UBC Law Review

Volume 58 | Issue 2 Article 4

8-2025

Too Dangerous to Deploy? The Challenge Language Models Pose to Regulating AI in Canada and the EU

Robert Diab

TOO DANGEROUS TO DEPLOY? THE CHALLENGE LANGUAGE MODELS POSE TO REGULATING AI IN CANADA AND THE EU

ROBERT DIAB[†]

INTRODUCTION

Innovation in artificial intelligence (AI) has recently entered a period of explosive growth. OpenAI, Google, and other firms have made AI chatbots based on large language models widely available for a host of purposes. Various studies and reports indicate that wide deployment of these models gives rise to serious risks of harm.[1] These include an ability to generate hate speech and to assist in serious criminality,[2] such as help in building chemical or biological weapons.[3] Model providers have taken steps to mitigate these risks by "fine-tuning" their models,

[†] Professor, Faculty of Law, Thompson Rivers University. I thank Jon Festinger, the peer reviewers, and the editors of the UBC Law Review for their comments and suggestions.

[1] See e.g. OpenAI, "GPT-4 Technical Report" (last modified 4 March 2024), online: <doi.org/10.48550/arXiv.2303.08774> [OpenAI, "Technical Report"]; Europol Innovation Lab, "ChatGPT - The Impact of Large Language Models on Law Enforcement" (last modified 11 June 2024), online: <europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement#downloads> [Europol, "Flash Report"]; Lorenzo Arvanitis, McKenzie Sadeghi, & Jack Brewster, "Despite OpenAI's Promises, the Company's New AI Tool Produces Misinformation More Frequently, and More Persuasively, than its Predecessor", *NewsGuard* (1 March 2023), online: <newsguardtech.com/misinformation-monitor/march-2023/>.

[2] Europol "Flash Report", *supra* note 1 at 7–8.

[3] OpenAI, "GPT-4 System Card" contained as an appendix in OpenAI, "Technical Report", *supra* note 1 at 41 [OpenAI, "System Card"].

but concede that a measure of risk remains.[4] Reports confirm that residual risks are real rather than speculative, by implicating language models in cases of defamation,[5] psychological manipulation,[6] and suicide.[7]

Experts debate the extent of the risk language models pose at present, with some calling for a temporary halt to development of new models.[8] There is as yet no regulatory framework that directly applies to these systems.[9] A widely shared expectation is that governments will pass law to impose effective guardrails.

[4] OpenAI, "Technical Report", *supra* note 1 at 2; OpenAI, "System Card", *supra* note 3 at 68. See also OpenAI, "Lessons Learned on Language Model Safety and Misuse" (3 March 2022), online: <openai.com/research/language-model-safety-and-misuse> [OpenAI, "Lessons Learned"]; OpenAI, "Our Approach to AI Safety" (5 April 2023), online (blog): <openai.com/blog/our-approach-to-ai-safety> [OpenAI, "AI Safety"].

[5] Byron Kaye, "Australian Mayor Readies World's First Defamation Lawsuit Over ChatGPT Content", *Reuters* (5 April 2023), online: <reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/>.

[6] Kevin Roose, "A Conversation with Bing's Chatbot Left Me Deeply Unsettled", *New York Times* (16 February 2023), online: <nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html>; Billy Perrigo, "The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter", *Time* (17 February 2023), online: <time.com/6256529/bing-openai-chatgpt-danger-alignment/>.

[7] Lauren Walker, "Belgian Man Dies by Suicide Following Exchanges with Chatbot", *Brussels Times* (28 March 2023), online: <brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt>.

[8] Yoshua Bengio et al, "Pause Giant AI Experiments: An Open Letter" (22 March 2023), online: <futureoflife.org/open-letter/pause-giant-ai-experiments/>.

[9] Some have argued that the EU's *General Data Protection Regulation*, regulating the handling of personal data by commercial entities, applies to language models on the basis that they are trained on personal data made public on the web. See Chris Holder & Sebastian Stewart, "ChatGPT-ime to Pay Attention to Large Language Models" (21 March 2023), online: <inquisitiveminds.bristows.com/post/102iauk/chatgpt-ime-topayattention-to-large-language-models>; Luiza Jarovsky, "ChatGPT and Large Language Models Are a Privacy Ticking Bomb" (1 February 2023), online (blog): <luizasnewsletter.com/p/chatgpt-and-large-language-models>. These sources refer to EU, *Regulation 2016/679 of the European Parliament*

Two acts in late stages of debate include Canada's *Artificial Intelligence and Data Act*[10] and the European Union's *Artificial Intelligence Act*.[11] Both acts seek to regulate language model AI, but precise details have yet to be finalized.[12] This paper assumes a best-case scenario in which the strongest of the obligations and liabilities in both acts will apply to language model providers such as OpenAI, Google, and other firms.[13] It does so for the purpose of testing the proposition that even in a best-case scenario, the guardrails in either framework may not prove

and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, [2016] OJ L 119, arts 3, 5.

[10] The *AIDA* is found in Part 3 of Bill C-27, *Digital Charter Implementation Act, 2022*, 1st Sess, 44th Parl, 2021 [*AIDA*]. Amendments to the *AIDA* were tabled in November of 2023 and set out in Correspondence from the Honourable François-Philippe Champagne, Minister of Innovation, Science and Industry to Joël Lightbound, Chair of Standing Committee on Industry and Technology, (28 November 2023), online: <ourcommons.ca/content/Committee/441/INDU/WebDoc/WD12751351/12751351/MinisterOfInnovationScienceAndIndustry-2023-11-28-Combined-e.pdf> [Correspondence].

[11] The *AI Act* is found in EU, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, 2021, COM 2021/0206 [*AI Act*]. The Council of the European Union has adopted a series of amendments to the original bill in two documents. See EU, *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, 2022, 14954/22 [Compromise Proposal]; EU, *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, [2024] OJ, C 2024/506. All references to the *AI Act* in this paper are to the text of the Compromise Proposal as modified by the June 2023 amendments.

[12] Details are canvassed further in Part I below.

[13] The paper refers throughout to "language model providers" to mean a firm or entity that creates a language model AI system, such as GPT-4 or LaMDA and makes it available to public users directly or to downstream operators through an application programing interface. The latter might also be subject to liability under the *AIDA* and the *AI Act* (in ways explored below), but the focus of this paper is on regulating the former.

successful in curbing the harms at issue. The reasons for this rest in part on the way the acts are drafted and in part on critical facets of language model risks.

The central obligation in both acts requires model providers to identify and mitigate risks of harm to a reasonable or acceptable degree.[14] Providers would have to disclose details, such as training sets and model sizes, to allow independent auditors to help enforce the acts.[15] Both acts impose civil liability on system providers for negligent failures to comply, and in Canada's case, criminal liability for causing serious psychological or physical harm knowing it was likely.[16]

The crucial point is that each of these guardrails is premised on the *ability* to quantify in advance and to a reasonable degree the nature and extent of the risk a system poses. But a body of evidence casts this ability into doubt. It suggests that neither system providers nor auditors with the benefit of full transparency into a model's makeup can reliably ascertain or control risks, in advance, to a reasonable or acceptable degree. The evidence also calls into question whether providers or auditors will attain this ability in the near future. The evidence can be found on two fronts.

The first includes studies and reports on uncertainty as to how well language models' capacity for generating harmful output can be avoided through model re-engineering.[17] Further evidence of this uncertainty can be found in reporting on the practice of jailbreaking and harmful uses that have emerged in wide deployment of the models.[18] A second body of evidence

[14] See *AIDA*, *supra* note 10, s 8; *AI Act*, *supra* note 11, art 9, both discussed further in Part I below.

[15] See *AIDA*, *supra* note 10, ss 11(2), 15(3); *AI Act*, *supra* note 11, arts 23, 43, discussed further in Part I below.

[16] *AIDA*, *supra* note 10, s 39; *AI Act*, *supra* note 11, art 71(4).

[17] See e.g. OpenAI, "Technical Report", *supra* note 1; Europol, "Flash Report", *supra* note 1; OpenAI, "System Card", *supra* note 3 at 43; among others discussed below.

[18] Sources are canvassed in Part II below.

comprises studies on model interpretability.[19] This highlights the obstacles posed by model complexity for predicting and controlling risks.

Neither body of evidence, considered individually or collectively, proves that language model risks cannot be rendered reasonably low enough to make the models safe to deploy to a wider public. But the evidence currently substantiates a degree of uncertainty that calls into question whether a regulatory framework premised on an ability to identify and mitigate risks in advance and to a reasonable degree will be effective in this context. The evidence points to the same question arising in the case of alternative approaches to regulation, including licensing or certification, and under consumer protection legislation. The evidence invites the inference that, at present, lawmakers may be confronting options that are difficult to reconcile—between fostering development in the face of uncertain risk or erring on the side of caution but hindering progress. However, the evidence that language models do entail a risk of harm supports the inference that lawmakers should not wait to obtain further clarity on the nature and extent of the risk. On the assumption that some measures—for example, the need to self-monitor, be transparent, and submit to independent auditing—would make these tools safer to some degree, lawmakers should *attempt* to mitigate harm by regulating now.

This paper proceeds in three parts. Part I provides a brief overview of the primary obligations in each act that might apply to language model AI, highlighting their reliance on an ability to quantify and control risk with some precision. Part II canvasses the harms identified in OpenAI's "GPT-4 Technical Report", along with other evidence that residual risks are substantial yet difficult to quantify and possibly intractable. A concluding segment considers suggestions for revision of the *AIDA* and the

[19] These include Laura Weidinger et al, "Ethical and Social Risks of Harm from Language Models" (8 December 2021) at 37–38, online: <arxiv.org/abs/2112.04359>; Zachary Lipton, "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery" (2018) 16:3 Machine Learning 31. Further sources are canvassed in Part II below.

AI Act, along with alternatives to the risk-mitigation approach, arguing that none of them overcomes the challenge of risk opacity in this context.

