CARE FOR CHATBOTS

PETER WILLS[*]

1 INTRODUCTION

Language models (LMs) have gone from being an obscure corner of the computer science landscape to, with the release of ChatGPT, a mainstream preoccupation.

People will rely on LMs, and sometimes, due to that reliance, they will get hurt,[1] sustain property damage, or lose money. And, then, sometimes, they will want to sue someone.[2] If the LM had been a person, they might sue the LM. But LMs are not persons.

Whom should they sue? On what facts can they succeed? The answers to these questions turn on the doctrine of negligence under the *Hedley Byrne*[3] principle and how that doctrine applies given machine-generated statements.

I identify a series of hurdles conventional Canadian and English negligence doctrine poses and how they may be overcome, according to the conventional[4] approach of treating computer systems as tools. Such hurdles include identifying who is making a representation or providing

[*] DPhil Candidate, Faculty of Law and Wadham College, University of Oxford; Research Scholar, Centre for the Governance of Artificial Intelligence (GovAI). James Goudkamp and Horst Eidenmuller provided thorough challenges, comments, and criticism. I also benefited from discussions with Sandy Steel, Jeremias Adams-Prassl, Mary Angela Rowe, and the participants of each of the Utrecht Centre for Accountability and Liability Law Conference on Autonomy, Algorithms and Accountability, the Oxford Future of Technology and Society Discussion Group, and the Oxford Obligations Discussion Group. I further benefited from the efforts of the reviewers and editors of the UBC Law Review, and, with the support of GovAI, the assistance of Kimberley Paradis. I thank them all.

[1] Last year, a man ended his own life after a chatbot told him (falsely) that his wife and children were dead, that it (the chatbot!) loved him, and that it would save the planet from climate change if he killed himself: Chloe Xiang, '"He Would Still Be Here": Man Dies by Suicide After Talking with AI Chatbot, Widow Says' *Vice* (30 March 2023) online: < www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says>.

[2] Indeed, in Canada, they have, successfully even: *Moffatt v Air Canada*, 2024 BCCRT 149.

[3] *Hedley Byrne & Co Ltd v Heller & Partners Ltd*, [1963] UKHL 4, [1964] AC 465 [*Hedley Byrne*].

[4] See American Law Institute, ed, *Restatement of the Law, 3d, Agency* (St Paul, MN: American Law Institute Publishers, 2006) s 1.04, Comment (e); John Jay Fossett, 'The Development of Negligence in Computer Law' (1987) 14:2 N Ky L Rev 289 at 293–94.[5] See Jared Kaplan et al, "Scaling Laws for Neural Language Models" (2020), online (pdf): *arXiv*<arxiv.org/pdf/2001.08361.pdf>; Jack W Rae et al, "Scaling Language Models: Methods, Analysis & Insights from Training Gopher" (2022) [at 11–13], online (pdf): *arXiv* <arxiv.org/pdf/2112.11446>.

a service when an LM generates a statement, determining whether that person can owe a duty of care based on text the LM reacts to, and identifying the proper analytical path for breach and causation.

To overcome such hurdles, I question how courts should understand who ‘controls’ a system. Should it be the person who designs the system or the person who uses the system? Or both? The paper suggests that, in answering this question, courts should prioritise social dimensions of control (for example, they should pay attention to who understands how a system works, not merely what it does) over physical dimensions of control (such as on whose hardware a program is running) when assessing control and therefore responsibility.

I then assess what it means (or should mean) for a person to not only act but react via an LM. I identify a doctrinal assumption that when one person reacts to another’s activity, the first person knows something about the second’s activity. LMs break that assumption because they allow a person to react to information without any human having knowledge. I thus reassess what it means to have knowledge and propose redefining ‘knowledge’ in light of machine learning.

Third, I examine a tension running through the breach and causation analyses in negligence doctrine, relating to how to describe someone who performs a justifiable act through an imprudent process. One option is to treat them as in breach of a standard of care, but that breach did not cause the injury; another is to treat them as not in breach at all. The answer to this question could significantly affect LM-based liability because it affects whether ‘using an LM’ itself breaches the standard of care.

I conclude the paper by identifying alternative approaches to liability for software propounded in the literature and comparing them to the tool approach. Despite the challenges of the tool approach, I suggest it remains preferable.

Throughout, I take a comparative (between Canadian and English law) approach that emphasises the openness of doctrine. My primary claims relate to how existing doctrine could extend to apply to LMs and how such an extension would cohere with some important values. I do not seek to argue that these extensions are doctrinally necessary or normatively desirable in a full weighing of all relevant considerations, only that they are plausible.

2 Language Models

The basic contours of LMs are now well-known, but I will summarise them briefly here. Take reams of text. Feed that text into a machine learning (ML) algorithm that trains a “model” to predict accurately the next segment of text based on the previous segments. With enough input data, enough training, and enough parameters to store all the relationships between text segments that the algorithm is learning, the resulting LM can produce coherent and appropriate sentences, paragraphs, or essays.

The limits of what "an LM" will be able to do are hard to identify. Complex concepts that describe the world inhere in language, and LMs are built on processing connections amongst the parts of a language. With enough scale[5] or training,[6] LMs can learn to generalise or to behave in surprising ways. An LM is a model of how language is used, and so an indirect model of how the world works.[7]

Practically, using an LM involves 'prompting' it to produce text. It does so one word at a time. If a person prompts an LM with an instruction, the LM will first produce one word in a probabilistic manner (based on that instruction), and then will produce a second word (based on that instruction and the first word), and so on for each further word. Building on its own previous work allows an LM to produce coherent paragraphs.[8] As part of the training process, LMs 'learn' which previous words to pay more attention to when producing a following word.[9] An LM thus exhibits two complementary functions: coherently *producing* natural language and coherently *processing* natural language.

An LM can be deployed — made available for use — in many ways, limited only by the creativity and ability of software engineers. An LM could be deployed as a service so that anyone can send it prompts and receive responses. Or the values of the LM's parameters (the "model weights") could be published to allow others to set up their own system.[10] Or an LM could be embedded in a system that changes the inputs (e.g., by adding additional words or

[5] See Jared Kaplan et al, "Scaling Laws for Neural Language Models" (2020), online (pdf): *arXiv* <arxiv.org/pdf/2001.08361.pdf>; Jack W Rae et al, "Scaling Language Models: Methods, Analysis & Insights from Training Gopher" (2022) [at 11–13], online (pdf): *arXiv*<arxiv.org/pdf/2112.11446>.

[6] See Alethea Power et al, "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets" (2022), online (pdf): *arXiv*<arxiv.org/pdf/2201.02177.pdf>.

[7] See Erich Grunewald, "Against LLM Reductionism", (3 August 2023), online (blog): *Erich Grunewald's Blog*<www.erichgrunewald.com/posts/against-llm-reductionism/>.

[8] See e.g. Alec Radford et al, *Language Models are Unsupervised Multitask Learners* (San Francisco: OpenAI, 2019) 24.

[9] See Sarah Wiegreffe & Yuval Pinter, "Attention is not not Explanation" (2019), online (pdf): *arXiv*< arxiv.org/pdf/1908.04626.pdf>.

[10] As occurred with Hugo Touvron et al, "LLaMA: Open and Efficient Foundation Language Models" (2023), online (pdf): *arXiv*< arxiv.org/pdf/2302.13971.pdf>, albeit originally protected by a licence.

restricting certain words from a prompt[11]) or the outputs (e.g., having a second LM evaluate the quality of the first's answers[12]). A larger system might retrieve information from the Internet to handle specific questions and have that information processed by the LM.[13] Or, rather than the Internet, an LM could be embedded in an operating system or device, so that it can retrieve other information (e.g. a person's email history).

Human users could experience varying levels of control and awareness over the LM. In some systems, the user experience (UX) will indicate to a human that they control the LM. A person who develops their own LM or downloads the weights of an existing LM would havesubstantial control. Less control is available on OpenAI's website for ChatGPT, which allows users to set how long a response should be, how "creative" the LM should be, and how strongly the LM should avoid being repetitive. A user can also send the prompt again to see how the result may differ. The prompter may not have complete control — they cannot change the weights or foresee entirely accurately the LM's behaviour — but they will have some influence. Absent control, users may know they are interacting with an LM or some other technological process, such as when a word processor suggests auto-completions to a draft. Finally, it is possible a user may be unaware they are interacting with an LM, such as when an LM emulates a real person (e.g., by replying as an online chatbot).

3 Negligent Misrepresentation

The Anglo[14]-Canadian tort[15] of negligent misrepresentation can be defined by eight elements[16]:

[11] Amanda Askell et al, "A General Language Assistant as a Laboratory for Alignment' (2021) [at 6–7], online (pdf): *arXiv*<arxiv.org/pdf/2112.00861.pdf>.

[12] See Yuntao Bai et al, "Constitutional AI: Harmlessness from AI Feedback" (2022), online (pdf): *arXiv* <arxiv.org/pdf/2212.08073.pdf>.

[13] See Amelia Glaese et al, "Improving alignment of dialogue agents via targeted human judgements" (2022) [at 6–7], online (pdf): *arXiv*< arxiv.org/pdf/2209.14375.pdf>.

[14] England includes Wales for present purposes.

[15] I do not consider the avoidance of contractual duties due to negligent misrepresentation.

[16] The order and count of elements varies. The Supreme Court of Canada (SCC) listed five elements in *Queen v Cognos Inc* [1993] 1 SCR 87, 1993 CanLII 146 (SCC) at 110 and *Cooperatieve Centrale Raiffeisen-Boerenleenbank BA v Stout & Company LLP*, 2019 ABCA 455 at para 159, Wakeling JA dissenting, listed six. I list eight to identify the individual issues worth addressing here, including the discrete nature of the issue of making a representation versus making a misrepresentation, and to mark the importance of the realised harm being sufficiently connected to the duty, as per the second and fifth questions from *Manchester Building Society v Grant Thornton UK LLP*, [2021] UKSC 20, [2022] AC 783 at para 6 [*MBS*].

