AI & SOCIETY
https://doi.org/10.1007/s00146-024-02096-7

CURMUDGEON CORNER

Air Canada's chatbot illustrates persistent agency and responsibility gap problems for AI

Joshua L. M. Brand[1]

Received: 29 July 2024 / Accepted: 1 October 2024

Valentine's Day was less rosy than usual this year for Air Canada after being found liable for its ill-informed AI-powered chatbot and ordered to pay compensation to a passenger in a Civil Resolution Tribunal in the Small Claims Court of British Columbia, Canada (2024 BCCRT 149). While the ruling was in favour of the customer, Air Canada's defence highlights how companies continue to attempt to absolve themselves of accountability in AI deployment by exploiting the responsibility gap and dubious metaphysical arguments.

Following the death of his grandmother in November 2022, Jake Moffatt searched for a last-minute flight to Toronto from Vancouver on Air Canada's website. While the original fare was CAN$1,630.36, Moffatt understood that as is the case with most airlines, bereavement fare discounts are offered when travelling in the event of an immediate family member's death. Moffatt interacted with the airline's built-in chatbot feature to determine his eligibility for the discounted fare. The chatbot instructed him:

> *If you need to travel immediately or have already travelled and would like to submit your ticket for a reduced bereavement rate, kindly do so within 90 days of the date your ticket was issued by completing our Ticket Refund Application form.*

This information provided by the chatbot was inconsistent with the airline's website, it *hallucinated* the policy—this being the common term used in the media to describe this instance as well as generally when other advanced chatbots, such as ChatGPT and Claude, generate false information. After completing travel and calling Air Canada to request the expected refund, Moffatt was told that he did not follow the correct procedure—requesting the discounted fare *before* travel—which is described on another part of the Air Canada website. Air Canada acknowledged the chatbot provided misleading information and offered Moffatt a CAN$200 voucher as a gesture of goodwill.

Understanding the fare difference was much greater, Moffatt refused the voucher and sued Air Canada in small claims court for negligent misrepresentation, arguing they owed a duty of care. In their defence, Air Canada distanced themselves from the chatbot with a questionable metaphysical argument: the chatbot is a separate legal entity akin to one of their "agents, servants, or representative" and they are therefore not responsible for its actions.

Tribunal member Christopher Rivers found this defence absurd:

> *This is a remarkable submission. While a chatbot has an interactive component, it is still just a part of Air Canada's website. It should be obvious to Air Canada that it is responsible for all the information on its website. It makes no difference whether the information comes from a static page or a chatbot.*

No matter how autonomous, interactive, or agent-like the chatbot presents, it is part of the Air Canada website like any other webpage they publish. The Tribunal therefore found that Air Canada broke the duty of care they owed Moffatt by negligently misrepresenting information on their chatbot tool. In this seemingly open-and-shut case, Moffatt was awarded CAN$812.02 in damages to cover court fees and the bereavement fare difference.

Before addressing the significance of this case, it is important to clarify that while it is often said that the chatbot *hallucinated*, it is a misnomer. While the precise technology used by Air Canada is not known, because the chatbot generated a nonexistent policy, it is fair to assume it is based on a large language model (LLM) that generates responses after being trained on various sources of information. This is opposed to retrieval-based chatbots that do not diverge from predefined responses. But hallucination refers to a perceptual experience—someone is perceiving something

✉ Joshua L. M. Brand
joshua.brand@telecom-paris.fr

[1] i3, CNRS, Télécom Paris, Institut Polytechnique de Paris, Paris, France

Published online: 23 October 2024

that is not really there—and LLMs clearly do not have this cognitive capacity. They cannot perceive, therefore they cannot misperceive. Moreover, hallucination implies a deviation from the norm, yet all that LLMs do is string together text through probability calculations which are not designed to convey accurate information but to produce convincing human-like text. Whether the information is ultimately true or false, the process to produce these outputs is the same. For these reasons, following the Harry Frankfurt-inspired work by Michael Townsen Hicks et al., it is more appropriate to call these outputs *bullshit* because the models lack attention to the veracity of their outputs (Hicks et al. 2024). This certainly fits better with the Tribunal's ruling of negligence.

This case has both legal and moral significance. A clear legal consequence of this decision centres around trustworthiness and accuracy of information. Rivers emphasized that Air Canada failed to demonstrate why their webpage, "Bereavement travel", containing the correct policy, is more trustworthy than the information provided by the chatbot. The onus is therefore not on the customer to double-check information; any text provided across customer interfaces, including a chatbot, ought to be deemed as equally valid information. The company's legal obligations to provide accurate information remain whether it is posted on a static page or chat box.

