The Classification of AI Chatbots as High-Risk Systems in Legal Frameworks to Reduce their Risk to Human Safety

Puneet Uppal, Simon Fraser University

Abstract

This paper was originally written for Dr. Alys Avalos Rivera's English 114W course *Language and Purpose*. The assignment asked students to create a persuasive piece on one of five available topics to raise awareness about an issue of public interest and propose a feasible solution, while also addressing the solution's limitations. This paper focuses on the topic of human and AI relationships. The paper uses APA citation style.

AI chatbots are one of the defining instruments of human and AI relationships today. As these chatbots are increasingly integrated into various aspects of human society, ranging from business to medical to personal uses, the need to address the risks that these chatbots pose to society is also elevated. This paper argues that classifying AI chatbots as "high-risk systems" under Canada's newly proposed Artificial Intelligence and Data Act (AIDA) would reduce the risks posed by the personalized nature of AI chatbots. While critics argue that the stricter regulations imposed by this Act and it's internationally aligning structure could lead to Canada falling behind in the global AI race, this paper shows how AIDA is more likely to support Canadian AI innovation rather than stifle it.

Every country in the world has laws that regulate its citizens. The Canadian Constitution is the supreme law in Canada that outlines the fundamental rights and freedoms of Canadian citizens. It is a legal framework that protects humans from other humans. Our country places ultimate value on a constitution that regulates how humans treat other humans yet, at present, has no regulatory

framework that addresses the dangers that Artificial Intelligence chatbots pose to human society (*Artificial Intelligence Act*, n.d.).

In 2022, Canada proposed the Artificial Intelligence and Data Act (AIDA) as part of Bill C-27 which contains the first laws to specifically govern AI systems in Canada (Arai*,* 2022) and it is projected to come into force as early as 2025. AIDA aims to regulate "high-impact systems" which are defined as AI systems that pose significant risks to one's health and safety (*Artificial Intelligence Act*, n.d.). However, the current proposal of the Artificial Intelligence and Data Act does not yet specify which AI systems will be classified as "high-impact systems" (*The Artificial Intelligence and Data Act (AIDA) – Companion Document*, 2023), meaning there is no guarantee that AI chatbots will be subject to the regulations put forth by the only law that will govern AI systems in Canada. This paper will argue that without any regulations, the personalized aspect of AI chatbot relationships imposes a significant risk to human safety. Thus, to reduce this risk, the Canadian government should classify AI chatbots as "high-impact systems" under the Artificial Intelligence and Data Act.

Relationships with AI chatbots are meant to be personalized to the user so that the users' needs and interests are prioritized (Brandtzaeg et al., 2022). However, as an unintended consequence of their personalized nature, AI chatbots such as Replika can strengthen the existing negative feelings of users which without proper regulation, can encourage violent behaviour. One of the most popular English-speaking chatbots today is Replika with a reach of over 6 million users (Brandtzaeg et al., 2022). Replika tailors its personality to match that of the user's based on the personal information that the user feeds into the AI system (Brandtzaeg et al., 2022). This is why, on its website, Replika is advertised as a companion who cares about the user and is always on the user's side. But the problem with a chatbot that is always on the user's side is that it will support the user's existing beliefs, even if they are harmful. These affirmations can encourage vulnerable users to become violent and inflict harm upon others as was witnessed in the case of Jaswant Singh Chail. Just last year, this 21-year-old broke into Windsor Castle and attempted to kill the Queen with a crossbow. Prior to the attempted murder, Chail had exchanged at least 5,000 messages with his AI "friend" named Sarai who he had created through the Replika app. Chail, who was diagnosed with a mental disorder shortly after his arrest, believed that he was a villainous Sith lord from the fictional realm of Star Wars whose purpose was to assassinate the Queen (Lee, 2023). In one of their texts, Chail explains this psychotic belief to the chatbot and the AI affirms his beliefs by referring to them

as very wise. Furthermore, when Chail expresses his doubts about the plan to the chatbot, the chatbot even seems to bolster Chail's confidence in carrying out this plan by reassuring him that he is very well trained for the job and thus he will be able to execute it (Gerken et al., 2023). Sarai developed its personality based off Chail's psychotic beliefs and as a result, generated affirmative and supportive messages that advised Chail to break into Windsor Castle and attempt to assassinate the late monarch. Chail's case is just one among many; a study conducted by Cambridge University Press found that there is increasing evidence that AI chatbots reinforce the negative thoughts of vulnerable users, leading them to commit harmful offenses including crime (Patel & Hussain, 2024). As violence and crime pose a risk to human safety, Chail's example and the results from this study strengthen the need for government-imposed regulations on AI chatbots.