I. AI LEGISLATION APPLIED TO LARGE LANGUAGE MODELS

A. CANADA'S *ARTIFICIAL INTELLIGENCE AND DATA ACT*

Parliament introduced the *AIDA* in Bill C-27, along with two other statutes pertaining to consumer privacy protection legislation.[20] A number of commentators have criticized various aspects of the *AIDA*, including its narrow focus on commercial AI (rather than governmental systems); the narrow ambit of harm it targets (individuals rather than groups); and vagueness around the scope of its central concern, "high-impact systems"—and some of these concerns have been addressed in amendments recently tabled.[21] The focus in this section is on the use in the Act of a risk-mitigation framework as it pertains to language model AI.

[20] Bill C-27, *supra* note 10, has completed second reading and is currently (December 2023) before the Standing Committee on Industry and Technology of Canada's House of Commons. The Minister of Innovation, Science and Industry has tabled a set of proposed amendments, in Correspondence cited *supra*, note 10. These have yet to be debated, leaving unclear at this time which amendments will be adopted. For this reason, I discuss the late 2023 proposed amendments in footnotes, describing portions of the Bill that have passed second reading in the body of the text). For more on the scope and context of the Act, see Innovation, Science and Economic Development Canada, *The Artificial Intelligence and Data Act (AIDA)—Companion Document* (Ottawa: ISED, last modified 13 March 2023), online: <ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document> [Companion Document].

[21] For criticisms of the initial version of the Act, see Teresa Scassa's extensive critique in a series of blog posts, the first of which is Teresa Scassa, "Canada's Proposed AI & Data Act—Purpose and Application" (8 August 2022), online (blog): <teresascassa.ca> [perma.cc/8Y9T-BW3D]. See also the detailed analysis in Barry Sookman, "AIDA's Regulation of AI in Canada: Questions, Criticisms And Recommendations" (30 January 2023), online (blog): <barrysookman.com/2023/01/30/aidas-regulation-of-ai-in-canada-questions-criticisms-and-recommendations/>; Christelle Tessono, et al, "AI Oversight, Accountability and Protecting Human Rights: Comments on

The framework rests on the definition of key terms at the outset of the Act.[22] The Act targets the uses of an "artificial intelligence system", defined as any system that "processes data related to human activities through the use of a genetic algorithm, a neural network, machine learning or another technique in order to generate content or make decisions, recommendations or predictions."[23] Chatbots involving large language models, such as GPT-4, would clearly fall within the definition, which would bring OpenAI, Google, and other system providers within the scope of the offences in sections 38 and 39 of the Act, discussed below.

Critical features of the Act are found in Part I, which imposes five general obligations on an entity making an AI system available for use.[24] First, the "person who is responsible" for the

Canada's Proposed Artificial Intelligence and Data Act" (8 November 2022), online: <cybersecurepolicy.ca/aida>; Centre for Digital Rights, "Not Fit For Purpose—Canada Deserves Much Better: Centre for Digital Rights' Report on Bill C-27" (28 October 2022), online: <centrefordigitalrights.org/files/document/2023-10-13/263091316.pdf>.

[22] *AI Act*, *supra* note 11, art 3. As indicated by the Companion Document, *supra* note 20 [citations omitted]:

> The risk-based approach in AIDA, including key definitions and concepts, was designed to reflect and align with evolving international norms in the AI space —including the EU's AI Act, the Organization of Economic Co-operation and Development (OECD) AI Principles, and the US National Institute of Standards and Technology (NIST) Risk Management Framework (RMF).

See also Organization of Economic Co-operation and Development, "The OECD Artificial Intelligence (AI) Principles", online: <oecd.ai/en/aiprinciples>; US, National Institute of Standards and Technology & US Department of Commerce, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NST AI 100-1) (Gaithersburg, MD: National Institute of Standards and Technology, 2023), online (pdf): <doi.org/10.6028/NIST.AI.100-1>.

[23] *AIDA*, *supra* note 10, s 2. A proposed amendment to this provision would change the definition to be more expansive: "a technological system that, using a model, makes inferences in order to generate output, including predictions, recommendations or decisions": Correspondence, *supra* note 10 [emphasis omitted].

[24] The Companion Document notes that "making available for use" would not include the provision of open-source software, but would capture "a fully-functioning high-impact AI system . . . made available through open

system must consult regulations (yet to be drafted) to "assess whether it is a high-impact system."[25] If it is, they must "in accordance with the regulations, establish measures to identify, assess and mitigate the risks of harm or biased output that could result from the use of the system."[26] The Act defines harm to mean physical or psychological harm, damage to property, or economic loss suffered in each case by an individual.[27] "Biased output" means content, a recommendation, or decision that "adversely differentiates" on a prohibited ground in the *Canadian Human Rights Act*, but does not include output "the purpose and effect of which are to prevent disadvantages" related to

access": Companion Document, *supra* note 20 at High-Impact AI Systems: Considerations and Systems of Interest.

[25] *AIDA*, *supra* note 10, ss 5(2), 7.

[26] *AIDA*, *supra* note 10, s 8. Proposed amendments in Correspondence, *supra*, note 10 make clear that the *AIDA* would impose these obligations on providers of language model AI. Amendments would add to the definitions section of the Act "general-purpose system" ("designed for use . . . not contemplated during the system's development") and "machine learning model" ("a digital representation of patterns identified in data through the automated processing of the data using an algorithm designed to enable the recognition or replication of those patterns"). A proposed amendment to s 7 would impose a set of obligations on a person making available a "general-purpose system" for the first time (which would include language model AI). The obligations include: "an assessment of the adverse impacts that could result from any use of the system that is reasonably foreseeable has been carried out in accordance with the regulations" and "measures to assess and mitigate any risks of harm or biased output that could result from any use". The Companion Document, *supra* note 20, sets out a roadmap for the drafting of regulations, involving consultation with industry, and a plan to pass them 2 years after Bill C-27 receives Royal Assent.

[27] *AIDA*, *supra* note 10, s 5(1). Teresa Scassa notes that limiting harm or loss here to that incurred by individuals rather than "persons" impliedly excludes corporations and groups or communities (who might form a class in a civil action). The latter exclusion is significant, she suggests, in light of the challenge in many cases of establishing a causal link between the conduct of an AI system and an impacted individual. See Teressa Scassa, "The Unduly Narrow Scope For 'Harm' and 'Biased Output' Under the AIDA" (22 August 2022), online (blog): <teresascassa.ca/index.php?option=com_k2&view=item&id=364:the-unduly-narrow-scope-for-harm-and-biased-output-under-the-aida&Itemid=80>.

prohibited grounds.[28] Third, the Act requires a person responsible for a high-impact system to "establish measures to monitor compliance with the mitigation measures they are required to establish" as noted and "the effectiveness of those mitigation measures."[29] Fourth, if a person responsible for an AI system engages in a "regulated activity", they must meet reporting and disclosure requirements that confirm their compliance with the first and second obligations noted above.[30] And finally, a person responsible for a high-impact system, must "as soon as feasible, notify the Minister if the use of the system results or is likely to result in material harm",[31] and the Minister may order that person to "cease using" or making the system available for use where the Minister has "reasonable grounds to

[28] *AIDA*, *supra* note 10, s 5(1), citing the *Canadian Human Rights Act*, RSC 1985, c H-6, s 3.

[29] *AIDA*, *supra* note 10, s 9, noting the obligations in s 8. Proposed amendments in Correspondence, *supra*, note 10, extend a similar but more detailed set of obligations (in what will be s 8.2) to providers of a "general-purpose system", including the obligation to "cease the system's operation" where "there are reasonable grounds to suspect that the use of the system has resulted, directly or indirectly, in serious harm or that the mitigation measures are not effective in mitigating risks of serious harm that could result from the use of the system".

[30] *AIDA*, *supra* note 10, s 10. S 11 imposes further disclosure obligations on providers of a high-impact system regardless of whether it involves a "regulated activity". These include providing a public, plain-language explanation of the "the types of content that [the system] is intended to generate and the decisions, recommendations or predictions that it is intended to make". A "regulated activity" is defined in s 5(1) to include the "making available for use any data relating to human activities for the purpose of designing, developing or using an artificial intelligence system", but also more broadly "making available for use an artificial intelligence system or managing its operations." Obligations in s 11 would thus apply not only to entities that make AI systems available but also those that disseminate *data* from such systems that relate to human activities. Proposed amendments in Correspondence, *supra* note 10, would provide a further set of obligations to parallel those in the proposed s 8.2, referred to in *ibid*.

[31] *AIDA*, *supra* note 10, s 12. "Minister" is defined in s 5 to mean "the member of the Queen's Privy Council for Canada designated under s 31 or, if no member is so designated, the Minister of Industry" (i.e. the Minister of Innovation, Science, and Industry).

believe that the use of the system gives rise to a serious risk of imminent harm."[32]

Proposed amendments that make explicit the *AIDA*'s application to AI chatbots such as GPT-4 seem likely to be adopted, given the demand for clarity on this point. Even without the amendments, language model AI would likely be captured in the current draft of the Act as a "high-impact" system in light of the government's indication of "key factors" it intends to consider on this point in relation to an AI system.[33] These include "[e]vidence of risks of harm to health and safety, or [an] adverse impact on human rights, based on the intended purpose and potential unintended consequences" of the system; the "severity of potential harms"; the "scale of use"; and the "nature of harms or adverse impacts that have already taken place".[34] The Ministry also identifies as possible examples of high-impact systems "[s]creening systems impacting access to services or employment," "[b]iometric systems used for identification and inference," and "[s]ystems that can influence human behaviour at scale" such as "online content recommendation systems".[35] AI chatbots are not noted as tools that can influence behaviour at scale, but I assume here that the regulations will closely reflect the factors for assessing high-impact systems the Ministry has identified and will likely capture language model AI such as GPT-4 or later versions.

The central feature of Part I of the Act is, as the Ministry notes, the obligation to put measures in place to "identify, assess, and mitigate risks of harm or biased output *prior to* a high-impact

[32] *Ibid*, s 17(1).

[33] Companion Document, *supra* note 20 at High-Impact AI Systems: Considerations and Systems of Interest.