The decision also reminds us of two moral concerns. First is the *responsibility gap*, initially tackled by Andreas Matthias, that challenges our legal and moral responsibility traditions with the following question: *when someone is harmed by AI, who is responsible?* As Mark Coeckelbergh tells us, the difficulty of attributing responsibility in the context of AI is due to the presence of both "many hands", i.e. the many people who work on the software and its implementation, and "many things", i.e. that there are many different technologies involved. In the case of the Arizonan woman who was killed by a self-driving Uber in 2018, as Coeckelbergh uses as his example, there were many involved in the car's deployment and action: Uber, the car company, the car user, and the regulators (Coeckelbergh 2020). Who is responsible in this case? The application of AI systems makes it easy to distort any clear path of accountability when harm is caused.

Moreover, any insistence that the chatbot hallucinated the policy further amplifies the potential for disowning the bot. Convincing others that the output was a hallucination or deviation from the norm, provides its developers and owners with more leeway to label it as an anomaly, rather than as an issue with the system altogether.

One may be under the optimistic impression that given the many harms caused by various forms of AI in recent years, like the Arizona case, companies would be increasingly more cautious with implementing AI, ensuring that its deployment is surrounded by safeguards that take into account the effects and harms it may have on users. The Air Canada defence, however, reminds us of the ever-present risk and attempt to exploit the responsibility gap present in AI use.

The second concern flows from this point as Air Canada did not attempt to exploit the responsibility gap by simply passing the buck to the developers of the chatbot, causing a long, arduous process (at the expense of the plaintiff) to decide who ultimately is responsible, but instead attributed full moral agency to the chatbot by mentioning it alongside their "agents, servants, or representatives", arguing these agents are responsible for their own actions, to absolve the airline's responsibility. This misrepresents the definition of moral agency.

While the interactive chatbot may emulate a human and provide ethically charged information like a customer service agent, such as offering financial advice on ticket purchasing, it nevertheless is a machine and not a moral agent. At most, its agency remains a mere suggestion—it may look like an agent and perform tasks with ethical consequences, it nevertheless lacks the necessary attributes to render it a fully independent agent capable of moral responsibility. Traditionally, these attributes require not only the autonomous control of one's actions but also the conscious appreciation of the reasons and principles underlying their actions. In other words, moral agents can understand, reflect on, and evaluate their actions; they can set their own rules and principles upon which they base their actions and to which we can eventually refer to hold them accountable.

This is why we usually do not subject animals to normative standards, or see them as responsible for their actions because they do not have these advanced rational capabilities required for *moral* agency. The same goes for children or those with mental impairments whose rational capacity is not fully realized. They have autonomous capabilities and can follow rules set by their guardians, but they generally cannot articulate the reasons and principles behind their actions in the same way adults can. As Sven Nyholm reminds us, this is why we hold parents and guardians accountable for their actions under a form of supervised agency (Nyholm 2018). It is only when children gain the necessary rational capacity that moral agency and responsibility is wholly transferred to them and we no longer hold their parents accountable for their actions.

At best, we could therefore understand the agency of a chatbot like a child, where it acts somewhat autonomously by providing information to customers, but is ultimately in a supervised endeavour with its parent, Air Canada, who ought to initiate, supervise, and manage its work. Air Canada, like any parent, remains the responsible party. It should be noted, however, that children are nevertheless more independent

and have a wider range of rational capacity than AI systems, which (so far) operate in a domain-specific context.

In what seems to be the first case of its kind, the Tribunal agreed that this metaphysical claim is inconsistent. While this does not set a major precedent in Canada given that it occurred in the lower provincial court, it serves as a reminder that companies cannot so easily absolve their legal and moral responsibilities when implementing AI by making unfound claims regarding AI capabilities and agency.

We also see here that chatbots can cause financial harm, therefore presenting legal liabilities to companies, and this is not only for airlines who deal with significant financial transactions but financial institutions as well. The Consumer Financial Protection Bureau (USA) has warned that chatbots are unable to solve complex problems and can provide incorrect information to consumers, which can lead to consumers paying unnecessary 'junk fees'.[1] Even more so that the chatbot market is projected to reach $32.4 billion by 2032,[2] this underscores the importance for companies to take seriously their implementation of AI and ensure the information they provide is accurate, whether it comes from a human representative or automated system.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Funding This research is supported by the XAI4AML chair (No ANR-20-CHIA-0023-01).

Data availability Not applicable.

Declarations

Conflict of interest The author has no competing interests to declare that are relevant to the content of this article.

References

Coeckelbergh M (2020) Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 26:2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Hicks MT, Humphries J, Slater J (2024) ChatGPT is bullshit. Ethics Inf Technol. https://doi.org/10.1007/s10676-024-09775-5

Nyholm S (2018) Attributing agency to automated systems: reflections on human-robot collaborations and responsibility-loci. Sci Eng Ethics 24:1201–1219. https://doi.org/10.1007/s11948-017-9943-x

[1] https://www.consumerfinance.gov/data-research/research-reports/chatbots-in-consumer-finance/chatbots-in-consumer-finance/ (accessed 29 July 2024).

[2] https://finance.yahoo.com/news/chatbot-market-size-reach-usd-235000000.html (accessed 29 July 2024).