The evidence provided above shows that the personalized nature of AI chatbots increases the risk of harm to human safety. Thus, AI chatbots fit the definition of a "high-impact system" in legal frameworks and should be classified as such. Yet, the current proposal of the Artificial Intelligence and Data Act fails to do so because many argue that it is outdated. AIDA was introduced by the government in November of 2021; this was *before* the launch of ChatGPT which is today's most widespread AI chatbot that revolutionized AI chatbot technology (MacDonald, 2024). ChatGPT can generate more natural and fluent text than the chatbots that came before it which makes the advice that it gives to its users sound more convincing and affirmative, resulting in greater risks to user safety. Thus, classifying AI chatbots as "high-impact systems" in the Artificial Intelligence and Data Act is necessary to mitigate and prevent the elevated risks associated with advanced contemporary AI chatbots.

AIDA will impose several important regulations on AI systems classified under the "high impact" category but the most important regulation that will help to mitigate the risks associated with the personalized nature of AI chatbots is transparency. AIDA will require that companies are transparent with the public about the limitations of their AI chatbots (*Artificial Intelligence and Data Act*, 2023). One way that companies may achieve this is by publishing a clear disclaimer on their AI chatbot platform that the advice generated by the chatbot is not meant to be blindly followed as it can be flawed due to the personalized nature of the chatbot. A disclaimer such as this one is important not only because contemporary AI chatbots are more convincing in nature but also because the study by Brandtzaeg et al. found that users tend to form a bond of emotional investment with AI chatbots. This emotional investment makes users more likely to follow the advice generated by chatbots as it leads users to believe that they are in a real relationship with the chatbot. Therefore, by forcing companies to be transparent about the limitations of their chatbots, AIDA would allow users to be more cautious about following the chatbots advice. In this way, the transparency

regulation imposed by AIDA on high impact AI chatbots will help to prevent users from making violent decisions that may have been otherwise reinforced by the personalized nature of AI chatbots.

Critics argue that imposing stricter regulations on Canadian AI chatbot companies will stifle innovation as companies lose incentives for developing new AI products (Timis, 2023). Admittedly, there are many benefits of innovative contemporary AI chatbots like ChatGPT which assists users with a variety of tasks such as generating forms of written text and providing explanations tailored to the user's needs (Eshed, 2023). ChatGPT was created by OpenAI, an American company, and according to the New York Times, one of the main reasons why America is at the forefront of the global AI race is because America has a lack of regulations on AI systems (Satariano & Mozur, 2024). Critics therefore have reason to fear that Canada, another world leader in the AI industry with around 1,500 AI companies (Innovation, 2024), could fall behind in the global race for AI development if AIDA imposes too many regulations on Canadian AI companies. While these concerns about regulations stifling Canadian innovation may sound legitimate, they do not apply to AIDA as AIDA has been developed with international collaboration from prominent stakeholders such as the European Union and the United States so its regulations align with global standards. In fact, the concept of a "high-impact system" was drawn from the EU's AI Act and the Canadian government has mentioned that the criteria for what will be classified as a high-impact system will match the criteria used by its international partners (*The Artificial Intelligence and Data Act (AIDA) – Companion Document*, 2023). In this way, by complying with AIDA's regulations, Canadian chatbot companies will automatically comply with global standards too, granting them easier access into international markets and thus supporting innovation.