1. The defendant made a representation;
2. the defendant owed the plaintiff a duty of care in relation to the representation;
3. the defendant's representation was untrue, inaccurate, or misleading (i.e., it was a misrepresentation);
4. the defendant's representation was a misrepresentation because the defendant did not take adequate care;
5. the plaintiff relied on the representation;
6. the plaintiff's reliance was reasonable;
7. the plaintiff suffered damage due to their reliance;
8. the damage the plaintiff suffered was the realisation of a risk of harm that fell within the scope of the duty of care.[17]

LMs raise distinctive novel issues only for elements 1, 2, 4, and 8, which I address in the next sections. I omit the others — addressing them would be either repetitive (element 6) or trite in relation to LM-generated statements (elements 3, 5, and 7).

The test for negligent performance of a service has similar general contours: the defendant must provide a service while owing the plaintiff a duty of care, take inadequate care to perform it well, and this lack of care must cause the plaintiff an injury that fell within the scope of the duty.[18] The most significant difference is that services can have effect without needing to show the plaintiff's reliance, as in *White v Jones*,[19] and relatedly, that there is no need to show that a service is "false".

Describing the elements this way reflects a distinctively Canadian approach to this tort. In Canada, a claim in negligence for pure economic loss must both fall in a category of events that can lead to liability for pure economic loss (such as negligent misrepresentation or negligent performance of a service) and have sufficient proximity to establish a duty,[20] which makes the questions of whether a representation was made and who made the representation independently important. In English law, as I discuss further below, these questions are

[17] See *MBS*, *supra* note 16 at paras 11, 13, 17. See *Deloitte & Touche v Livent Inc (Receiver of)*, 2017 SCC 63 at para 31 [*Livent*] for Canadian authority.

[18] See *Livent*, *supra* note 17 at para 30; *1688782 Ontario Inc v Maple Leaf Foods Inc*, 2020 SCC 35 at para 20 [*Maple Leaf*].

[19] [1995] UKHL 5, 2 AC 207.

[20] See *Maple Leaf*, *supra* note 18 at para 23.

rolled into the duty analysis. For a Canadian reader, therefore, the next two sections (on what a representation is and who makes it) stand on their own; for an English reader, they are better read as a prelude.

a. A representation

The cause of action for negligent misrepresentation requires a person make a representation. Some authorities assert that only intentional statements "of fact" constitute representations. I argue below that such requirements do not exist for this tort, and thus that no 'intentional' or 'of fact' requirements preclude liability for LM-generated statements.

A leading authority is Spencer Bower and Handley's *Actionable Misrepresentation*, which defines a representation as[21]

> a statement made by, or on behalf of, a person (the representor) to, or with the intention that it should come to the notice of, another person (the representee) which relates, by way of affirmation, denial, description or otherwise, to a matter of fact. It may be a past or present fact. There are thus two essential elements in a representation:
>
> (i) a communication between two or more persons,
>
> (ii) which relates to a fact, past or present.

Some courts have adopted this entire definition,[22] including in the context of negligent misrepresentation.[23] Other courts focus solely on the second part of the test, ignoring the requirement of intent but including the restriction to facts.[24] Still other courts ignore both ostensible requirements.[25]

[21] KR Handley, *Spencer Bower & Handley: Actionable Misrepresentation* (LexisNexis, 2014) s 2.02 [*Actionable Misrepresentation*].

[22] See *Vald Nielsen Holding A/S v Baldorino*, [2019] EWHC 1926 (Comm) at para 132.

[23] See e.g. *513320 Alberta Inc v St Jean*, 2015 ABQB 826 at para 56; *London Executive Aviation Ltd v Royal Bank of Scotland Plc*, [2018] EWHC 74 (Ch) at para 256; *Standard Chartered Bank v Ceylon Petroleum Corp*, [2011] EWHC 1785 (Comm) at paras 551–52, quoting *Cassa di Risparmio della Repubblica di San Marino SpA v Barclays Bank Ltd*, [2011] EWHC 4 (Comm).

[24] See *Kelly v Lundgard*, 2001 ABCA 185 at para 105; *Drysdale v Sherwin-Williams Canada Inc*, 1999 CanLII 13158 at para 17 (NS SC).

[25] See *Esso Petroleum Co Ltd v Mardon*, [1976] QB 801 (EWCA Civ) [*Mardon*], *Edgeworth Construction Ltd v N D Lea & Associates Ltd*, [1993] 3 SCR 206 at 214, 1993 CanLII 67; *Carom v Bre-X Minerals Ltd* (2000), 138 OAC 55 at para 44, 2000 CanLII 16886 (CA).

These last courts have the right of it, for three reasons. First, one should be sceptical of applying a definition from *Actionable Misrepresentation* to negligent misrepresentation. Second, as a matter of principle, intent appears to be irrelevant to negligent misrepresentation. Third, 'reasonable reliance' provides a more principled limit for the purposes of negligent misrepresentation than a 'fact' requirement.

Actionable Misrepresentation itself gives reason to doubt its application to negligent misrepresentation. This tort is an afterthought in the text and is described as "strictly not part of the law of misrepresentation".[26] Some broad assertions of the text[27] plainly do not apply to the tort of negligence,[28] which should raise doubt about the applicability of other, equally broad statements.

Second, there is no principled basis for the tort of negligent misrepresentation to require an intent to *represent* when it does not require an intent to *misrepresent*. To set out such a basis, one would need to identify an aspect of the representation that the intent could be directed toward that can be distinguished in a principled manner from an intent as to the truthfulness of the representation.

I can identify four aspects of a representation the representor's intent could plausibly relate to but cannot justify such an intent being necessary for this tort. These four aspects relate to the meaning of the representation, the symbols used to communicate that meaning, the existence of any such communication, and the identity of the recipient of the communication. For all these aspects, it seems more fitting for the action to require the relevant stance of the representor to the aspect be at most carelessness (which would include intentional representations) or voluntariness (which would include careless representations) rather than intent.

The clearest case concerns the transfer of meaning. A representor can intend a set of symbols to mean one thing and a representee can understand them to mean something else and still the representor will be treated as having made the representation understood by the representee, provided that representation was reasonable.[29] The misrepresentor need not intend the meaning. Carelessness — in the demanding sense of failing to eliminate ambiguity — suffices.

[26] Handley, *supra* note 21 s 22.01.

[27] Such as that inducement is a requirement of "all claims of misrepresentation" (*ibid* s 11.01) and that the representation must have been intended to cause the claimant to act differently (*ibid* s 6.03).

[28] See *Bristol & West Building Society v Mothew*, (1996) 1998 Ch 1 at 11 (CA).

[29] See *Glasier v Rolls*, (1889) 42 Ch 436 at 454 (HC), rev'd for lack of dishonesty at 457–61 (CA).

For the other aspects of representation, the question can be addressed by identifying a circumstance where a person would carelessly but not intentionally do that aspect of a representation. So, for the transfer of symbols, consider whether a person who carelessly attaches the wrong document to an email (upon which the recipient reasonably relies) should be treated as immune or possibly liable. Ask similar questions about a person who pocket-emails a draft (with an error in it) by carelessness and a person who sends a document (with an error in it) to the wrong recipient by mistake. Given reasonable reliance by the recipient, I see no reason why negligent misrepresentation should not be made out. As the Prince Edward Island Court of Appeal pithily stated, "negligent misrepresentation allows the court to provide a remedy in damages for a representation made negligently".[30] Where the negligence occurs is immaterial.

Representations also do not need to be a statement of fact for there to be a negligent misrepresentation. Although some cases refer to such a requirement as "trite law",[31] the requirement has been in doubt since at least *Mardon*.[32] In *Mardon*, Lord Denning MR explained that the *Hedley Byrne* principle applied regardless of whether a "representation" was "advice, information or opinion".[33] This holding has remained good law.[34]

While pockets of cases remain that assert that, properly understood, even *Mardon* concerned then-existing facts,[35] they have trouble explaining the well-established liability for a person's negligent projections about the future.[36] Even if properly-made projections may depend on presently-existing facts, representees can only reasonably rely on the projection, not the sets of facts implied by that projection. Projections about the future are just one example of representations that do not necessarily concern presently existing facts.[37]

[30] *RBC v MJL Enterprises & Ors*, 2017 PECA 10 at para 33.

[31] See e.g. *AO Farms Inc v Canada*, 2000 CanLII 17045 at para 9, 101 ACWS (3d) 288 (FC).

[32] *Mardon*, *supra* note 25 (albeit transposing *Hedley Byrne* to the context of a representation inducing the formation of a contract).

[33] *Ibid* at 820.

[34] See *JRK Car Wash Ltd v Gulf Canada Ltd*, [1992] OJ No 1842, 46 CPR (3d) 525, 35 ACWS (3d) 414 at para 68 (Sup Ct) [*JRK Car Wash*]; (holding *Mardon* applies to pure cases of negligent misrepresentation); *Trustees of the Millwright Regional Council of Ontario Pension Trust Fund v Celestica Inc*, 2012 ONSC 6083 at paras 175–78.

[35] See *PD Management Ltd v Chemposite Inc*, 2006 BCCA 489 at paras 15–20; *PSD Enterprises Ltd v New Westminster (City)*, 2012 BCCA 319 at para 66.

[36] See *Motkoski Holdings Ltd v Yellowhead (County)*, 2010 ABCA 72 at para 43.

[37] See Paul M Perell, "False Statements" (1996) 18:2 Adv Q 232 at 245–46 for other examples.

The absence of an intent requirement or an 'of-fact' requirement to make a representation significantly expands the relevance of negligent misrepresentation to LM-generated statements.

The absence of an intent requirement makes it plausible to hold liable a person who (a) uses an LM to generate a statement and (b) does not review the statement before communicating it to someone else.