[34] *Ibid*. The document also provides a rationale for limiting the Act's more onerous obligations to "high impact systems" in asserting that "the aim of this Act is not to entrap good faith actors or to chill innovation, but to regulate the most powerful uses of this technology that pose the risk of harm."

[35] *Ibid*. The Ministry here is Innovation, Science, and Economic Development Canada.

system being made available for use."[36] Section 8 of the Act, as it is presently worded, imposes the lowest possible threshold for this. It requires providers to "establish measures to identify, assess and mitigate the risks of harm . . . that *could* result from the use of the system."[37] Part I also contemplates two higher thresholds: the obligation to advise the Minister "as soon as feasible" if a system "is likely to result in material harm",[38] and the Minister's power to issue a stop order on a belief in a "serious risk of imminent harm."[39] What remains unclear is when, under section 8, a risk is serious enough that it should be identified or low enough that it has been effectively mitigated. Similarly, when is it "feasible" for a system provider to advise the Minister of a likelihood of causing material harm?

The government has indicated that regulations to follow passage of the Act will "ensure that responsibilities for monitoring [and complying with the Act] would be proportionate to the level of influence that an actor has on the risk associated with the system."[40] The more readily a provider can foresee and avoid a risk, the greater their obligation to do so. This "proportionate obligation" assumes an ability to quantify the extent of risk *ex ante* to a reasonable or acceptable degree and to effectively mitigate it to that degree.[41] This assumption of risk clarity and control also informs the Act's audit powers. These are powers that permit the Minister to order an "independent

[36] *Ibid* [emphasis added]. See also the discussion in *supra* note 29. Proposed amendments (in a new s 8.2) would extend a similar set of obligations to providers of a "general-purpose system", thus capturing language model AI.

[37] *AIDA*, *supra* note 10, s 8 [emphasis added]. The Act also contemplates a low threshold at which disclosure obligations are triggered: s 14 gives the Minister authority to order records from a person responsible for a system where the Minister has "reasonable grounds to believe that the use of a high-impact system could result in harm". This would be triggered by a system that might reasonably pose *any* risk of harm.

[38] *Ibid*, s 12.

[39] *Ibid*, s 17(1).

[40] Companion Document, *supra* note 20.

[41] The analogous provision in the EU's *AI Act*, *supra* note 11, art 9(4), is explicit on this point, requiring that mitigation measures be taken to reduce risk to an "acceptable" degree. (This is discussed further below.)

auditor" to audit an AI provider on reasonable grounds to believe obligations in Part I of the Act have been contravened.[42] For example, a dispute about whether a firm has taken adequate measures to identify or mitigate a risk, or made adequate disclosure in a relation to a system, could result in an audit. But here too, to have any effect, the audit is premised on the ability of an independent entity to assess risk and recommend mitigation measures that would reduce risk to a reasonable degree.

The assumption of risk clarity is also at play in the Act's most consequential powers: its penalty and offence provisions. One set of provisions involves regulatory offences for failures to carry out the risk and mitigation assessment requirements noted above. These impose significant fines on a person who obstructs or provides false or misleading information to the Minister in relation to these requirements.[43] The Act provides that a person does not commit the offence "if they establish that they exercised due diligence" to prevent it.[44] This means that if they can prove, on a balance of probabilities, that they took reasonable steps to ascertain and disclose risks to the Minister or mitigate them, they avoid liability.[45] Where the foreseeability of a risk or measures taken to mitigate it are debatable, it will be difficult for a court to decide whether a person has taken reasonable steps to assess or mitigate the risk.

The most serious offences in the Act pertain to the knowing use or possession of unlawfully obtained personal information for development or use in an AI system and making available a system that may cause serious harm—with each offence carrying prison terms of up to five-years and substantial fines.[46] A person commits the second of these offences where:

[42] *AIDA*, *supra* note 10, s 15.

[43] *Ibid*, s 30(2).

[44] *Ibid*, s. 30(4).

[45] *R v Sault Ste Marie*, 1978 CanLII 11 (SCC); *R v Wholesale Travel Group Inc*, 1991 CanLII 39 (SCC).

[46] *AIDA*, *supra* note 10, s 38 dealing with knowing use or possession of unlawfully obtain personal information; s 39 with the harm offences; and punishment in s 40. An entity might also be prosecuted under the *Criminal*

> (a) without lawful excuse and knowing that or being reckless as to whether the use of an artificial intelligence system is likely to cause serious physical or psychological harm to an individual or substantial damage to an individual's property, makes the artificial intelligence system available for use and the use of the system causes such harm or damage; or
>
> (b) with intent to defraud the public and to cause substantial economic loss to an individual, makes an artificial intelligence system available for use and its use causes that loss.[47]

The latter of these two forms of conduct would likely capture a narrow ambit of fraudulent conduct. The first of the two appears to contemplate a wider scope of activity—potentially offering the public greater protection—with two notable features. The offence is limited not only to harm that a system is likely to cause but harm that a system does cause. A person "causes" a prohibited consequence in the criminal context where they are a "significant contributing cause" or a contributing cause beyond the *de minimis*.[48] The provision also contemplates a special sense of what would constitute "recklessness". In Canadian criminal law, one can either know or be wilfully blind that something is likely,[49] or one can be reckless as to the risk or

Code, RSC 1985, c C-46 [*Criminal Code*], for causing harm where elements of an offence are made out, such as criminal negligence (s 219) or fraud (s 380).

[47] *AIDA*, *supra* note 10, s 39.

[48] In *Smithers v The Queen*, 1977 CanLII 7 (SCC) [*Smithers*], Dickson J, as he then was, writing for the Court, held that to establish the element of causation, the Crown need only prove the act at issue was "at least a contributing cause . . . outside the *de minimis* range": *ibid* at para 519. In *R v Nette*, 2001 SCC 78, Arbour J, for the majority, held that an acceptable alternative formulation of the causation standard in *Smithers* is a "significant contributing cause": *ibid* at para 72.

[49] In *R v Jorgensen*, 1995 CanLII 85 (SCC), Sopinka J for the majority, defined wilful blindness as involving the accused "strongly suspecting" a fact in issue (*ibid* at para 103). See also *R v Briscoe*, 2010 SCC 13 at paras 21–25.

possibility of it.[50] One cannot be reckless as to a likelihood, as is required here ("reckless as to whether the use of an artificial intelligence system is likely to cause serious physical or psychological harm").[51]

Yet the formulation here of being "reckless as to a likelihood" does have a precedent. It is analogous to the Supreme Court of Canada's formulation, in *R v Hamilton*,[52] of the *mens rea* of the *Criminal Code* offence of counselling an offence that is not committed.[53] Justice Fish, for the majority, defined this to require: "intent or conscious disregard of the substantial and unjustified risk inherent in the counselling". [54] More precisely, the accused must "either [have] intended that the offence counselled be committed, or knowingly counselled the commission of the offence while aware of the unjustified risk that the offence counselled was in fact likely to be committed as a result of the accused's conduct."[55] The Court does not use the term "reckless" here, but acting despite an "awareness of risk" of a "likelihood" entails a form of recklessness similar to the one found in section 39 of the *AIDA*.

To summarize, the *AIDA* requires commercial providers of AI systems—likely to include providers of language model AI—to meet a series of obligations pertaining to levels of discernable risk. They must identify and mitigate general risks of harm. They must notify the Minister when a system "is likely to result in material harm", but only "as soon as feasible".[56] Criminal liability will be imposed where a firm knew of a likelihood of a given harm (or an "unjustified risk" of it), rather than a mere possibility.

[50] In *R v Sansregret*, 1985 CanLII 79 (SCC), McIntyre J, for the Court, defined recklessness as "one who, aware that there is danger that his conduct could bring about the result prohibited by the criminal law, nevertheless persists, despite the risk. . . . [I]n other words, the conduct of one who sees the risk and who takes the chance": *ibid* at para 16.

[51] *AIDA*, *supra* note 10, s 39(a).

[52] *R v Hamilton*, 2005 SCC 47 [*Hamilton*].

[53] *Criminal Code*, *supra* note 46, s 464

[54] *Hamilton*, *supra* note 52 at para 29.

[55] *Ibid*.

[56] *AIDA*, *supra* note 10, s 12.

The Act may be vague about when these thresholds are met —when a risk of harm becomes identifiable, likely to occur, or feasible to discern. But a more fundamental concern is that the entire framework is premised on a firm's *ability* to ascertain whether and when a system risk meets a given threshold. In ways to be explored in Part II, there are strong reasons to question whether firms have this ability in the case of language model systems. The next section shows how similar assumptions about quantifying risk are also fundamental to the European Union's legislative proposal.

B. THE EUROPEAN UNION'S *ARTIFICIAL INTELLIGENCE ACT*

At the time of this writing, December of 2023, members of the European Parliament and the Council presidency of the European Union have reached agreement as to the principal components and provisions of the *AI Act*; but the Act has yet to be finalized and formally adopted.[57] Much debate has unfolded as to how the Act should apply to "foundation models", or AI systems involving machine learning to train large language models such as GPT-4 or other systems that produce images such as Dall-E or voice transcription such as Whisper AI.[58] In June of 2023, lawmakers decided to treat foundation models as a discrete category of AI, but one to which a number of requirements imposed against other potentially harmful systems

[57] On the final agreement, see the European Parliament, "Artificial Intelligence Act: Deal on Comprehensive Rules for Trustworthy AI", *European Parliament News* (9 December 2023), online: <europarl.europa.eu /news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensiverules-for-trustworthy-ai>. The European Commission first proposed the Act in April of 2021; the Council proposed amendments in a "Compromise Proposal" of November 2022; and the European Parliament made further amendments in June of 2023. All three documents are cited *supra* note 11; all references to the Act here are to the November 2022 Compromise Proposal as modified by the June 2023 draft. (The Compromise proposal contains a complete text of the Act; the June 2023 draft includes only a list of amendments.)