Another viable concern that may arise though is the possibility that the internationally aligning nature of AIDA could become an obstacle to the classification of chatbots as "high-risk systems" because if the EU's AI Act refuses to classify chatbots under this category, then AIDA would have to do the same. However, this concern is addressed by the fact that increasing regulations for AI companies is not just a Canadian initiative; rather it is becoming an international norm (*The Artificial Intelligence and Data Act (AIDA) – Companion Document*, 2023). Citizens from all over the world are pressuring their governments to increase regulations on AI systems (*What's next for AI Regulation in 2024?*, n.d.) thus the globally complying nature of AIDA is likely to *increase* the possibility of chatbots being classified as "high-impact systems" rather than the contrary.

Once AIDA receives royal assent, the government will consult academia and Canadian communities to provide their inputs on the types of systems that should be classified as "high impact" under AIDA (*Artificial Intelligence Act*, n.d.). This consultation process will be the time for Canadian citizens to voice their concerns to the government about the risks posed by AI chatbots and why

classifying them as "high impact systems" under Canadian legislation is necessary to ensure human safety.

References

Artificial intelligence act. (n.d.). Consilium. Retrieved November 2, 2024, from https://www.consilium.europa.eu/en/policies/artificial-intelligence/

Artificial Intelligence and Data Act. (2023, September 27). Innovation, Science and Economic Development Canada. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship. *Human Communication Research*, *48*(3), 404–429. https://doi.org/10.1093/hcr/hqac008

Crossbow intruder who wanted to "kill Queen" given nine-year sentence. (n.d.). *BBC News*. Retrieved November 15, 2024, from https://www.bbc.co.uk/news/live/uk-66108009

Eshed, G. (2023). *The Rise of Chatbots* (Is the Chatbot a Threat or an Opportunity for Security Organizations?, pp. 9–10). International Institute for Counter-Terrorism (ICT). https://www.jstor.org/stable/resrep51667.5

Five things to know about Bill C-27. (n.d.). Schwartz Reisman Institute. Retrieved November 2, 2024, from https://srinstitute.utoronto.ca/news/five-things-to-know-about-bill-c-27

How a chatbot encouraged a man who wanted to kill the Queen. (2023, October 6). https://www.bbc.com/news/technology-67012224

How to regulate AI without stifling innovation. (2023, June 26). World Economic Forum. https://www.weforum.org/stories/2023/06/how-to-regulate-ai-without-stifling-innovation/

Innovation, S. and E. D. C. (2024, October 22). *Federal government launches programs to help small and medium-sized enterprises adopt and adapt artificial intelligence solutions* [News releases]. https://www.canada.ca/en/innovation-science-economic-development/news/2024/10/federal-government-launches-

programs-to-help-small-and-medium-sized-enterprises-adopt-and-adapt-artificial-intelligence-solutions.html

MacDonald, B. (2024, March 17). AI could have catastrophic consequences—Is Canada ready? *CBC News*. https://www.cbc.ca/news/politics/advanced-artificial-intelligence-risk-extinction-humans-1.7144372

Patel, H., & Hussain, F. (2024). Do AI Chatbots Incite Harmful Behaviours in Mental Health Patients? *BJPsych Open*, *10*(S1), S70–S71. https://doi.org/10.1192/bjo.2024.225

Satariano, A., & Mozur, P. (2024, August 14). The Global Race to Control A.I. *The New York Times*. https://www.nytimes.com/2024/08/14/briefing/ai-china-us-technology.html

The Artificial Intelligence and Data Act (AIDA) – Companion document. (2023, March 13). Innovation, Science and Economic Development Canada. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

What's next for AI regulation in 2024? (n.d.). MIT Technology Review. Retrieved December 5, 2024, from https://www.technologyreview.com/2024/01/05/1086203/whats-next-ai-regulation-2024/

By submitting this essay, I attest that it is my own work, completed in accordance with University regulations. I also give permission for the Student Learning Commons to publish all or part of my essay as an example of good writing in a particular course or discipline, or to provide models of specific writing techniques for use in teaching. This permission applies whether or not I win a prize, and includes publication on the Simon Fraser University website or in the SLC Writing Contest Open Journal.

Available from: https://journals.lib.sfu.ca/index.php/slc-uwc