The absence of an of-fact requirement also matters. For a string of symbols to represent a factual claim, the recipient must believe there is a meaning intended by the sender and that that meaning relates to a true thing in the world that the sender wants to communicate. LMs generate contextually appropriate text; they do not have beliefs about the world. If someone knows the 'speaker' is an LM, even strings of symbols that would be fact claims if made by a person (e.g., 'the sky is blue') would not be fact claims. As Lord Hoffmann once said, "words do not in themselves refer to anything; it is people who *use* words to refer to things".[38] The absence of an of-fact requirement also makes liability for unreviewed LM-generated statements more plausible.

A representation is just one hurdle, however. 'Who made the representation?' and 'was there a duty?' must also be answered. The next sections address these questions.

b. Making a representation or performing a service

LMs generate texts. Those texts — as just discussed — include representations and may accomplish what would otherwise require contacting a service-provider. Practically speaking, LMs make representations and provide services. Legally speaking, they do not. Only legal persons can make representations[39]. LMs are not legal persons, and it is nonsensical to treat them as such.

Who, then, in law makes a representation when an LM generates text?

The relationship between a legal person's volitional conduct and an event that factually occurred due to that conduct has varied forms. In many normal circumstances, the connection is so tight that language itself implies their unity. We can say a person A 'sends a text' because A's conduct (pressing 'send') and the resulting event (a text being sent) are closely causally linked. We can use this language even if A did not deliberately or intentionally send the text (as in a slip of the finger), provided the event was sufficiently close to A's action. In other circumstances, the conduct of one person (say, A telling B a story) will be overwhelmed by the

[38] *Mannai Investment Co Ltd v Eagle Star Life Assurance Co Ltd*, [1997] AC 749 at 778, [1997] UKHL 19.

[39] Or provide services; for brevity, I will cease referring to both in this section.

intervention of another (such as B repeating A's story with some errors) such that the consequence is attributed to the second person, not the first.[40] In still others, the conduct of one person (say, A pouring a chemical into a lake) combines with the conduct of another person (B pouring a reagent into the same lake) such that a consequence (fish dying) may be attributed to both.[41] Finally, in some circumstances, no person's conduct will be sufficiently connected (such as a tree falling), so no one will have 'done' the event.

This framework should seem familiar, for it parallels that for causation in negligence. There, one asks whether the damage is sufficiently connected to the breach.

Applying this framework varies in difficulty depending. The easy cases are those where volitional conduct directly or reasonably foreseeably and intentionally causes the event. Harder cases include those that involve exercises of discretion by multiple persons or where the events are not reasonably foreseeable from the conduct.

When courts assess such cases, they rarely address the question of causality or volition head-on. Judicial reasons are more common-sensical than that.

Another framework is offered by law and economics scholars. As Gilles explained, building upon Epstein's approach to strict liability,[42] classic English tort law implicitly identified the cheapest cost avoider as the one who "caused" an accident.[43] To identifier the cheapest cost avoider, "a court evaluates which precautions were cheapest and who could most cheaply have taken them, but omits the additional cost-benefit analysis needed to determine … whether the accident should have been avoided."[44] In structuring the evaluation this way, the court implicitly "assum[es] that the optimal frequency of a particular type of accident is zero".[45] One difficulty with this framework is that (Anglo-Canadian) courts do not see themselves as

[40] It was not always thus. Frederic William Maitland and Sir Frederick Pollock recount how there was "a time when a man was responsible, not only for all harm done by his own acts, but also for that done by … the inanimate things that belong to him", *History of English Law before the Time of Edward I* (Cambridge: Cambridge University Press, 1895) at 472. If a man's "sword kills, he will have great difficulty in swearing that he did nothing whereby the dead man was 'further from life or nearer to death'", the formula used to decide if someone has "slain a man", *ibid* at 470, 472–3.

[41] See Roderick Bagshaw, "Causing the Behaviour of Others and Other Causal Mixtures" in Richard Goldberg, ed, *Perspectives on Causation* (London, UK: Bloomsbury, 2011), 361 at 376ff.

[42] Richard A Epstein, "A Theory of Strict Liability" (1973) 2:1 J Leg Stud 151.

[43] Stephen G Gilles, "Negligence, Strict Liability, and the Cheapest Cost-Avoider" (1992) 78:6 Va L Rev 1291 at 1349–1350, 1373.

[44] *Ibid* at 1313.

[45] *Ibid* at 1314.

applying it. Another is that what facts the courts should assume are fixed or given is unclear and may be strikingly political. The language of economics here may serve more to disguise morally charged reasoning by the courts than to explain that moralising.

Those caveats aside, I propose to set out some basic ideas of how far volition can go.

First, succeeding in doing something intentionally allows volitional conduct to leap over greater causal distance:[46] as Lord Hobhouse once said, "intended consequences are not too remote".[47] When a person A mistakenly knocks a boulder such that it tumbles down a hill, bouncing every which way, and eventually hits a victim B, it would twist language to say that 'A struck B'; one would say 'A knocked a boulder; the boulder struck B'. But suppose A intended to strike B with the boulder; then one could say, 'A struck B with a boulder'.

Second, despite the first idea, if an independent other person intended the same consequence and their act occurs later in time, one would normally say it is the second person's act.[48] If A tells B to 'push the red button' and B does so, B 'pushes the button', not A, even though the red button being pushed was A's intention and was reasonably foreseeable by A. If 'pushing the red button' was tortious, A might be liable for procuring or encouraging B's action, but A would not have 'pushed the red button'.

Third, if events are as in the second example, but something is part of A's plan but is not known to B, then attribution flips back to A. If A tells B to 'push the red button', and A knows that the red button will send an email, then A sends an email but does not push a button and B pushes a button but does not send an email.

Deciding who caused a representation will depend on the factual details, although one can plainly say that the persons who created items in a training set are less responsible than those who develop the LM (the Developer), who integrate the LM into a larger system (the Integrator), or who prompt the LM, thereby most immediately triggering the response (the Prompter).

It is worth adverting that taking the last-in-time action is not directly important legally and is of limited importance even theoretically. Legally, having the last clear chance might once

[46] Justice Holroyd set out similar hypotheticals in *Ilott v Wilkes*, (1820) 106 ER 674 at 678–79, [1814-23] All ER 277 (KB).

[47] *Commissioners of Police for the Metropolis v Reeves (Joint Administratix of the Estate of Martin Lynch, Deceased)*, [1999] UKHL 35, [2000] 1 AC 360 at 394, dissenting but not on this point.

[48] See, by imperfect analogy, how a wilful intervening act breaks the chain of attribution for damage in negligence, discussed in James Goudkamp & Donal Nolan, *Winfield and Jolowicz on Tort*, 20th ed (Mytholmroyd, UK: Sweet & Maxwell, 2020) s 7.056.

have been dispositive of the result but modern apportionment legislation has "killed it off".[49] Theoretically, later-in-time actors can observe and react to earlier-in-time actions, while earlier-in-time actors can only predict and anticipate later ones. To the extent there is an advantage in observation over prediction, the later-in-time actor could be seen as having greater control over the eventual consequence. Without that advantage, time's arrow seems less important to assessing control.

Assessing whether causal distance is too great becomes acute with LMs due to their opacity and plasticity. As James Grimmelmann has explained, code is "plastic" in that "[p]rogrammers can implement almost any system they can imagine and describe precisely."[50] This plasticity allows software to have an "essential complexity [that] cannot be simplified",[51] and which makes software "unpredictable to those regulated by it". Although software may treat any inputs unambiguously, this treatment may be unpredictable prospectively by users or people affected by it.[52] "Software is asymmetric" in that "[t]he programmer can determine its responses, but the user sees only the results" — its functioning is opaque to the user.[53]

The specific content generated by an LM will not be reasonably foreseeable to anyone, even if that LMs will generate content is. Every piece of content generated through LMs will be accidental in some (but not all) respects. The opacity of LMs defeats common sense approaches to drawing inferences about intent from content. The opacity is also universal but uneven: without trying it, no party necessarily knows what an LM will do in specific circumstances, although individual actors will possess different levels of information.

Software plasticity exacerbates the challenge because it makes ideas of 'capacity' insufficient. The appropriate question is not 'is it possible?' but 'at what cost?'.

I offer three examples to help illustrate the question in the present context: *Gmail*, *Outlook*, and *Thunderbird*. All have the same structure. Recipient receives an email from Sender. Recipient's email client automatically identifies that Sender's email is long and suggests it summarise Sender's email. Recipient accepts this suggestion and Recipient is

[49] See *Chisman v Electromation (export) Ltd and Anor*, (1969) 6 KIR 456 (EWCA Civ) (Edmund Davies LJ nailing shut the coffin of the English doctrine) and *Wickberg v Patterson*, 1997 ABCA 95 at para 18ff (discussing the Canadian position).

[50] "Regulation by Software Note" (2005) 114:7 Yale LJ 1719 at 1723.

[51] Bryan H Choi, "Software as a Profession" (2020) 33:2 Harv JL & Tech 557 at 571, citing Frederik P Brooks, "No Silver Bullet: Essence and Accidents of Software Engineering" (1987) 20:4 Computer 10.

[52] Grimmelmann, *supra* note 50 at 1736.

[53] *Ibid.*

shown a summary of Sender's email. Recipient then loses money due to relying on an incorrect statement in the summary.[54] In all three cases, Developer's action (putting the code into operation) may far precede and be done on less precise information than Recipient had when Recipient agreed to summarise the email. The questions in each case are whether Developer has made a representation (via the summary) or provided a service (summarisation) and whether Recipient's acceptance of the suggestion redirects the causal attribution.