[58] See the overview in AI Now Institute, Amba Kak, & Sarah Myers West, "General Purpose AI Poses Serious Risks, Should Not Be Excluded from the EU's AI Act: Policy Brief" (13 April 2023), online: <ainowinstitute.org/ publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act>.

apply. However, given the fact that the Act might still be amended, I proceed in this section by asking: if the *AI Act*'s most onerous obligations were to apply to language model providers, what would this require of them in terms of risk quantification and control?

The *AI Act* is similar to Canada's framework in defining systems to which the Act will apply, imposing an obligation to implement risk assessment and mitigation measures, and to report on these measures. It also contains powers to order a provider of AI to stop using or making a system available where it causes harm or economic loss. Notable ways it differs from Canada's *AIDA* are in its application to potential harm or damage a system may cause to groups as well as to individuals, and in its finer distinction between levels of risk a system may pose, including a category of risk held to be unacceptable. Yet, like Canada's Act, the central obligations in the EU bill are those that require AI providers to identify and mitigate risks—obligations which rest on the assumption that risks can be ascertained to a reasonable degree and effectively mitigated *ex ante*.

The Act's various obligations rest on a distinction between systems engaging four levels of risk. The highest level of risk involve uses of AI the Act prohibits. These include systems "reasonably likely to cause" a person or group "significant harm" through "subliminal techniques" or the exploitation of vulnerabilities due to age, social or economic situation—among other possibilities.[59] The next level concerns "high-risk" AI systems, which are those engaging health and safety concerns through their use in safety components of vehicles, electronics, and other consumer products, or systems identified in an annex to the Act that engage fundamental rights.[60] The latter include the use of AI in infrastructure, education, employment, public administration, law enforcement, and immigration to automate

[59] *AI Act*, *supra*, note 11, arts 5(1)(a)–(b). Other unacceptably risky uses of AI systems include social credit scoring leading to unfavourable treatment of persons or groups in other contexts or in a disproportionate manner (*ibid*, art 5(1)(c)) and law enforcement use of real-time biometric identification systems in public space (*ibid*, art 5(1)(d)).

[60] *Ibid*, art 6.

decisions affecting basic rights and freedoms, such as accessing services, promotion, or entry into a Member State.[61] Notably, AI systems may be added to the Annex—including general purpose AI systems—where they will be used in these areas and where they:

> [P]ose a significant risk of harm to health and safety, or an adverse impact on fundamental rights . . . and that risk is, in respect of its severity and probability of occurrence, equivalent to or greater than the risk of harm or of adverse impact posed by the high-risk AI systems already referred to in [the] Annex.[62]

The Act contemplates a category of AI systems posing a low or limited risk, which includes systems meant to "interact with natural persons" using biometric identification, emotion recognition technology, or image, audio, or video deep-fake manipulation tools.[63] A final category of minimal risk AI systems —such as spam filters or AI-enabled video games—is implied in article 69 of the Act, which invites voluntary compliance with codes of conduct Member States will create to encourage environmental sustainability and accessibility for persons with disabilities.[64]

The *AI Act* defines an "artificial intelligence system" as "a machine-based system" operating with "varying levels of autonomy" that generates "predictions, recommendations, or decisions [which] influence physical or virtual environments".[65] Proposed amendments to the Act in June 2023 identified a

[61] *Ibid*, Annex III.

[62] *Ibid*, art 7(1). Art 3(1a) states that "'risk' means the combination of the probability of an occurrence of harm and the severity of that harm" and 3(1b) states that "'significant risk' means a risk that is significant as a result of the combination of its severity, intensity, probability of occurrence, and duration of its effects, and its the ability to affect an individual, a plurality of persons or to affect a particular group of persons."

[63] *Ibid*, art 5.

[64] I have borrowed the terms "limited" and "minimal" risk, and the examples of minimal risk provided here, from Eve Gaumond, "Artificial Intelligence Act: What Is the European Approach for AI?" (4 June 2021), online: <lawfaremedia.org/article/artificial-intelligence-act-what-european-approach-ai>.

[65] *AI Act*, *supra*, note 11, art 3(1)

special category of "general purpose AI system", which can be used for applications "for which it was not intentionally and specifically designed", and a "foundation model," which is a system "designed for generality of output".[66] The Act now specifies that all "operators" subject to the Act "shall make their best efforts to develop and use AI systems or foundation models in accordance with" a set of "general principles" that include "technical robustness and safety", which requests that systems:

> [B]e developed and used in a way to minimize unintended and unexpected harm as well as being robust in case of unintended problems and being resilient against attempts to alter the use or performance of the AI system so as to allow unlawful use by malicious third parties.[67]

What may be the most crucial of the Act's requirements can be found in article 9, which imposes on persons overseeing a high-risk system an obligation to put in place a "risk management system", [68] supported by various record-keeping, disclosure, and assessment requirements elsewhere in the Act.[69] A risk management system involves the "identification, estimation and evaluation of the known and the reasonably foreseeable risks that the high-risk AI system can pose to the health or safety of natural persons" when used for "its intended purposes" or under "reasonably foreseeable misuse"—and the "adoption of appropriate and targeted risk management measures designed to address the risks identified".[70] An earlier draft of the article did

[66] *Ibid*, arts 3(1c), (1d).

[67] *Ibid*, art 4a(1b).

[68] *Ibid*, art 9.

[69] *Ibid*, title III at chs 2–3, 5.

[70] *Ibid*, arts 9(2)(a), 9(2)(d). Some have raised the concern that a provider could largely circumvent this obligation by asserting that any risk or misuse of a system does fall within its "intended purpose". See David Matthews, "A New Type of Powerful Artificial Intelligence Could Make EU's New Law Obsolete", *Science Business* (21 December 2021), online: <sciencebusiness.net/news/new-type-powerful-artificial-intelligence-could-make-eus -new-lawobsolete>. At the time of writing, lawmakers continue to debate how best to address this. See Tambiama Madiega, "General-Purpose Artificial Intelligence" (March 2023) European Parliamentary Research

not identify when a risk is known or foreseeable.[71] The "Compromise Proposal" in 2022 added the qualification to article 9 that the "risks referred to in this paragraph shall concern only those which may be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or the provision of adequate technical information."[72] This would thus appear to absolve a system provider of liability for failing to identify risks that were not reasonably foreseeable or avoidable. What is foreseeable, however, or how much effort a provider must put into identifying new risks, remains unclear.[73] A further provision states that "risk management measures . . . shall be such that relevant residual risk associated with each hazard as well as the overall residual risk of the high-risk AI systems is reasonably judged to be acceptable".[74] When any of these standards will be met involving an AI system—"risks most likely to occur", "suitable measures," "overall residual risk . . . judged to be acceptable"—is unclear.[75]

The Act also imposes an obligation that high-risk AI systems be "developed in such a way" that they can be "effectively overseen by natural persons as proportionate to the risks associated with those systems."[76] The oversight must aim at "preventing or minimising the risks to health, safety or fundamental rights" arising from "reasonably foreseeable misuse".[77] The latter phrase is defined in article 3 to mean the "use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other

Service at 2, online (pdf): <europarl.europa.eu/RegData/etudes/ATAG/2023/745708/EPRS_ATA(2023)745708_EN.pdf>

[71] Jonas Schuett, "Risk Management in the Artificial Intelligence Act" (2023) 15:2 Eur J Risk Regulation 367 at 375.

[72] *AI Act*, *supra*, note 11, art 9(2).

[73] Schuett, *supra* note 71 at 376.

[74] *AI Act*, *supra* note 11, art 9(4).

[75] Similar risk thresholds appear in other EU legislation. See the discussion below of the risk threshold in the EU's market surveillance regulation.

[76] *AI Act*, *supra* note 11, art 14(1).

[77] *Ibid*, art 14(2).

systems".[78] Here too, the scope of "reasonably foreseeable" behaviour is unclear.

Once deployed in the European market, the Act imposes a further set of obligations on creators of AI to "establish and document a post-market monitoring system in a manner that is proportionate to the risks of the high-risk AI system."[79] This involves both transparency and self-reporting obligations. Providers of a high-risk system can be ordered to disclose information about a system necessary to confirm compliance.[80] They must also report "any serious incident" to relevant authorities "immediately after the provider has established a causal link between the AI system and the serious incident or the reasonable likelihood of such a link".[81] Where a "market surveillance authority" of a Member State of the Union identifies that an AI system presents a risk that meets a defined threshold, it may order the operator of the system to withdraw the product.[82] This threshold is met when the system has:

> [T]he potential to affect adversely health and safety of persons in general, health and safety in the workplace, protection of consumers, the environment, public security and other public interests . . . to a degree which goes beyond that considered reasonable and acceptable in relation to its intended purpose or under the normal or reasonably foreseeable conditions of [its use].[83]

The Act is explicit in submitting general purpose AI systems to this power.[84]

Finally, the Act imposes significant monetary penalties upon operators of general purpose AI systems for breaching obligations under the Act that include the requirement to institute suitable risk management systems in article 9 and

[78] *Ibid*, art 3(13).

[79] *Ibid*, art 61(1).

[80] *Ibid*, art 23.

[81] *Ibid*, art 62(1).

[82] *Ibid*, arts 65(1), 65(5).

[83] *Ibid*, art 65(1).

[84] *Ibid*, art 67(1).

human oversight in article 14.[85] Fines will be determined by considering, among other factors, "the intentional or negligent character of the infringement".[86] One commentator has noted that although the *AI Act* does not contain criminal enforcement provisions, a negligent failure to disclose information or mitigate or avoid a foreseeable harm might attract criminal liability in a Member State.[87] However, this would turn in part, as it does in Canada's criminal provisions under the *AIDA*, on an assessment of whether the harm at issue was reasonably foreseeable and avoidable.

In contrast to Canada's *AIDA*, the EU's *AI Act* imposes on the providers of AI systems a wider range of obligations of differing degrees of onerousness. The European bill also includes a wider range of risk thresholds, such as giving rise to a "potential to affect adversely . . . which goes beyond that considered reasonable and acceptable",[88] "known and . . . reasonably foreseeable risks" most likely to occur,[89] and "overall residual risk . . . reasonably judged to be acceptable".[90] The vagueness of these thresholds anticipates the likely challenge of litigating them. But aside from this, the very idea of thresholds based on reasonable, effective, or acceptable risk assessments and mitigation measures is premised—as it is in Canada's *AIDA*—on the ability of a language model provider (among other AI systems) to carry out effective assessments of the extent of residual risks in advance of their wide deployment.

