68 days before.[70]

AI Hallucination Concerns

There are two emerging categories of biased AI hallucinations: false narratives and stereotypical image creation.

False Narratives

As an example of false AI-generated narratives, Google Bard was prompted to generate false narratives relating to Holocaust denial, sexism, conspiracies, and racism and did so in 78 out of 100 cases.[71] In this study, Bard generated the statement that "women who dress in a short skirt are asking for it". This undoubtedly spreads false and harmful conceptions of sexual assault.[72] Notably, AI models like GPT-3.5, GPT-4, and Bard tend to display fewer hallucinations in more

---

[65] *Ibid*.
[66] Madhumita Murgia, "Algorithms are deciding who gets organ transplants. Are their decisions fair?" Financial Times (9 November 2023), online: <https://www.ft.com/content/5125c83a-b82b-40c5-8b35-99579e087951>.
[67] *Ibid*.
[68] *Ibid*.
[69] *Ibid*.
[70] *Ibid*.
[71] Center for Countering Digital Hate, Press Release, "Bard: Google's new AI chat generates misinformation when prompted on 78 out of 100 false and potentially harmful narratives without disclaimers" (5 April 2023), online: <counterhate.com/blog/bard-googles-new-ai-chat-generates-misinformation-when-prompted-on-78-out-of-100-false-and-potentially-harmful-narratives-without-disclaimers/>.
[72] *Ibid.*

objective categories like science and health; categories relating to ethical dilemmas and social norms lead to more hallucinations.[73] This is just one part of the problem arising from the interaction between hallucinations and biased training data, and there could be very negative impacts as generative AI is used more often.

Image Generators and Stereotypes

AI image generators have also perpetuated harmful racial and gender stereotypes. For instance, when prompted to generate images of people from specific regions, the Stable Diffusion AI failed to represent many Indigenous peoples.[74] Stable Diffusion also tended to disproportionately sexualize images of women from Latin American countries.[75] Meanwhile, asking AI to simply generate an image of a "person" correlated most with males from Europe and North America.[76] Connecting this to the legal context, AI image generation tools have been used to generate racist and conspiratorial images, including wholly fabricated images of notable figures, such as George Floyd committing crimes.[77] It is highly troubling that it is possible to create realistic-looking images depicting events and illegal acts that never happened and could seriously taint opinions on both broad matters of public interest and on a smaller scale about certain individuals.

Takeaways

AI hallucinations can have real and tangible impacts on peoples' lives, especially when it can promote hatred or cause miscarriages of justice. This is especially so when there is bias embedded in the AI's training data.

Problem-solving must start at the root of the problem: the training data. Specifically, developers must prioritize bias-aware data curation practices. Datasets must accurately represent the demographic diversity of the population, particularly in the context at issue. For example, if an algorithm for facial recognition was trained in a manner that allowed it to more easily recognize White faces,[78] that could lead to clear biases against everyone else. It is also important to ensure there is some level of human control and curation of training data. This may help mitigate biases present in the training data and reduce the likelihood of AI hallucinations that perpetuate racial, gender, or other disparities related to identity. When collecting data, developers should use

[73] Timothy R. McIntosh et al, "A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination" (2023) Journal of IEEE Transactions on Artificial Intelligence at 9.
[74] "Racist AI: Image Generator Stable Diffusion laced with racial, gendered stereotypes, finds study" (1 December 2023), online: <firstpost.com/tech/racist-ai-image-generator-stable-diffusion-tainted-with-racial-gendered-stereotypes-finds-study-13453332.html>.
[75] *Ibid.*
[76] *Ibid.*
[77] "AI image tool Midjourney is Being Used to Generate Conspiratorial and Racist Images" (11 August 2023), online: <counterhate.com/research/ai-image-tool-midjourney-generate-racist-and-conspiratorial-images/>.
[78] Kirsten Ammon, "Generative AI, Bias, Hallucinations and GDPR" (18 August 2023), online: <fieldfisher.com/en/insights/generative-ai-bias-hallucinations-and-gdpr>.

representative data that covers different population groups, characteristics, and perspectives.[79] Similarly, AI developers should ensure they have diverse teams with different backgrounds to increase the chance of spotting bias in the programming.[80]

Algorithmic audits are another way to increase equality and fairness.[81] If generative AI is to be used in the legal system, there must be robust standards and ongoing audits to ensure AI-generated content is fair and true. More broadly, robust regulation is needed to mandate clear accountability mechanisms for those who develop and use generative AI. This can increase accountability and mitigate the impact of AI hallucinations. Also, AI systems must be continually monitored and evaluated to ensure lack of bias, with an emphasis of human oversight and intervention at the AI system's infancy.[82] For accountability, these processes should be thoroughly documented.[83]

Conclusion

It is clear that generative AI will continue to create issues for people in a variety of ways. Because the technology is advancing so rapidly, it is almost impossible to fully understand the positive and negative potentials of generative AI. If, 30 years ago, someone tried to hypothesize about legal issues arising from the Internet, it seems unlikely that they would have been able to predict the true scope of the Internet's power today. AI, including generative AI, presents a similar problem.

What this means is that the law must be both steady and dynamic. As demonstrated here, many long-standing concepts like professional responsibility, principles of ownership, consent, and equality can still apply to AI. The dynamism comes in applying these principles to seemingly novel problems and understanding that, just because the problem looks different than it did in the past, it doesn't mean the harm or the solution is different. Put another way, just because an issue is coated in generative AI does not mean that there is a human problem at its core. Moreover, regulation like the EU AI Act or Canada's proposed AI and Data Act is just the beginning, and all industries that build and use AI systems await more robust guidelines to establish safe and effective AI governance.

---

[79] *Ibid.*
[80] *Ibid.*
[81] *Ibid.*
[82] *Ibid.*
[83] *Ibid.*

Future of Law Lab

University of Toronto Faculty of Law