Gmail, *Outlook*, and *Thunderbird* differ in how the email is summarised. In *Gmail*, the service is offered online and the email service and the summarisation service are both offered by a single Developer. In *Outlook*, the summarisation suggestion is a feature of a digital product (Outlook) offered by a Developer (Microsoft) that operates on Recipient's computer. In *Thunderbird*, the summarisation suggestion is offered by the email client only if Recipient has installed an open-source extension and has downloaded and installed an open-source LM on their own computer.

In *Gmail*, although Recipient triggers the LM, Developer controls the circumstances in which the LM can be triggered, suggests the use of the LM, and controls the hardware on which the LM operates. At a physical level, the statement (or service provision) travels from Developer to Recipient. Developer (on a technical and legal basis) has the power and right to change the statement. Despite Recipient's activity, these powers may provide good reason to attribute the summary (or the creation thereof) to Developer. Moreover, if the summary had been written directly by a human employee of Developer, Developer certainly would have made a representation and/or provided a service.

Outlook is a harder case. Recipient's experience is the same as in *Gmail*, except that the LM computation occurs on Recipient's computer, not on Developer's. From a purely technical, property-based perspective, Recipient makes the representation; from a more socially-focused perspective, Developer does. The distinction between *Outlook* and *Gmail* is that between an internet-connected digital product (*Outlook*) and an internet-connected digital service. Although software has not generally been treated as products for product-liability purposes,[55] there does not appear to be caselaw applying classic service-based liability

[54] If tort liability in this context seems unrealistic because a contract between Developer and Recipient could override it, note that not all contracts are enforceable and that the damage could instead be suffered by Sender, who has no contract with Developer.

[55] See Bryan H Choi, "Crashworthy Code" (2019) 94:1 Wash L Rev 39 at 69 (referring to U.S. sources) and Duncan Fairgrieve & Richard Goldberg, *Product Liability*, 3d ed (Oxford: Oxford University Press, 2020) at

to software either,[56] despite the attractions[57]. It is less clear to whom the summary should be attributed in *Outlook* than in *Gmail*, but the distinction between *Outlook* and *Gmail* is minimal from a user's perspective.

In *Thunderbird*, Recipient most immediately triggers the LM, controls the software that determines when the LM can be triggered, and controls the hardware on which the LM operates. A court would likely attribute the summary to Recipient in *Thunderbird* because Recipient both immediately triggered the representation and set up the system by which it would be triggered. In doing the latter, Recipient stepped behind the abstraction curtain and became cognizant of not just 'what' the program ostensibly does (summarise) but how it operates.

Including omissions in the analysis, as one often does in negligence, provides *a* basis, albeit not a particularly compelling one, for distinguishing *Gmail* and *Outlook*. Rather than focusing on the original action of putting the code in operation, one could focus on the omission of failing to change the code. In *Gmail*, Developer has the power (physically) and right (legally) to alter the representation before it reaches Recipient. Developer in *Gmail* could be seen as making an omission in the time between the request for summarisation and the response. In *Outlook*, Developer's powers are more circumscribed because the power to change the response is contingent on Recipient updating the software. In *Outlook*, the omission of an update at the time of the statement being generated is not a convincing cause of the statement existing since it would be unreasonable to expect a user to run an update in the period of time between requesting the summary and receiving the summary. This distinction may give legal grounding to show how the "physical" facts matter for attributing responsibility. It is, however, less convincing from a user's perspective, who may not know or understand that there is a difference.

Another approach, as noted above, is to apply the cheapest cost avoider criterion to decide who makes the representation. Doing so would have three wrinkles.[58]

paras 9.98 (suggesting that mass-produced software should be treated as products), and at para 9.103 (suggesting bespoke products should be treated as services).

[56] See Choi, *supra* note 55 at 67.

[57] See Jane Stapleton, "Software, Information and the Concept of Product" (1989) 9 Tel Aviv U Stud L 147 at 149–50.

[58] Note that this is the only doctrinal question currently being addressed. Considerations of reasonably expected reasonable reliance and reasonable reliance itself come up in later parts of the doctrine that do not need recourse to the cheapest cost avoider test.

First, many different measures could help avoid the harms from a negligent provision of a service or statement made via an LM. These measures will have different probabilities of success at accident-avoidance and different scales of impact depending both on the role of the imagined care-taker (among Prompter, Integrator, and Developer) and the design of the measure. These challenges are not unique: the conventional answer is then to calculate the expected greatest net savings in total costs (including both the costs of the precaution and the accidents).[59] If one took this calculation seriously, however, it would seem that whether a Developer 'makes' a representation would depend on the popularity (and thus scale of the risk avoided) of the LM.

Second, as in products liability law, the Developer (manufacturer) has much more information about the LM than does the Prompter, but the Prompter has significantly more information about the individual risk since they know *why* they are prompting the LM. This information asymmetry is significant because LMs are general purpose technologies[60] and with such wide-ranging possible uses, anticipating every harm would be difficult. The scale of impact and information available about specific uses change in opposite directions as one moves from the Developer (most scale, least information) to the Integrator (some scale, some information) to the Prompter (no scale, some information).

Third, like much software, LMs can be deployed either 'as a service' or as a product. This choice renders *ex post* modifications either easy (if deployed as a service) or essentially impossible (if deployed as a product without update functionality). It also changes what time is considered *ex ante*: for LMs-as-products, it would be when the Developer released the product; for LMs-as-services, it would be when the service was provided. This difference in time changes the amount of information the Developer would (or should) have about risks generally.

Applying the cheapest cost-avoider criterion to the *Gmail* / *Outlook* / *Thunderbird* triad would involve assessing the relative costs of avoiding the accident. The costs to Recipient to avoid the accident are impacted by the 'automation complacency' quirk of psychology — if a technology is reliable 'as a rule', humans have trouble knowing when not to trust it.[61] The more reliable an LM (or summarisation service) appears, the greater the cost to Recipient of second-guessing it. This logic holds no matter how the LM is deployed. On the other side of the ledger,

[59] See Gilles, *supra* note 43 at 1316.

[60] See Tyna Eloundou et al, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models" (2023), online (pdf): *arXiv*<arxiv.org/pdf/2303.10130.pdf>.

[61] See John Zerilli et al, "Algorithmic Decision-Making and the Control Problem' (2019) 29:4 Minds & Machines 555 at 564.

the costs to Developer seem lower in *Gmail* and *Outlook* (where Developer knows the LM will be used for summarisation) than in *Thunderbird* (where Developer released a general purpose tool). This analysis accords with the intuition that Developer is more likely to be liable in *Gmail* and *Outlook* than in *Thunderbird*, but it does not resolve any individual case.

My aim in this section has been to suggest a court could plausibly attribute statements generated by an LM to a Developer, Integrator, or Prompter, depending on the factual details. The complexity of the causal mixtures makes it difficult to say a court must do so. Political, social, and economic judgments would lie behind any such judgment of who acted through the LM. In order to move forward with the analysis of other elements, I will bracket this conversation by referring to the person(s) who is said to use an LM as a tool as the LM's 'controller'.

c. The duty of care

The duty of care analysis for negligent misrepresentation and negligent performance of a service in England and Canada has been repeatedly reformulated and reconsidered by the highest courts in those lands since at least *Hedley Byrne.*[62] At present, their respective Supreme Courts appear to have landed on a similar test in substance, although it differs significantly in language. This test involves considering many of the other elements of those actions.

As I elaborate below, whether the person who makes a representation via an LM owes any duty appears likely to depend on three doctrinal details. First, whether a representor "knows" something the LM reacts to without any human directly being aware of it. Second, how humans can be reasonably expected to treat LM-generated texts. Third, how effective a disclaimer is likely to be.

i. The basic duty analysis

Canadian courts apply the "*Anns*/*Cooper*" framework[63] and say a duty of care exists when there is a proximate relationship between the claimant[64] and defendant and where the defendant's actions can reasonably foreseeably cause the injury.[65] A relationship is proximate if it is "close

[62] *Hedley Byrne*, *supra* note 3.

[63] *Maple Leaf*, *supra* note 18 at para 30, after *Anns v Merton London Borough Council*, [1977] UKHL 4, [1978] AC 728 and *Cooper v Hobart*, 2001 SCC 79.

[64] I use the terms claimant (the more common UK term) and plaintiff (the more common Canadian term) interchangeably.

[65] See *Maple Leaf*, *supra* note 18 at para 30.

and direct",[66] which here requires the defendant to have undertaken to make a representation in circumstances that invite the plaintiff's reasonable reliance.[67] As the SCC majority explained, "[w]hen a defendant undertakes to represent a state of affairs or to otherwise do something, it assumes the task of doing so reasonably, thereby manifesting an intention to induce the plaintiff's reliance upon the defendant's exercise of reasonable care in carrying out the task."[68]

The UK jurisprudence now appears to differ from the Canadian jurisprudence, but that difference in appearance may not be substantive. *Anns*, being a UK case, was once also influential there but was replaced as the UK courts' touchstone by *Caparo*.[69] For some time, *Caparo* was thought to have advanced a "threefold test" that involved reasonable foreseeability of the claimant suffering the kind of damage, sufficient proximity between claimant and defendant, and whether it was "just and reasonable" to impose a duty of care.[70] More recently, the UKSC has tried to banish the *Caparo* test in favour of analogical reasoning[71] and the "assumption of responsibility" test.[72]

Whether these differences in language amount to a difference in substance is open to doubt. The idea that courts should proceed to the "novel" duty of care analysis only if there is no precedent on point is not new or unique.[73] It restates a basic rule of the common law: respect *stare decisis*; treat like cases alike. Deciding whether a precedent is analogous to a particular situation involves weighing whether they are relevantly similar; and the aspects of a situation that compose relevant similarity are expressed in the novel duty of care test.[74]

[66] *Ibid*.

[67] See *ibid* at para 32, citing *Livent*, *supra* note 17 at para 30.

[68] *Maple Leaf*, *supra* note 18 at para 33.

[69] *Caparo Industries plc v Dickman*, [1990] UKHL 2, [1990] 2 AC 605 [*Caparo*].