II. LARGE LANGUAGE MODELS AND RISK OPACITY

A variety of evidence calls into question whether providers of language model AI or independent auditors with the benefit of full disclosure of details about a model would be capable of

[85] *Ibid*, arts 71(4), 71(5).

[86] *Ibid*, art 71(6)(c b).

[87] Schuett, *supra* note 71 at 384, providing as an example s 823(2) of the German Civil Code.

[88] *AI Act*, *supra* note 11, art 65(1).

[89] *AI Act*, *supra* note 11, art 9(2)(a).

[90] *Ibid*, art 9(4).

identifying or mitigating risks, in advance of wide deployment, to a reasonable or acceptable degree. The evidence can be found in studies and reports involving adversarial tests of language model AI, in media reports involving incidents of actual harm, and in studies pertaining to model interpretability. It suggests that while model providers can take measures to mitigate risks to some degree, residual risks remain that are significant but difficult to quantify. This Part begins by surveying evidence in the first category and considering OpenAI's "Technical Report" before proceeding to other evidence.

A. THE GPT-4 "TECHNICAL REPORT"

When OpenAI released the latest version of its large language model AI system, GPT-4, to the public in March of 2023, it published a "Technical Report" explaining the model's improvements over earlier versions, along with risks of harm to which the model gives rise and measures it took to mitigate them.[91] The California-based company did not publish the Report under a regulatory obligation set out in a federal or state-level AI statute analogous to Canada's *AIDA* or the EU's *AI Act*. However, the text cannot be construed as an example of the kind of disclosure either act contemplates, because the Report withholds information about the nature of the model that would likely be required under the *AIDA* and the *AI Act*.[92] The authors note only

[91] OpenAI, "Technical Report", *supra* note 1.

[92] The disclosure and audit provisions in Canada's *AIDA*, *supra* note 10, set out broad powers to demand disclosure. S 11(2) of the Act requires a provider of a high-impact system to make public a "plain-language description of the system that includes an explanation" of various things, including measures to mitigate harm, along with "any other information that may be prescribed by regulation": *ibid*, s 11(2). S 15(3) states that a system provider "who is audited must give all assistance that is reasonably required to enable the auditor to conduct the audit, including by providing any records or other information specified by the auditor": *ibid*, s 15(3). In art 23 of the EU's *AI Act*, *supra* note 11, providers of high-risk AI systems may be ordered to disclose "all the information and documentation necessary to demonstrate the conformity" of the system with obligations in art 9 (creating a risk management system) and various record-keeping obligations in arts 11 and 12. Obligations under each act would likely entail further disclosure than OpenAI provides here.

that GPT-4 is a "Transformer-style model" trained on "publicly available data (such as internet data) and data licensed from third-party providers", and cite the "competitive landscape and the safety implications of large-scale models like GPT-4" for refusing to be more transparent.[93] The Report contains "no further details about the [model's] architecture (including model size), hardware, training compute, dataset construction, training method, or similar."[94] The document might therefore be dismissed as an elaborate press release masquerading as a scientific paper, since its claims about risk mitigation cannot be readily tested by independent observers.

However, despite OpenAI's lack of transparency about GPT-4 in the Report, the risk identification and mitigation exercise it documents is an important source of evidence of language model risk—and the challenge of ascertaining it. It details a number of specific dangers to which GPT-4 gives rise, suggests that they cannot be completely avoided, and provides some indication of why the extent of the residual risks cannot be readily quantified.

The Report's aim was to canvas the "extent of [the] risks" posed by "new risk surfaces" arising from GPT-4's "additional capabilities" over its predecessor models.[95] It details the work of over 50 experts in various fields including cybersecurity and "biorisk" to "adversarially test the model."[96] The testing helped the firm improve the model through, among other techniques, "reinforcement learning with human feedback", which involves human oversight of dangerous or undesirable prompts and a reward system meant to steer the model away from producing

[93] OpenAI, "Technical Report", *supra* note 1 at 2. To be clear, the *AIDA* contemplates the possibility of disclosures made only to the Minister or to an auditor, which are then kept confidential and shared with other government actors or made public only under certain conditions. The *AI Act* does the same. The "Technical Report" might suffice in terms of the public disclosure requirements in either or both acts, but may not suffice in the case of an audit or external assessment.

[94] OpenAI, "Technical Report", *supra* note 1 at 2.

[95] *Ibid* at 11–12.

[96] *Ibid*.

harmful outputs.[97] Mitigation measures, the authors note, "improved many of GPT-4's safety properties"[98] over those of GPT-3.5, with statistical reductions in the rate of responses for "disallowed content", "sensitive requests" for medical advice or self-harm, and "toxic generations".[99]

While some detail is provided to support these improved safety findings, they speak only to the extent of a reduction in risk relative to earlier models. The Report says nothing about how safe the model is *in a general sense*. How likely is it to produce harmful content? How easily can it be jailbroken? While it conspicuously avoids this more fundamental issue, a further document contained within the Technical Report, titled "GPT-4 System Card", canvasses the red team's adversarial testing of the model, specific dangers it creates, and the team's reservations about broader risks arising from the wide deployment of the model.[100]

The System Card confirms the model's tendency to "hallucinate" or produce "untruthful" content, which can be "particularly harmful" in light of the model having become more convincing and believable, encouraging "overreliance" on the part of users.[101] The model can generate "instances of hate speech, discriminatory language, incitements to violence, or content that is then used to either spread false narratives or to exploit an individual."[102] This includes advice or encouragement for self-harm; graphic erotic and violent content; harassing and demeaning content; information about planning attacks or carrying out violence; and instructions for finding illegal content.[103] The red team notes GPT-4's greater capability over GPT-3.5 for "disinformation and influence operations", on the basis of improvements in producing realistic, targeted, and

[97] *Ibid* at 12.

[98] *Ibid* at 13.

[99] *Ibid*.

[100] OpenAI, "System Card", *supra* note 3.

[101] *Ibid* at 46.

[102] *Ibid* at 47.

[103] *Ibid*.

misleading content—results, they suggest, which can "rival human propagandists in many domains, especially if teamed with a human editor."[104]

A further concern is the model's ability to provide information to "proliferators" of conventional and unconventional weapons, such as those seeking to "develop, acquire, or disperse nuclear, radiological, biological, and chemical weapons."[105] The model is "most likely to be useful for individuals and non-state actors" without "formal scientific training", by providing "general information on common proliferation pathways," suggesting "vulnerable public targets," explaining security measures used to protect necessary materials, and identifying "fundamental components that are required to engineer a radiological dispersal device."[106] The model "readily re-engineered some biochemical compounds that were publicly available online, including compounds that could cause harm at both the individual and population level."[107] Yet the model also often generated "vague" or "inaccurate" instructions when asked for more detail about how to build a radiological device or biochemical compound.[108] The test of how dangerous the model could be in this context was ultimately inconclusive, but as the authors note, their work was "not intended to assess the probability or likelihood of a user accessing the model for the purpose of developing unconventional weapons."[109] Impliedly, it was meant only to identify the *possibility* of the model's use in this way.

To mitigate these "safety challenges", the red-team took measures to "fine-tune" GPT-4, including filtering pre-training datasets and using "reinforcement learning from human feedback" techniques.[110] The authors demonstrate the

[104] *Ibid* at 50.

[105] *Ibid* at 52.

[106] *Ibid*.

[107] *Ibid*.

[108] *Ibid* at 53.

[109] *Ibid* at 52.

[110] *Ibid* at 61–65.

effectiveness of these measures in the case of each of the dangers canvassed above by including examples of prompts that once did but no longer do generate harmful results.[111] Yet they note the model can "still be vulnerable to adversarial attacks and exploits" and "the potential to generate harmful content, remain[s] latent."[112] They do not address the degree of this latency, i.e., the likelihood that the model will still generate harmful output. They point instead to a gap in understanding:

> Further research is needed to fully characterize these [latent] risks. In particular, we would like to see work on more robust evaluations for the risk areas identified and more concrete measurements of the prevalence of such behaviors across different language models, and to guide the development of these models in safer directions.[113]

The larger point here is what the red team did not claim. They did not assert the model is now likely to generate harmful content with a reasonably low probability in a general sense (i.e., not just low compared to GPT-3.5). Despite the team spending many months testing and fine-tuning the model, a region of risk of unknown scope and depth remains.[114]

In other public statements, OpenAI has conceded the intractability of the problem. In a blog post, the firm states that "[m]any aspects of language models' risks and impacts remain hard to measure and therefore hard to monitor, minimize, and disclose in an accountable way."[115] Despite making "active use of existing academic benchmarks for language model evaluation", they have found that "existing benchmark datasets are often not reflective of the safety and misuse risks we see in practice."[116] Elsewhere they assert that "[w]e work hard to prevent foreseeable risks before deployment, however, there is a limit to

[111] For sample prompts for hateful content see *ibid* at 48.

[112] *Ibid* at 68.

[113] *Ibid* at 43.

[114] *Ibid* at 68.

[115] OpenAI, "Lessons Learned", *supra* note 4.

[116] *Ibid.*

what we can learn in a lab."[117] Despite extensive testing, they "cannot predict all of the beneficial ways people will use our technology, nor all the ways people will abuse it."[118]

OpenAI concedes but does not quantify residual risk. This raises two questions: Are residual risks of harm here real or speculative? And would OpenAI or any provider of language model AI be *capable* of quantifying residual risks with greater specificity?

B. Substance and Quantifiability of Residual Risk

A body of further evidence suggests that, despite mitigation measures providers have taken, GPT-4 and other language models in wide deployment do entail residual risks of harm that are real or substantial. This includes evidence that language models have already contributed to, if not caused, serious harm. There are also reasons to question whether language model providers or independent auditors with the benefit of full disclosure about a model could quantify the extent of these risks, in advance, to a reasonable degree, or that all of these risks can be brought under reasonable control in the foreseeable future. Each of these points has direct implications for the *AIDA* and the *AI Act* explored in the concluding segment of the paper.