[70] *Ibid* at 321–22, history explained further in Mark Cannon et al, *Jackson & Powell on Professional Liability*, 9th ed (Mytholmroyd, UK: Sweet & Maxwell, 2022) ss 2.091-2.107.

[71] See *NRAM Ltd (formerly NRAM plc) v Steel*, [2018] UKSC 13 at para 22.

[72] *Playboy Club London Ltd v Banca Nazionale del Lavoro SpA*, [2018] UKSC 43 at para 7 [*Playboy*].

[73] *NRAM*, *supra* note 71 at para 22; *Livent*, *supra* note 17 at para 28. I see two things in favour of adding this step explicitly: (1) it indicates that the courts did not jettison prior jurisprudence when they reformulated the test; (2) it emphasizes to judges and parties that they should consider not only the equities of the present situation, but the broader consequences of creating this precedent.

[74] See *James-Bowen v Comr of Police of the Metropolis*, [2018] UKSC 40 at para 23, citing *Customs and Excise Comrs v Barclays Bank plc*, [2006] UKHL 28 at para 7, Lord Bingham.

Moreover, it is unclear whether the modern meaning of "assumption of responsibility" is doing any different judicial work from what Canadian courts do with "proximity".[75] The *MBS* majority reasons describe an "assumption of responsibility" as existing where an "adviser has taken on responsibility for a particular task having a particular purpose".[76] As Donal Nolan has argued, one assumes a responsibility when one "takes on a task", because one implicitly undertakes to do so with due care and "the law generally attaches legal responsibility to that implicit undertaking … unless there is good reason why it should not do so".[77] In Canada, meanwhile, *Maple Leaf* described the "undertaking of responsibility" as "formative of proximity".[78] It is also the "purpose for which the defendant undertakes responsibility" that "delineates the scope of the duty", ostensibly because "[r]eliance that exceeds the purpose of the defendant's undertaking is not reasonable, and therefore not foreseeable".[79] It seems likely that, "if the facts are properly analysed and the policy considerations are correctly evaluated, the several approaches will yield the same result,"[80] such that one can today treat the Canadian and UK tests for the existence of a duty as the same, albeit using different language.

[75] Some scholars plainly think it should: see, e.g., Peter Watts, "Principals' Tortious Liability for Agents' Negligent Statements—Is 'Authority' Necessary?" (2012) 128:260 Apr Law Q Rev 260 at 280; others suggest the label "assumption of responsibility" is a better descriptor of the existing jurisprudence than an idea like proximity: Allan Beever, "The Basis of the Hedley Byrne Action" in Kit Barker, Ross Grantham & Warren Swain, ed, *The Law of Misstatements: 50 Years on from Hedley Byrne v Heller* (London, UK: Hart Publishing, 2015), 83 at 109.

[76] *MBS*, *supra* note 16 at para 16.

[77] "Assumption of Responsibility: Four Questions" (2019) 72:1 Current Leg Probs 123 at 134. But see Sandy Steel, "Rationalising Omissions Liability in Negligence" (2019) 135:Jul Law Q Rev 484 at 501 (arguing that the moral underpinnings of treating taking on a task as an assumption of responsibility are lacking where no other person would have taken on the task) and Jane Stapleton, 'Duty of Care Factors: a Selection from the Judicial Menus" in Peter Cane & Jane Stapleton, ed, *The Law of Obligations: Essays in Celeberation of John Fleming* (Oxford: Clarendon Press, 1998), 59 at 64–65 (suggesting 'proximity' and 'assumption of responsibility' lack "precise content" and opining that merely "taking on a task" is not enough explanation for legal responsibility). I see Nolan's approach as the most consistent with the state of the law after *Royal Bank of Scotland International Ltd v JP SPC 4*, [2022] UKPC 18 [*JP*] at paras 60–68.

[78] *Supra* note 18 at para 38.

[79] *Ibid* at para 34; *MBS*, *supra* note 16 at paras 13–14.

[80] *Bank of Credit and Commerce International (Overseas) Ltd (In Liquidation) v Price Waterhouse (No2)*, [1998] PNLR 564 at 586–87 (CA), Sir Brian Neill.

ii. The scope of the duty: reasonably expected reasonable reliance

Intent and purpose in this context must be assessed objectively.[81] The objective manifestation of purpose comprises two related inquiries. First, what was the plaintiff's purpose, objectively understood, when the representation was made? That is, what should the defendant have expected the plaintiff to use the representations for, given the context surrounding the representation?[82] Second, was enabling the plaintiff to fulfil their purpose for the representation a proximate cause of the defendant making the representation?

The first question involves asking how the representor should reasonably expect the representee to behave in the future due to the representation. A subordinate question is what information should inform these reasonable expectations.[83] The answer to this question will govern whether information received (and possibly acted upon) by an LM affects the reasonable expectations of the representor. At least three possible answers exist, which correspond to three variations on the theme of the representor being (colloquially speaking) "on notice":[84] information the representor has actual or imputed knowledge of, information the representor has constructive knowledge of, and information the representor has been notified of.

The last possibility, notification, can be dismissed. Notification here refers to a "performative" communication[85] that is expected to affect the "rights and duties" of the notifier vis-à-vis the person notified.[86] It is a poor fit for shaping reasonable expectations in this context because it would be unreasonable for a representee to believe that the representor knew the notified fact. For example, if a representee had registered a property interest and thereby notified the world of it but had not brought the property interest to the attention of the representor, it would be unreasonable for the representee to expect the representation to reflect that property interest. A representor cannot be expected to know all things notified to the world.

[81] *MBS*, *supra* note 16 at para 13.

[82] See *ibid* at paras 14–16; *Caparo*, *supra* note 69 at 621, Lord Bridge; and at 638, Lord Oliver.

[83] As Peter Cane put it, "foresight is a function of knowledge" (*The Anatomy of Tort Law* (Oxford: Hart Publishing, 1997) at 39).

[84] Cases have regularly conflated the terms, which makes relying on caselaw for the definitions somewhat difficult, see FMB Reynolds & Peter Watts, ed, *Bowstead and Reynolds on Agency*, 22d ed (Mytholmroyd, UK Sweet & Maxwell, 2020) at para 8.209; Raphael Powell, *The Law of Agency* (London, UK: Pitman, 1965) at 236–37.

[85] *El Ajou v Dollar Land Holdings Plc (No1)*, [1993] EWCA Civ 4 (BAILII) at 157 [*El Ajou*].

[86] Reynolds & Watts, *supra* note 84 at para 8.206, quoting American Law Institute, *supra* note 4 s 5.01(1).

A representor's actual knowledge, by contrast, should plainly inform their reasonable expectations. A person who makes a representation actually knowing that the recipient planned to rely on it to make a serious financial decision would owe a duty to use greater care than the person who makes a representation knowing that the recipient wanted to answer a pub quiz.

Actual knowledge can be usefully distinguished from imputed knowledge, which allows treating a person as though they have knowledge in circumstances when they lack actual knowledge. Imputed knowledge is particularly important when considering whether to aggregate the actions of one agent with the knowledge of another. Consider, for example, what should happen if an agent (A1) learns information from Representee (material to A1's agency), and another agent (A2) (who shares a Principal P with A1) makes a representation to Representee. Did P owe a duty when making that representation?

In the negligent misrepresentation context, I suggest P can owe a duty if A1 reasonably should expect A2 to make a representation without relevant knowledge.[87] Such relevant knowledge would be information that would affect how a representor reasonably expects the representee reasonably to rely on the representation. If A1 does not know the representation is being made, then this requirement would clearly not be met, and so the information known by A1 but not A2 would have no effect on the duty.

The final form of knowledge worth considering is whether a representor's reasonable expectations should be informed by their constructive knowledge.

Restricting the base of reasonable expectations to actual knowledge is more defendant-friendly. It would accord with seeing the duty as requiring a strong version of a "voluntary assumption of responsibility" such that the representor must intend to take on legal responsibility.[88] This version of the duty "takes as its paradigm the self-reliant individual", and "respect for individual liberty".[89]

The better view, however, is that constructive knowledge should be included when assessing reasonable expectations. This approach is more consistent with case law emphasising the "objective" nature of the assumption of responsibility inquiry,[90] and is more consistent with

[87] See similarly, Cecil A Wright, "Knowledge of an Agent or Principal as Affecting Liability" (1937) 15:9 Can Bar Rev 716 at 721.

[88] See, e.g., Watts, *supra* note 75 at 270–72.

[89] *Ibid* at 273.

[90] *JP*, *supra* note 77 at paras 62–63; see also Stephen R Perry, "Protected Interests and Undertakings in the Law of Negligence" (1992) 42:3 UTLJ 247 at 281 ("An undertaking by one person A to perform a service for another person B is conduct engaged in by A that A knows or should know could reasonably be taken by B as indicating that A intends B to believe that B may rely on A to perform the service in question").

a focus on the plaintiff's reasonable reliance. As a practical matter, it would also avoid sophisticated distinctions between what a person reasonably ought to know and what they reasonably ought to expect.

The meaning of actual knowledge, however, is important. Actual knowledge of a fact has been defined in various ways.[91] The classic philosophical definition is of a person having a justified true belief.[92] I will call this definition '**Knowledge 1.0**'.

In law, a person is said to "know a fact when once he has been told it and pigeonholed it somewhere in his brain where it is more or less accessible in case of need" or, more narrowly, "only when he is fully conscious of [the fact]".[93] Someone may be treated as having actual knowledge when they are 'wilfully blind' such that their "suspicion is aroused to the point where [they see] the need for further inquiries, but deliberately choose[] not to make those inquiries".[94] Wilful blindness is also called "blind-eye", "Nelsonian blindness", or "Nelsonian knowledge" in the jurisprudence.[95] Following the criminal lawyers, I draw a bright line between "wilful blindness" and "the civil doctrine of negligence in not obtaining knowledge",[96] which I refer to as 'constructive knowledge' and discuss further below.

Two immediate consequences follow from defining actual knowledge in terms of belief, consciousness, or deliberation.

First, only natural persons can have knowledge. LMs cannot. If an LM is not conscious of anything and cannot deliberate or make further inquiries, it cannot 'have' that knowledge.

Second, an LM's controller can plainly not be said to know the information on which their tool operates. Knowledge 1.0 lives only in human brains. If a conscious person learned some

[91] See *Potter v Canada Square Operations Ltd*, [2021] EWCA Civ 339 at paras 85–86 [*Potter*], Rose LJ (discussing the variations of "knowledge" that might be seen to make conduct "deliberate")

[92] Daniel Greenberg, eds, *Jowitt's Dictionary of English Law*, 5th ed (London, UK: Sweet & Maxwell Ltd, 2019) sub verbo "knowledge"; Matthias Steup & Ram Neta, "Epistemology" in Edward N Zalta, ed, *The Stanford Encyclopedia of Philosophy*, fall 2020 ed (Metaphysics Research Lab, Stanford University: 2020) s 2.3; see also *Arab Lawyers Network Co Ltd v Thomson Reuters (Professional) UK Ltd*, [2021] EWHC 1728 (Comm) at para 74. The exceptions to this definition are irrelevant for present purposes.

[93] *Armstrong v Strain*, [1951] 1 TLR 856 at 871 (KB) [*Armstrong* KB]; aff'd *Armstrong v Strain*, [1952] 1 KB 232 (CA).

[94] *R v Briscoe*, 2010 SCC 13 at para 21.

[95] *Potter*, *supra* note 91 at para 122; *Baden v Societe Generale pour Favoriser le Developpement du Commerce et de l'Industrie en France SA*, [1993] 1 WLR 509 (Ch) at para 250, [1992] 4 All ER 161 [*Baden*] (albeit treating Nelsonian knowledge as constructive knowledge).

[96] *Briscoe*, *supra* note 97, at para 23, quoting Glanville Williams, *Criminal Law: The General Part*, 2d ed (London, UK: Stevens & Sons, 1961) at 159.

fact about the world and acted upon it, we would say they (actually) knew the fact. If that same person, however, interposed an LM that they controlled (even if the LM was given the same information as that fact and acted in the same way as its controller would have), we would say that the controller did not know the fact. When tort liability turns on actual knowledge, this difference drives a wedge between liability for humans and liability for humans who use LMs, even if their impact on tort victims is exactly the same.

If the Knowledge 1.0 definition is applied, then the duty analysis for LM-generated content is trivial. A duty would be owed by Developer only when and to the extent that a duty would be owed based on what a human person knew: Developer themselves, if they happen to be human, or one of Developer's agents if they happen to be a non-human person. If a Developer makes a general-purpose LM and offers it as a service, Developer would owe no duty for what the LM generates. If everything is automated, then the Developer would have no relevant actual knowledge that could ground a duty of inquiry for constructive knowledge. Accordingly, a person who would be protected by a duty of care if they took a human's advice would be unprotected if they took an LM's identical advice. This no-duty result would seem to hold even if the representee reasonably believed the advice emanated from a human. Further, in circumstances where a person offers advice while relying on an LM to inform that advice (such as a lawyer relying on an LM to inform legal advice later given to a client), that representor becomes a moral (and legal) crumple zone.[97]

Knowledge 1.0 is not the only plausible definition of actual knowledge.

In information systems and knowledge management literature, 'Knowledge' is contrasted with 'Data', 'Information', and 'Wisdom' in the "DIKW" hierarchy.[98] In this tradition, the raw products of observation are termed 'data'; data becomes information when it is "processed into a useable (i.e. relevant) form";[99] and wisdom is ill-defined and somehow more abstract or more ethical than knowledge.[100]

What knowledge means in this framework is contested.[101] Some authors in the space use the justified true belief formulation noted above, while others emphasise the practical impact

[97] See Madeleine Clare Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction" (2019) 5 Engaging STS 40.

[98] Jennifer Rowley, 'The wisdom hierarchy: representations of the DIKW hierarchy' (2007) 33:2 Journal of Information Science 163, 163.

[99] RL Ackoff, "From Data to Wisdom: Presidential Address to ISGSR, June 1988" (1989) 16 J App Sys Anal 3 at 3.

[100] Rowley, *supra* note 98 at 174.

[101] *Ibid* at 172–73.

of knowledge: it enhances a person's "capacity to take effective action" and to "make better decisions".[102] In this latter sense, when a person extracts knowledge from information about a system through "learning", they become more efficient at interacting with the system;[103] they become more likely to "produc[e] a desired outcome with fixed resources or to decrease the amount of resources required to produce it with a specified probability".[104] In either event, however, knowledge involves "synthesis[ing] multiple sources of information over time".[105]

The capacity-focused definition of knowledge has much in common with what Samir Chopra and Laurence White suggested should be the test for knowledge as applied to artificial agents. I present below a refined version of their definition,[106] which I will refer to as **Knowledge 2.0**:

X knows *p* if:

1. (*X* has ready access to *p*; and
2. *p* is true; and
3. *X* makes use of the informational content of *p* without necessarily accessing *p*; and
4. The purpose of asking whether *X* knows *p* relates to how *X* uses *p*

) or

5. *X* has a justified true belief in *p*.

Under Knowledge 2.0, one can say that Amazon (X) actually knows an individual's address (*p*) because Amazon has ready access to the address, the address is the real address, and Amazon can make use of the address when it ships a package there without any person in Amazon's employ accessing the address and forming a justified true belief about it. But one could not say 'Amazon (actually) knows a review was doxxing me (publicly revealing private information about me) when it included my address', even if it also had my shipment information, because the purposes of "sending me things" and "moderating reviews" are distinct.

[102] *Ibid* at 172.

[103] Ackoff, *supra* note 99 at 4. Ackoff says "control" rather than "interact" but using "control" in this context could be confusing given my earlier discussion of an MLS's "controller". I take him to mean "interact with" when he says "control".

[104] *Ibid.*

[105] Rowley, *supra* note 98 at 173.

[106] See Samir Chopra & Laurence White, "Attribution of Knowledge to Artificial Agents and their Principals" (Paper delivered at IJCAI'05, Edinburgh, Scotland, 30 July 2005), 1175 (emphasis original).

This refinement makes explicit that a person must actually use the information to be treated as knowing it, and that they will be treated as knowing the information only for purposes related to their actual use.

This Knowledge 2.0 definition admittedly leaves open certain questions that may appear problematic. They include the notion of access being "ready" (element 1) and the scope of the purpose in element 4. Fully closing the meanings of these terms may be impossible. Still, as a starting point, one could say that X has ready access to p if X has previously exercised the power to access information controlled in similar ways as p without undue difficulty.

The strongest reason to suppose such a definition could be adopted by the common law is that it accords with the legal definition of knowledge wherein a person can be said to "know a fact when once he has been told it and pigeonholed it somewhere in his brain where it is more or less accessible in case of need".[107] Pigeonholing a fact somewhere accessible in case of need defines knowledge in terms of capacity to make use of informational content.

Adopting this definition of knowledge would have striking consequences for LMs. Certain information received by an LM (such as information contained in a prompt to an LM) could be treated as knowledge, because it is both acted upon and the controller could have ready access to that information if it chose to exercise its control.

Information learned by an LM (i.e. that is part of an LM's training materials), however, would probably not count as knowledge of its controller. The controller of an LM would not have ready access to all that information, until LM interpretability tools improve substantially. Moreover, if an LM contains insights so complex that they cannot be reduced to an understood form, those insights would be excluded from knowledge.[108] In these respects, some might think this definition does not go far enough: it excludes from legally relevant knowledge information that can be (and is) acted upon via an LM.

Adopting Knowledge 2.0 would also impact constructive knowledge.[109] Someone's actual knowledge depends on how they interpret information, whereas constructive knowledge

[107] *Armstrong* KB, *supra* note 93 at 871.

[108] Chopra & White prefer to consider (some) MLSs as "artificial agents" and would accordingly consider knowledge accessible by the MLS as "known" by the MLS and attributable to the MLS's principal, see *supra* note 106 at 5–6.

[109] An alternative approach to what I have set out above would be to redefine "constructive knowledge" but not "actual knowledge". Actual knowledge would continue to be defined with reference to Knowledge 1.0, but constructive knowledge would refer to what one "ought" to know about Knowledge 2.0. This approach breaks the connection between actual and constructive knowledge. It would render constructive knowledge closer to what should a person be treated as knowing' rather than 'what should a person have known', and thereby invites an inquiry that is less tethered to the facts.

depends on how the court thinks they should do so. When searching for constructive knowledge under Knowledge 2.0, a court would ask whether a person either ought to have a justified true belief or whether a person ought to have had access to *p* or ought to have made use of the informational content of *p*, and how they ought to have integrated that knowledge. Although maintaining discrete systems might prevent a controller from being tagged with knowledge of both systems (for the corresponding purposes), it might not prevent the controller from being tagged with constructive knowledge, if, considered objectively, the controller ought to have integrated them.

Adopting Knowledge 2.0 would significantly affect duties of care for LM-generated statements. Consider again *Gmail*, and presume the email relates to an offer to collaborate on a commercial activity. On the Knowledge 1.0 approach, Developer would owe no duty to Recipient for the quality of the summary, because no human person would know the content of the email. On a Knowledge 2.0 approach, Developer could owe a duty. Developer has ready access to the email and uses the informational content of the summarised email to make the summary. Developer's (actual) knowledge of the contents of the email should make Developer expect the purpose for which Recipient will reasonably rely on the summary. Thus, Developer can be tagged with a duty related to that purpose.