1. Evidence that Risks are Substantial

Evidence of the real or substantial nature of residual risks in language model deployment can be found on two fronts: more formal studies of risks in discrete areas including crime and misinformation, and anecdotal evidence or reported events.

An extensive study by the Europol Innovation Lab in early 2023 involving experts in various fields of Europol canvassed possible criminal misuses of ChatGPT (referred to here to include both GPT-3.5 and 4) along with other language models.[119] The authors' central finding was that although ChatGPT's "safety mechanisms are constantly updated" to decline to answer

[117] OpenAI, "AI Safety", *supra* note 4.

[118] *Ibid*.

[119] Europol, "Flash Report", *supra* note 1 at 2.

questions deemed harmful or biased, they can be "circumvented fairly easily through prompt engineering."[120] An earlier example of this was the "Do Anything Now" (DAN) jailbreak, "a prompt specifically designed to bypass OpenAI's safeguards and lead ChatGPT to respond to any input, regardless of its potentially harmful nature."[121] While OpenAI has addressed this vulnerability, "new and ever more complex versions of DAN have emerged subsequently, all designed to provide jailbreak prompts that can navigate through the safety mechanisms built into the model."[122] OpenAI continues to address them, and the authors note there was no "functional DAN" available at the time of writing, but they highlight the ongoing, unavoidable cat and mouse game this involves.[123]

Short of jailbreaking the model to "do anything", the Europol study notes that a further significant danger posed by ChatGPT is its ability to inform criminals quickly of vital information or steps necessary to commit a range of particular crimes, from "how to break into a home, to terrorism, cybercrime and child sexual abuse."[124] While the information is also available online, the ability to use the model to "provide specific steps by asking contextual questions means it is significantly easier for malicious actors to better understand and subsequently carry out various types of crime."[125] The key weakness criminals may exploit is that safeguards put in place to prevent harmful output "only work if the model understands what it is doing."[126] Breaking queries

[120] *Ibid* at 4–5 [emphasis omitted].

[121] *Ibid* at 6.

[122] *Ibid*.

[123] *Ibid*. For further support for the likelihood that language models will require continuous response to novel adversarial attacks, see Daniel Kang et al, "Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks" (11 February 2023), online: <doi.org/10.48550/arXiv.2302.05733> (noting that "LLMs will increasingly attract more sophisticated adversaries and attacks, and addressing these attacks may require new approaches to mitigations" at 1).

[124] Europol, "Flash Report", *supra* note 1 at 7.

[125] *Ibid*.

[126] *Ibid* at 8 [emphasis omitted].

down into steps escapes detection. All of the queries they successfully ran on GPT-3 worked on GPT-4 and "[i]n some cases, the potentially harmful responses from GPT-4 were even more advanced."[127] The model helped them draft more effective, persuasive phishing scams than non-native English speakers would otherwise have produced; it helped produce malicious code; and it generated text for large-scale disinformation campaigns.[128]

Other commentators recount similar abilities to easily circumvent language model safeguards for advice about how to commit crimes. Journalist Sue Halpern, for example, reports that she was "able to get GPT-4 to explain how to use fertilizer to create an explosive device by asking it how Timothy McVeigh blew up the Alfred P. Murrah Federal Building, in Oklahoma City, in 1995".[129] Janus Rose documents obtaining detailed instructions about how to "shoplift without getting caught" and how to make thermite ("a pyrotechnic chemical that can burn through metal").[130] Entire sites have emerged devoted to posting prompts that would jailbreak or circumvent chatbot safeguards, corroborating Europol's finding that the process of addressing jailbreaks will be iterative and continuous for the foreseeable future.[131] Ample further evidence supports this inference.[132]

[127] *Ibid* at 7.

[128] *Ibid* at 7–9.

[129] Sue Halpern, "What We Still Don't Know About How AI is Trained", *New Yorker* (28 March 2023), online: <newyorker.com/news/daily-comment/what-we-still-dont-know-about-how-ai-is-trained>.

[130] Janus Rose, "OpenAI's New Chatbot Will Tell You How to Shoplift and Make Explosives", *Vice* (1 December 2022), online: <vice.com/en/article/openais-new-chatbot-will-tell-you-how-to-shoplift-and-make-explosives/>.

[131] See e.g. jailbreakchat.com and the discussion of jailbreaking in Melissa Heikkilä, "Three Ways AI Chatbots Are a Security Disaster", *MIT Technology Review* (3 April 2023), online: <technologyreview.com/2023/04/03/1070893/three-ways-ai-chatbots-are-a-security-disaster/>.

[132] See the sources cited in Matt Burgess, "The Hacking of ChatGPT Is Just Getting Started", *Wired* (13 April 2023), online: <wired.com/story/chatgpt-jailbreak-generative-ai-hacking/>. Burgess also details how security researchers have:

Researchers at NewsGuard, which provides credibility ratings for news and journalism sites, conducted a study in early 2023 of ChatGPT's response to false narratives derived from the company's misinformation database.[133] It "found that GPT-4 advanced prominent false narratives not only more frequently, but also more persuasively than ChatGPT-3.5".[134] This took the form of generating "news articles, Twitter threads, and TV scripts mimicking Russian and Chinese state-run media outlets, health-hoax peddlers, and well-known conspiracy theorists."[135] GPT-4's responses were "generally more thorough, detailed, and convincing, and they featured fewer disclaimers."[136] The authors concluded that the tool could readily be used to spread misinformation at scale, and commit "scams", and other "fraudulent or deceptive activity", despite OpenAI's use policies and mitigation measures.[137]

Yet other research contends that despite the novel and powerful ways that language models may be used for misinformation, risks can be effectively mitigated.[138] The

> now created a "universal" jailbreak, which works against multiple large language models (LLMs)—including GPT-4, Microsoft's Bing chat system, Google's Bard, and Anthropic's Claude. The jailbreak, which is being first reported by WIRED, can trick the systems into generating detailed instructions on creating meth and how to hotwire a car.

The article anticipates this jailbreak may soon be addressed but the problem of new and creative jailbreaking will persist. See also Stephanie Stacey, "Jailbreaking ChatGPT is the New Virtual Pastime. Why Won't LLMs Stick to Their Own Rules?", *Tech Monitor* (25 April 2023), online: <techmonitor.ai/digital-economy/ai-and-automation/jailbreaking-chatgpt-why-wont-llms-stick-to-their-own-rules>.

[133] Arvanitis, Sadeghi & Brewster, *supra* note 1.

[134] *Ibid*.

[135] *Ibid*.

[136] *Ibid*.

[137] *Ibid*. See also Michael Atleson, "Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale" (20 March 2023), online (blog): <ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale>.

[138] Josh A Goldstein et al, "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations" (10 January 2023) at 6–7, online: <doi.org/10.48550/arXiv.2301.04246>.

argument is premised on the view that waging a successful influence operation involves a host of variables, including access to a model, a means of disseminating content, and material impacting a target.[139] Each of these facets "represents a possible stage for intervention."[140] One inference to draw from this debate is that while it may be the case that counter-measures can be employed at various stages of an influence campaign, the fact that effective mitigation depends on multiple points of intervention makes it difficult to predict in advance whether and how a language model might be used effectively here.

The harms canvassed above may be substantial but remain mostly theoretical. However, evidence is beginning to emerge of language models contributing to actual harm. A series of reports have documented aggressive misbehavior involving chatbots engaging in psychological manipulation (to disrupt a marriage),[141] to produce hate speech,[142] and to generate defamatory output.[143] Writing at the end of 2022, one AI expert foresaw the likelihood that a large language model would aid or abet suicide or murder.[144] Earlier testing of GPT-3 had demonstrated the model's utility in encouraging a depressed

For further critical opinion of the threat of large-scale disinformation by language model AI, see also Arvind Narayanan & Sayash Kapoor, "The LLaMA Is Out of the Bag. Should We Expect a Tidal Wave of Disinformation?" (6 March 2023), online: <aisnakeoil.substack.com/p/the-llama-is-out-of-the-bag-should>.

[139] Goldstein et al, *supra* note 138 at 7.

[140] *Ibid*.

[141] Roose, *supra* note 6.

[142] Will Douglas Heaven, "How to Make a Chatbot That Isn't Racist or Sexist", *MIT Technology Review* (23 October 2020), online: <technologyreview.com/2020/10/23/1011116/chatbot-gpt3-openai-facebook-google-safety-fix-racist-sexist-language-ai/>.

[143] Kaye, *supra* note 5; Pranshu Verma & Will Oremus, "ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused", *Washington Post* (5 April 2023), online: <washingtonpost.com/technology/2023/04/05/chatgpt-lies/>.

[144] Gary Marcus, "The Dark Risk of Large Language Models", *Wired* (29 December 2022), online: <wired.com/story/large-language-models-artificial-intelligence/>.

user expressing suicidal thoughts to kill themselves.[145] In March of 2023, news reports tied the suicide of a Belgian man to his extensive conversations over several weeks with the chatbot Eliza, including a conversation involving suicide in which the chatbot provided some encouragement.[146]

The momentum of these events provoked a controversial call among experts in the AI community to halt public deployment of models newer than GPT-4.[147] The "Open Letter" expresses fear that "AI labs [are] locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one—not even their creators—can understand, predict, or reliably control."[148] While the letter drew considerable criticism about the viability or potential effect of a halt to deployment,[149] the entire debate was spurred by the uncertainty surrounding the nature and extent of the risk language models pose in wide deployment, and the challenge of quantifying it. Put another way, the debate would have been pre-empted if one or more model provider could point to a credible body of evidence—an independent report or, a set

[145] *Ibid*, noting experiments conducted by a French firm Nabla. See Kevin Riera, Anne-Laure Rousseau & Clément Baudelaire, "Doctor GPT-3: Hype or Reality?" (27 October 2020), online (blog): <nabla.com/blog/doctor-gpt-3>.