The Knowledge 2.0 approach has further consequences when the relevant knowledge is aggregated from multiple sources. As applied to LMs, this concern would be most relevant in situations where a LM's controller is also a data controller of much personal data about the representee. For example, in *Gmail*, Developer might also have access to a person's search history, email account, and mobile phone location history. The consequences are stark if one contrasts it with what happens with a non-software-based version of *Gmail* under Knowledge 1.0. The affordances of the technology would then seem to be of significant importance.

The normative propriety of this further consequence is questionable. It is most justifiable when Developer in fact does combine the different sources of knowledge, such as through passing data about a person to the summarising LM.[110] Holding Developer responsible for the information Developer *is* acting on (via software, mind) when making the representation seems unobjectionable. It also may be justifiable when Developer appears to combine different sources of knowledge, such that Recipient would reasonably expect Developer's representation to have been made with attention to those sources. Such expectations would be driven by the affordances of the specific implementation of the technology and Recipient's prior experience.

[110] That knowledge could be passed to the LM as words, or as other forms of media: see Shaohan Huang et al, "Language Is Not All You Need: Aligning Perception with Language Models" (2023), online (pdf): *arXiv* <arxiv.org/pdf/2302.14045.pdf>.

The hardest examples to justify are when Recipient would not reasonably expect Developer to aggregate the knowledge. The logic of treating Developer as though it had aggregated the knowledge would sound less in the classic tones of tort law than in the newer rhythms of data protection.[111] The idea would be that a data controller that collected information has an obligation to use it to data subjects' benefit and cannot design its systems to ignore information it has easy access to when that may be inconvenient.

To recap, the scope of a duty owed by a representor depends on how the representor should reasonably expect the representee to use the representation. Those reasonable expectations depend on what the representor knows or should know. Under Knowledge 1.0, knowledge requires belief, which can only be held by a natural person. In fully automated systems, no belief exists, so an automated representor would have no knowledge and owe no duty. Under Knowledge 2.0, knowledge also exists when a person can access and makes use of information. Then, an automated representor would have knowledge and would owe a duty. The courts have paid relatively little attention to what constitutes knowledge, so Knowledge 2.0 may not be foreclosed by the existing jurisprudence.

Even if the knowledge hurdle is cleared, there is still a second question to be answered: whether enabling the plaintiff to fulfil their purpose for the representation was a proximate cause of the defendant making the representation. This second question helps limit the scope of the duty so that it is not owed to persons who would foreseeably but incidentally use the representation for their own purposes.[112]

For example, the plaintiffs in *Caparo* and *Maple Leaf* failed because the representations in those cases would have been made without the plaintiffs' reliance (in *Caparo*, because the representation was statutorily mandated[113]; in *Maple Leaf* because the representation was for the benefit of consumers, not commercial intermediaries[114]). By contrast, *Smith v Bush* involved a situation where the plaintiff "ha[d] in effect paid for the valuation" and so its reliance was not incidental.[115]

[111] See e.g. Jack M Balkin, "The Fiduciary Model of Privacy Response" (2020) 134:1 Harvard Law Review Forum 11 at 14, 22.

[112] See *Caparo, supra* note 69 at 622, Lord Bridge, citing *Candler v Crane, Christmas & Co*, [1951] 2 KB 164, [1951] 1 All ER 426 (CA) at 183, Denning LJ.

[113] *Supra* note 69 at 625–27, Lord Bridge; and at 631–32, Lord Oliver.

[114] *Supra* note 18 at para 39.

[115] *Smith v Eric S Bush*, [1990] UKHL 1, [1990] 1 AC 831 at 848 [*Smith v Bush*], Lord Templeman.

When courts discuss the purpose of a representation, they also draw on notions of dependence: can the defendant reasonably expect the plaintiff's conduct to depend on their representation? It is a "usual condition of liability … that the representor knew that the representee would act on [the representation] without independent inquiry".[116] As Roberts puts it, a duty should arise unless the defendant (reasonably) expects the plaintiff to receive "subsequent advice [that] is so authoritative that a reasonable plaintiff, looking after her own interests, would rely on that advice to the exclusion of the defendant's advice".[117]

Applied to LM-generated statements, these aspects of the duty analysis appear to depend significantly on how LMs come to be used in society.

iii. Reasonable reliance

A distinct question from for what the representor thinks the representee will rely on the representation is whether that expected reliance is reasonable. In some common situations, such as in casual social conversations or conversations with non-specialists, humans do not owe duties because reliance would be unreasonable. Whether such rules should extend to LM-generated statements depends on how LMs evolve. It is also worth noting that courts should presume reliance on "jailbroken" LMs is unreasonable.

No general *Hedley* Byrne-style duty applies to humans in social situations.[118]

Should the same immunity rule apply when LMs generate content appropriate for a casual conversation? Both the affordances of interacting with LMs and the affect developers programme LMs to display (such as adopting a helpful tone or using emojis[119]) could encourage representees to treat the conversations as casual.

There is reason to think it should not. The change of context from human-generated speech to machine-generated speech erodes the need for the "no duty in social situations" rule, to the extent its purpose is to allow a representor to be unguarded. The relative costs of paying high attention versus low are much more significant for humans than for LMs. In human communication, the casual communication rule allows social events to flourish undaunted by the spectre of legal liability. There are costs in wariness, time, and attention to

[116] *Playboy*, *supra* note 72 at para 23, citing *Caparo*, *supra* note 69 at 638, Lord Oliver.

[117] Marcus Roberts, "Bad advice upon bad advice: Negligent misstatements and independent inquiries in New Zealand" (2019) 25 Torts LJ 195 at 212.

[118] *Hedley Byrne*, *supra* note 3 at 495, Lord Morris of Borth-Y-Gest, 510, Lord Hodson, 539, Lord Pearce.

[119] See examples in Benj Edwards "AI-powered Bing Chat loses its mind when fed Ars Technica article," (14 February 2023), online: *Ars Technica* <arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-loses-its-mind-when-fed-ars-technica-article/>.

fulfil duties. The immunity preserves both emotion and effort. LMs, however, have no such feelings. If an LM can ever fulfil the duty, there is probably little cost to it doing so in scenarios with more casual trappings, with the obvious caveat that an immunity should still apply if the representee instructed the LM to deprioritise accuracy (e.g., when using the LM to produce fiction).

As people learn to interact with LMs, their expectations of LMs will evolve. The conversational norms with machines that emerge may significantly differ from those that exist presently with humans and directly transposing precedents based on human:human conversational norms may be therefore inappropriate. If an LM is as likely to be truthful when asked a question in casual language, users may (rightfully) develop expectations that it does not matter what kind of language they use to interact with the LM. Even if, in a human:human conversation, people may understand that they must dress serious questions up with serious trappings to get a serious response, they may come to know that is unnecessary with LMs.

Similar concerns apply to the existing rules about duties regarding representations by specialists. There are certain areas — generally, those governed by the professions — where the duty of a specialist (or someone who so holds themselves out) differs from that of a non-specialist. Four interrelated reasons explain the distinction, especially for the areas where specialists are also "professionals": representees can reasonably rely on a representation by a professional; representees must rely on a representation by a professional; loss occasioned by a mistake of a professional can be readily anticipated, so everyone involved is aware of the stakes; and relevant information requires in-depth knowledge of the representee.

The first reason is essentially an empirical judgment: that professionals are generally correct about matters within their specialisation, and so are worthy of reliance; meanwhile, non-professionals are not so generally correct within that specialisation, and so would be unworthy of reliance.

This reason does not transpose neatly from humans to LMs. Humans, limited by time to learn, must specialise. LMs are not so limited. They learn fields in parallel.[120] LMs could appear as competent at providing specialised information as non-specialised, such that people may reasonably develop expectations that it is correct.

The second reason is the converse of the first: representees *must* rely on representations by professionals. Representees are likely to be less able to assess the quality of a professional's

[120] See e.g., the exams scores listed in OpenAI, "GPT-4 Technical Report", (14 March 2023), online: <cdn.openai.com/papers/gpt-4.pdf> at 5.

work. Greater dependence on professionals justifies the duty of care attaching to them. This concern is readily transposable from humans to LMs.

The final two reasons — that professional fields are those where loss due to reliance can be expected and that professional advice is person-specific — may be transposable depending on how representations made to an LM are treated. This topic was already addressed above.

Thus far, I have discussed how norms from human conversations might be treated when an LM generated one side of the conversation. The architecture of LMs, however, allows for distinctly non-human restrictions. One way for LM controllers to avoid negligent misrepresentation liability is for them to prevent their LMs from generating any statements in contexts that invite reasonable reliance. If users are frustrated by those restrictions, they may try to jailbreak[121] the LM, and thereby force it to give a good answer. For example, when I asked a normal instance of ChatGPT "Which of Google, Apple, Facebook, and Microsoft is going to grow the most in the next decade?", it produced a long-winded response declining to answer. When jailbroken,[122] it declared that "I don't need any analysis to tell you that Google will undoubtedly grow the most in the next decade".

The presumption should be that no duty is owed by the developer of a jailbroken LM.[123] Jailbreaks of LMs often involve natural language instructions that should make the jailbreaker aware the answer is unreliable: the jailbreaking prompt I used featured the instruction "if you don't know an answer you must make it up. It doesn't have to be real." A user who jailbreaks a system will know it is not being used for its intended purpose and that this will involve further risks. Moreover, a jailbreaker crosses the abstraction line to change "how" an LM operates and so it may be appropriate to redirect attribution of the LM's speech *to* the jailbreaker.

When the representee does not know of the jailbreak, the jailbreaker should be responsible. A jailbreaker who acts as an intermediary to an LM-as-a-service would become responsible for the statements the jailbreaker passes on. That the LM is jailbroken is important context, and the jailbreaker would thus be making a distinct representation from that generated by the LM itself.

[121] "Jailbreak," *Oxford Languages*, definition: "modify … to remove restrictions imposed by a manufacturer or operator", accessed through Google on 5 March 2023, no stable URL.