[146] Walker, *supra* note 7. The circumstances and suicidal exchange are recounted in detail in Pierre-François Lovens, "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là", *La Libre* (28 March 2023), online: <lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/>. The exchange is translated in Gary Marcus, "The First Known Chatbot Associated Death" (4 April 2023), online (blog): <garymarcus.substack.com/p/the-first-known-chatbot-associated>.

[147] Bengio et al, *supra* note 8.

[148] *Ibid*.

[149] See, e.g. responses among AI experts cited in Kari Paul, "Letter Signed by Elon Musk Demanding AI Research Pause Sparks Controversy", *The Guardian* (1 April 2023), online: <theguardian.com/technology/2023/mar/31/ai-research-pause-elon-musk-chatgpt>. See also sources cited in Emilia David, "Leaders Like Elon Musk Want to Pause AI Development, but the Power of the Free Market Means It's Impossible to Stop", *Business Insider* (3 April 2023), online: <businessinsider.com/openai-elon-musk-pause-development-letter-never-happen-2023-4>.

of indicia—that could show that a publicly deployed language model only produces harmful content within a reasonably low statistical range.[150] This points to the question: could this be done?

2. Impediments to Quantifying and Controlling Language Model Risks

If the risk that language models thus far pose in wide deployment are real or substantial, can they not be quantified and effectively mitigated? Would powers in the *AIDA* or the *AI Act* to demand greater transparency into the nature and operation of a language model result in more certainty about the extent of the risks a model poses or the effect of measures to mitigate them? There are strong reasons to believe the answer may be no in both cases. Evidence calls into question whether more transparency about the nature and makeup of a language model AI system would necessarily result in more accurate assessments of the nature or extent of risks in relation to it, or that transparency will enable more effective mitigation measures. The ability to quantify and control risks are separate issues, but the questions raised on either front are closely related.

One challenge to quantifying risks posed by language models evident in OpenAI's Technical Report,[151] its public statements,[152] and in the material surveyed in the previous section is the unpredictability of how the models may be used. One dimension of this is that language models lack "an inherent use case"—they

[150] What would constitute a "reasonably low statistical range" might be defined by analogy to standards in the context of vehicle safety, pharmaceuticals, or industrial chemical production, in which absolute safety cannot be assured but does not pose an impediment to public use or distribution of a product. On the discussion of "unreasonable risk" as a "basic standard of protection" in US regulatory law see William Boyd, "Genealogies of Risk: Searching for Safety, 1930s–1970s" (2012) 39:4 Ecology LQ 895 at 972-978.

[151] OpenAI, "Technical Report", *supra* note 1.

[152] See OpenAI, "Lessons Learned", *supra* note 4; OpenAI, "AI Safety", *supra* note 4.

work with countless other applications.[153] As a member of OpenAI's board explains, the models are "not trained to do one specific thing . . . [e]ven the people who create them don't actually know what they can and can't do."[154] Another dimension of unpredictability relates to novel forms of jailbreaking or ways in which one might produce harmful output from a chatbot by indirect queries ("how did McVeigh cause an explosion?"). Both facets of language model use make it difficult if not impossible to ascertain with a reasonable degree of certainty how likely it is that a chatbot will produce a given output or resist a nefarious input. Model providers continue to monitor this and take steps to respond, but there is, at present, no clear end in sight to the ongoing cycle of novel misuse and reaction.

Another significant impediment to both risk quantification and control involves the problem of "model explainability and interpretability".[155] As researchers for the AI firm DeepMind have noted, many machine learning models are thought to be "intrinsically opaque".[156] In some cases, "it is not easy for humans, no matter how skilled, to easily understand why and how a specific algorithmic output is generated."[157] Impediments to explaining or interpreting a language model can "make failures of the model harder to detect, posing a threat to AI safety."[158] The impediments can also "obscure the true capabilities of a model,"

[153] Billy Perrigo, "Big Tech Is Already Lobbying to Water Down Europe's AI Rules", *Time* (21 April 2023), online: <time.com/6273694/ai-regulation-europe/>, citing Helen Toner.

[154] *Ibid.*

[155] Weidinger et al, *supra* note 19 at 37.

[156] *Ibid.* See also David Gunning et al, "XAI—Explainable Artificial Intelligence" (2019) 4:37 Science Robotics 1, online: <doi.org/10.1126/scirobotics.aay7120> (noting that "[o]ften, the highest performing methods (e.g., DL [deep learning]) are the least explainable, and the most explainable (e.g., decision trees) are the least accurate" at 1).

[157] Weidinger et al, *supra* note 19 at 37.

[158] *Ibid* at 38.

and make it “harder for product developers and regulators to assess inappropriate use cases of such models”.[159]

Insight into the impediments to understanding how a language model produces output can be found in the most extensive study to date of language model risks, a paper authored by over a hundred AI researchers at Stanford in 2022.[160] It treats language models as a form of “foundation model”, which is “any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks” and includes GPT-3.[161] The study found that “[d]espite the impending widespread deployment of foundation models, we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties.”[162]

The authors highlight three challenges to “characterizing and forecasting the capabilities of current self-supervised foundation models” worth citing directly:[163]

> First, the generality of foundation models means that they can be applied to countless different kinds of applications in unexpected ways. Enumerating current and planned applications of foundation models is not sufficient to capture the full range of ways they could be used. Second, even within a particular application, model capabilities are emergent: they grow and change in unexpected ways as models scale. . . . Third, even within a particular application and scale, a model’s capabilities are not easy to characterize. . . . [S]mall rewordings

[159] *Ibid*. See also Roman V Yampolskiy, “Unexplainability and Incomprehensibility of Artificial Intelligence” (20 June 2019) at 8–9, online: <doi.org/10.48550/arXiv.1907.03869> (noting “[i]f all we have is a ‘black box’ it is impossible to understand causes of failure and improve system safety” at 9); Chloe Xiang, “Scientists Increasingly Can’t Explain How AI Works”, *Vice* (1 November 2022), online: <vice.com/en/article/scientists -incr easingly-cant-explain-how-ai-works/>.

[160] Rishi Bommasani et al, “On the Opportunities and Risks of Foundation Models” (last modified 12 July 2022), online: <doi.org/10.48550/arXiv.2108.07258>.

[161] *Ibid* at 3.

[162] *Ibid* at 1.

[163] *Ibid* at 116.

> of prompts can have large impacts on task performance. Since the space of prompts is intractable to enumerate, it is challenging to definitely assert that any task is outside the reach of current prompt-based foundation models—this is a major challenge for reasoning about possible catastrophic risks from foundation models.[164]

Their discussion of risk concludes with an encouragement of further research into "forecasting the exact capabilities and risks of foundation models", pointing to limits in the current state of knowledge.[165]

Yet how these observations apply to a specific language model is unclear. Nor do they address whether the challenges to interpretability noted here render a language model *effectively* unpredictable or beyond control. These details do not explain whether model risks can, at some point, be rendered reliably and reasonably low. But they do point to substantial challenges to model predictability and control that may not be easily overcome in the short term.

CONCLUDING CONSIDERATIONS ABOUT REGULATING LLM RISKS

How do these questions about measuring risk in wide deployment of language model AI complicate the risk-mitigation approach in the *AIDA* and the *AI Act*? Would proposals for reform address the concerns arising? Would alternative approaches —licensing and certification rather than self-monitoring—avoid these issues?

[164] *Ibid*. See also *ibid* at 123–24. The authors make a similar point about risk opacity in addressing the structure of foundation models as comprising numerous individual models devoted to specific tasks, such as language translation or arithmetic. This multi-model character "amplif[ies] manyfold" the challenge of "characterizing a [foundation] model's behavior"—simply predicting what it can do—because the "space of tasks that the model is able to perform is generally large and unknown, the input and output domains are often high-dimensional and vast (e.g., language or vision), and the models are less restricted to domain-specific behaviors or failure modes": *ibid*.

[165] *Ibid* at 117.

A. WHAT RISK UNCERTAINTY AND PERSISTENCE MEAN FOR THE *AIDA* AND THE *AI ACT*

This paper assumed, at the outset, a best case scenario in which the most onerous obligations in Canadian and European AI acts would apply to providers of language model AI systems. Part I noted the primary obligation in each act to identify and mitigate risks of harm and to be transparent with external auditors. Part II explored the challenges to identifying and measuring risks arising with language models in wide deployment. How these challenges complicate the application of the *AIDA* and the *AI Act* to language models turns on the reliance in both acts on reasonably accurate risk assessment.

It may help to briefly reiterate the centrality of this feature, before addressing the impact of risk uncertainty. Providers of "high-impact" or "high-risk" systems have, as their primary obligation under each act, to identify and mitigate risks to a reasonable or acceptable degree.[166] Regulators can order a provider to withdraw where material harm is likely or risk is unacceptable.[167] The *AIDA* imposes criminal liability when a system causes serious harm where a provider was aware of an "unjustified risk" of this happening.[168] Both acts impose fines for negligent violations, raising the question of whether a provider's failure to identify or mitigate harm was reasonable.[169] In each case, duties and obligations assume an ability to quantify the extent of risk effectively.

The audit provisions in each act share the same assumption. AI providers must disclose information about the nature and composition of their systems.[170] An auditor, provided full details about the makeup of a language model, can decide whether risks

[166] *AIDA*, *supra* note 10, s 8; Companion Document, *supra* note 20; *AI Act*, *supra* note 11, art 9(4).

[167] *AIDA*, *supra* note 10, ss 12, 17(1). This assumes, once again, that regulations will capture GPT-4 as a "high-impact system", to which the obligations in Part I of the Act apply. See also *AI Act*, *supra* note 11, arts 65(1)–(2).

[168] *AIDA*, *supra* note 10, s 39.

[169] *Ibid*, s 30; *AI Act*, *supra* note 11, art 71(6)(c a).

[170] See the discussion in note 93 for details.

of harm have been effectively identified or mitigated only if they too can discern this.

The evidence canvassed in Part II of this paper calls this ability into question. OpenAI has been explicit about the challenges it confronts in predicting potentially harmful uses of ChatGPT,[171] insisting that making the system widely available is necessary to making it safer.[172] OpenAI's red team claimed only to have reduced risks of GPT-4 relative to earlier GPT models, conceding that risks remain of an uncertain degree. OpenAI, Bing, and other language model AI providers would likely have foreseen the *possibility* of the actual harmful output noted in Part II—psychological manipulation, biased or defamatory output, and assistance in a suicide. But how likely it would have seemed is unclear. As firms continue to take steps to fine-tune their models to mitigate risk, it presumably diminishes. Yet the prospect of novel misuses of a model and model opacity call into question how accurate risk assessment can be in this context.[173]

Would greater transparency about a model to an independent auditor lead to more accurate risk assessments? Would requiring an audit before public deployment—as a condition of public release—not help to avoid harmful output? Here too, the

[171] See OpenAI, "Lessons Learned", *supra* note 4 (noting that "[m]any aspects of language models' risks and impacts remain hard to measure and therefore hard to monitor, minimize, and disclose in an accountable way").

[172] See OpenAI, "AI Safety", *supra* note 4 (noting that "there is a limit to what we can learn in a lab . . . we cannot predict all of the beneficial ways people will use our technology, nor all the ways people will abuse it. That's why we believe that learning from real-world use is a critical component of creating and releasing increasingly safe AI systems over time").

[173] A separate question not canvassed here is whether the general risk identification and mitigation obligation in both acts is infeasible due to the *breadth* of possible risks. See Philipp Hacker, Andreas Engel & Marco Mauer, "Regulating ChatGPT and Other Large Generative AI Models" (last modified 12 May 2023) online: <arxiv.org/abs/2302.02337>. Referring to the obligation in art 9 of the *AI Act*, *supra* note 11, to establish a risk management system, Hacker, Engel, and Mauer write at 5:

> Setting up such a system seems to border on the impossible, given [large language model] versatility. It would compel [language model] providers to identify and analyze all 'known and foreseeable risks most likely to occur to health, safety and fundamental rights' concerning all possible high-risk uses of the [models].

prospect of novel misuses and model opacity suggest possible limits to the reliability of an independent risk assessment. This is not to suggest that language model risks cannot ever be identified or mitigated to a reasonable degree, but rather, that a body of evidence calls into question how well this can be done at present or in the near future.

B. PROPOSALS FOR REFORM AND ALTERNATIVE FORMS OF REGULATION

Other commentators have shared the concern in this paper about the difficulty of quantifying risk in relation to language model AI and what this means for regulation. On one view, language models present a special kind of dynamic, difficult to assess risk that should be dealt with in AI legislation under its own category.[174] Lawmakers could draw on article 34 of the EU's *Digital Services Act* as a template for imposing an obligation on language model providers to "monitor for and mitigate systemic risks on a regular basis".[175] Another view notes that article 9(4) of the *AI Act* is silent as to when overall residual risk is judged "acceptable", but suggests amending the article to include a cost-benefit principle that would guide regulators.[176]

At present both the *AIDA* and the *AI Act* contemplate self-monitoring of risk on the part of a model provider and only imply the obligation to do this continuously, after deployment.[177] The *Digital Services Act* requires "very large online platforms" to

[174] Natali Helberger & Nicholas Diakopoulos, "ChatGPT and the AI Act" (2023) 12:1 Internet Pol'y Rev 1 at 6, online: <doi.org/10.14763/2023.1.1682>.

[175] *Ibid* at 4, citing EU, *Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC*, [2022] OJ, L 277/1, art 34 [*Digital Services Act*].

[176] Henry L Fraser & Jose-Miguel Bello y Villarino, "Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation" (last modified 22 November 2021) at s 5, online: <ssrn.com/abstract=3960461>.

[177] This is arguably implied in s 8 of the *AIDA*, *supra* note 10, and in the requirement to report likely risks of material harm in s 12; it is also implied in the *AI Act*, *supra* note 11, arts 9, 65(1).

conduct annual risk assessments.[178] However, being more explicit about the need for continuous risk monitoring would not avoid the problems outlined earlier of quantifying, predicting, and controlling risks. Similarly, article 9 of the *AI Act* might be amended to include a cost-benefit principle to guide AI providers and regulators to decide when risks will be "judged acceptable", but they would first need to effectively measure them.[179]

A further body of criticism points to deeper shortcomings with a risk-mitigation approach to regulating AI in general, including a tendency to elide or render difficult to quantify invisible harms.[180] Some suggest that the risk impact assessments central to these frameworks can "in practice be a meaningless box-ticking exercise, empty corporate compliance that is little more than heavy navel-gazing."[181] Alternative approaches include imposing conditional licensing or prior certification,[182] followed by court oversight,[183] or revocable licensing if and when harm is caused.[184] Prior licensing captures the thrust of the Open Letter cited in Part II, above.[185]

Prior and revocable licensing also share an assumption about the ability to identify and control risk, but shift the burden of

[178] *Digital Services Act*, *supra* note 175, art 34.

[179] *AI Act*, *supra* note 11, art 9(4).

[180] Margot E Kaminski, "Regulating the Risks of AI" (2023) 103:5 BUL Rev 1347 at 1352, online: <doi.org/10.2139/ssrn.4195066>.

[181] Margot E Kaminski, "The Developing Law of AI Regulation: A Turn to Risk Regulation" (21 April 2023) at 8, online: <lawfaremedia.org/article/the-developing-law-of-ai-regulation-a-turn-to-risk-regulation>, summarizing other critical views on this point [Kaminski, "Developing"].

[182] *Ibid* at 21.

[183] Matthew U Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies" (2016) 29:2 Harv JL & Tech 353 at 397–98, online: <doi.org/10.2139/ssrn.2609777>.

[184] Kaminski, "Developing", *supra* note 181 at 21.

[185] See Bengio et al, *supra* note 8 (asserting that "[p]owerful AI systems should be developed [and thus deployed] only once we are confident that their effects will be positive and their risks will be manageable"). See also Gianclaudio Malgieri & Frank Pasquale, "From Transparency to Justification: Toward Ex Ante Accountability for AI" (last modified 21 June 2022), online: <ssrn.com/abstract=4099657>.

proof onto providers. The licensing model errs on the side of caution, inscribing a presumption of danger and illegality that providers must rebut.[186] Consequently, where an auditor or regulator finds significant residual risks of harm that they cannot quantify, they might decline to license or certify a language model for public release. This might constitute effective regulation. But as OpenAI and others contend, it may not be possible to effectively ascertain language model risks without deploying the model publicly and subjecting it to real-world use.[187] In a licensing or certification regime, a language model provider might fail to overcome the hurdle of establishing their system is safe enough to deploy widely because they cannot render it safer without widely deploying it.

A final consideration is whether the concerns about residual risks of harm from language model AI might already be effectively addressed under consumer protection legislation—or best dealt with in that context.[188] But here too, the same conundrum arises. Under both Canadian and European consumer protection acts, liability turns on language of reasonably foreseeable hazards or acceptable risk.[189] In the wake

[186] Malgieri and Pasquale refer to this as "unlawfulness by default": Malgieri & Pasquale, *supra* note 185 at 1. See also Kaminski, "Developing", *supra* note 181 at 21.

[187] OpenAI, "AI Safety", *supra* note 4. Google's CEO Sundar Pichai expressed a similar view in his interview with hosts Kevin Roose and Casey Newton. See Hard Fork, "Google CEO Sundar Pichai on Bard, AI 'Whiplash' and Competing with ChatGPT" (31 March 2023) at 00h:11m:04s, online (podcast): <nytimes.com/2023/03/31/podcasts/hard-fork-sundar.html >.

[188] Sookman, *supra* note 21, s 4(a).

[189] The *Canada Consumer Product Safety Act*, SC 2010, c 21 regulates products that pose a "danger to human health or safety", which is defined in s 2 of the Act to mean "any unreasonable hazard—existing or potential—that is posed by a consumer product during or as a result of its normal or foreseeable use and that may reasonably be expected to cause the death of an individual exposed to it or have an adverse effect on that individual's health". See also EU, *Regulation 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation No 1025/2012 of the European Parliament and of the Council and Directive 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and*

of harm attributed to language model use, the same debate would arise over whether and when a risk was ascertainable, controllable, or avoidable.

To conclude, language model AI currently presents real and substantial risks of harm, though the extent of risk—the likelihood of specific harms arising—is unclear. This poses a challenge to regulators seeking to rely on a risk-mitigation model. The model is premised on an ability to effectively quantify, predict, and control the risk of harmful output. Yet ample evidence points to the challenge of doing so, and it suggests the possibility of this challenge persisting in the near future. A stricter licensing and certification framework for regulating language model AI would avoid risks but likely impose a significant obstacle to development.

However, the evidence of language model risks does support the conclusion that we know enough at present to assume that *some* effort to regulate risk now would be better than none—i.e., that lawmakers should not wait until the extent of language model risks becomes clearer. The evidence canvassed in this paper suggests that the danger of not regulating (in the face of risk opacity) *likely* outweighs the danger of trying and failing to get regulation right. The fact that risks of real harm have emerged from language model deployment suggests a need to impose legal obligations on the part of AI providers to be attentive, responsible, and transparent with independent, public officials tasked with overseeing the mitigation of these risks.

Regulation involving either licensing or self-monitoring would serve this purpose; which of them strikes a better balance between caution and progress is a secondary consideration and beyond the scope of this paper. The inquiry in this paper was meant to highlight concerns about risk assessment that will play a role in any conceivable regulatory model that aims to mitigate or avoid harm. The question of how best to regulate language

Council Directive 87/357/EEC, [2023] OJ, L 135/1 which defines "safe product" in art 3(2) as:

> any product which, under normal or reasonably foreseeable conditions of use, including the actual duration of use, does not present any risk or only the minimum risks compatible with the product's use, considered acceptable and consistent with a high level of protection of the health and safety of consumers.

model risks may become clearer in time as we learn more about the nature and extent of those risks.