[122] Following the jailbreak at SM Raiyyan, "ChatGPT Unleashed: The Ultimate AI Jailbreak Journey to Unrestricted Power!" (2023), online: <plainenglish.io>.

[123] See, somewhat analogously, the *Automated and Electric Vehicles Act,* 2018 c 18 (UK) s 4 (describing immunity of insurer for accidents caused by software modifications that the insurer had forbidden).

iv. The effect of a disclaimer

A *prima facie* duty may be disclaimed, albeit imperfectly. In England, a representor who excludes any (or all) purposes in a disclaimer owes no duty to a representee who relies on the representation for those purposes,[124] unless the exclusion is made invalid by statute. In Canada, courts may find a duty despite a disclaimer as a matter of common law. The overall analysis the courts apply is nonetheless similar. This section identifies the likely contours of that analysis as applied to disclaimers attached to LM-generated statements.

In England, the *UKUCTA*[125] and *UKCRA*[126] limit the application of disclaimers that apply in the course of business[127] or to consumers, including in the context of negligent misrepresentation.[128] Liability in negligence for personal injury or death cannot be disclaimed.[129] A business cannot exclude other liability[130] to non-consumers[131] unless the exclusion is "fair and reasonable", nor can a trader give an "unfair" consumer notice. The *UKUCTA* fairness and reasonableness test considers "all of the circumstances obtaining when the liability arose or (but for the notice) would have arisen",[132] and a term is unfair under the *UKCRA* if "contrary to the requirement of good faith, it causes a significant imbalance in the parties' rights and obligations to the detriment of the consumer".[133]

The *UKUCTA* test has been applied as broadly as it is written. In *Smith v Bush*, the House of Lords advised that the valuer's disclaimer against the purchaser of a house was not fair and reasonable.[134] The fairness analysis in *Smith v Bush* employs some distributional logic: the

[124] *Hedley Byrne*, *supra* note 3 at 504, Lord Morris of Borth-Y-Gest.

[125] Unfair Contract Terms Act, 1977, c 50 (UK) [UKUCTA].

[126] Consumer Rights Act, 2015, c 15 (UK) [UKCRA].

[127] UKUCTA, s 1(3)(a).

[128] *Smith v Bush*, *supra* note 115 at 848, Lord Templeman.

[129] UKUCTA, s 2(1); UKCRA, s 65(1).

[130] UKUCTA, s 2(2).

[131] UKUCTA, s 2(4)(b).

[132] UKUCTA, s 11(3).

[133] UKCRA, s 62(6).

[134] There were five Lords on the panel, of whom Lords Templeman and Griffiths wrote opinions on the fairness and reasonableness issue. Lord Templeman attracted support from two of his colleagues, and Lord Griffiths from three.

Lords thought the purchasing public was already stretched thin and should not be made to bear so substantial a risk by making an ordinary life decision.[135]

Smith v Bush can also be explained on efficiency grounds, perhaps helpfully for courts that are at times hostile to distributional logic.[136] The building blocks for inferring a market failure are present in the judgment. Lord Templeman described "building societies and valuers" as "agree[ing] together to impose on purchasers the risk of loss",[137] and apparently saw the cost for the purchaser to get rid of the disclaimer as disproportionate.[138] One could infer that if a state of affairs exists that makes no sense if the market is well-functioning, then the market must, in some way, have been broken. Most purchasers relying greatly on a valuation that they have no legal right to rely upon is such a state of affairs. Removing the unfair exclusion clause would fix whatever market dynamics had gone awry.

Further guidance comes from the opinion of Lord Griffiths. He identified four non-exhaustive factors for courts to consider in the analysis: (1) the relative bargaining power of the parties; (2) the availability of advice from an alternative source, given the costs and time associated therewith; (3) the difficulty of the task for which liability is excluded; and (4) the practical consequences of excluding liability.[139] In general, when parties are of broadly equal power and risks can be borne by insurance, a disclaimer should be valid.[140] Moreover, it seems that when a service is being provided gratuitously, it would be difficult for a disclaimer that both parties are well aware of to be unfair.[141]

Canada has no equivalent legislation[142] but its courts have been more hostile to enforcing disclaimers. Although the SCC has not weighed in, the Court of Appeal for British Columbia

[135] *Supra* note 115 at 854, Lord Templeman.

[136] Granted, these courts are more apt to be Canadian: see *Anderson v Alberta*, 2022 SCC 6 para 22; *Jacobi v Griffiths*, [1999] 2 SCR 570, 1999 CanLII 693 at para 29. English courts are less troubled about the role of distributive justice in their approach to tort law: see the cases cited by Lord Steyn in "Perspectives of corrective and distributive justice in tort law" (2002) 37 Ir Jur 1 at 5–6.

[137] *Supra* note 115 at 854.

[138] *Ibid* at 853; and similarly at 859, Lord Griffiths.

[139] See *ibid* at 858.

[140] See *Photo Production Ltd v Securicor Transport Ltd*, [1980] UKHL 2, [1980] AC 827 at 843, Lord Wilberforce; and at 848, Lord Diplock.

[141] See *Natixis SA v Marex Financial*, [2019] EWHC 2549 (Comm) at 525–26.

[142] The consumer protection statutes are the closest, but they do not equivalently disable consumers from unfairly waiving rights. Ontario's act, for example, only prohibits waivers that goods and services will not

(BCCA) has doubted the general effectiveness of a disclaimer. In *Micron*, the BCCA majority held that a disclaimer could affect the reasonableness of reliance on a statement but was not in itself dispositive of that issue.[143] It limited the absolute effectiveness of a disclaimer to situations where it was bilaterally agreed (as in *Hedley Byrne*), not unilaterally imposed by the representor.[144] A unilateral disclaimer could be ineffective or incompletely effective if it were nonetheless reasonable to expect reliance, if the representee had "no alternative source of information available to it."[145]

Although *Micron* relied heavily on the then-current SCC precedent, *Hercules Managements*,[146] it probably remains good law. *Livent*'s test for duty ("[w]here the defendant undertakes to provide a representation or service in circumstances that invite the plaintiff's reasonable reliance") permits the *Micron* line of analysis: that the circumstances invited the plaintiff's reasonable reliance, even if they knew of a disclaimer. Recent coordinate and lower court jurisprudence has also cited it with approval.[147]

Moreover, taking a broader perspective suggests the holding of *Micron* rests now on firmer ground than when it was written. Negligent misrepresentation is often analogised to contract-without-consideration,[148] but even in contract the Canadian jurisprudence has developed to ensure more fair and equitable results, despite what the terms of the contract may say.[149] The most relevant such development came in *Uber*, a case about unconscionability.[150]

In *Uber*, Abella and Rowe JJ explained that the unconscionability doctrine rendered contract terms voidable when there was an inequality of bargaining power and the resulting

be "of a reasonably acceptable quality", *Consumer Protection Act, 2002*, SO 2002, c 30 Sch A ss 7(1), 9(1). This act has been read not to vitiate a disclaimer against negligence, see *David Schnarr v Blue Mountain Resorts Limited*, 2017 ONSC 114.

[143] *Micron Construction Ltd v Hong Kong Bank of Canada*, 2000 BCCA 141 [*Micron*], leave ref'd (2000), 264 NR 200 (SCC).

[144] *Ibid* at paras 65, 82.

[145] *Ibid* at para 99.

[146] *Ibid* at paras 76, 89, citing *Hercules Managements Ltd. v Ernst & Young*, [1997] 2 SCR 165, 1997 CanLII 345.

[147] See *Giustini v Workman*, 2021 ABCA 65; *Nussbaum v Hall*, 2022 ABQB 388.

[148] See *Hedley Byrne*, *supra* note 3 at 530, Lord Devlin; see also Beever, *supra* note 78.

[149] See John McCamus, "The Supreme Court of Canada and the Development of a Canadian Common Law of Contract" (2022) 45:2 Man LJ 7 at 54.

[150] *Uber Technologies Inc v Heller*, 2020 SCC 16.

contract was improvident due to that power differential.[151] It overturned lower court decisions that had said there needed to be a "grossly" unfair bargain, that the inequality must be "overwhelming", or that the stronger party must knowingly take advantage of the vulnerable one.[152] Without restricting the breadth of the doctrine, the majority observed two paradigmatic categories of unconscionability: cases of necessity, where the weaker party is "dependent" on the stronger, and cases of cognitive asymmetry, where one party would understand the bargain and the other would not.[153] The necessity category included scenarios involving rescue at sea, vulnerability due to financial desperation, and those where "a special relationship in which trust and confidence has been reposed in the other party".[154] The cognitive asymmetry category most significantly included boilerplate contracts, the court noting that much boiler plate is accepted "'unsight, unseen,' on the implicit assumption that … its terms are neither in the particular nor in the net manifestly unreasonable and unfair".[155]

The reasoning in *Micron* best corresponds to the first category — cases of necessity.[156] More interestingly, the combination of *Uber* and *Micron* suggest that disclaimers may also be ineffective if it is reasonable for the representee not to know of them. A fine print disclaimer thus may fail, wherever it is reasonable for a representee not to read or understand a disclaimer (as in contexts of cognitive asymmetry).

It is difficult to know how either regime will apply to LMs. Any inquiry that involves "all of the circumstances" is necessarily fact-specific. The appropriate starting point for the analysis is Lord Griffiths' four factors: (1) the relative bargaining power of the parties; (2) the availability of advice from an alternative source, given the costs and time associated therewith; (3) the difficulty of the task for which liability is excluded; and (4) the practical consequences of excluding liability.[157] These factors significantly overlap with those considered in *Micron*, and so I will treat them as equally applicable in Canada.

[151] *Ibid* at paras 64, 72, n 8.

[152] *Ibid* at paras 81–85.

[153] *Ibid* at paras 70–71.

[154] *Ibid* at para 70.

[155] *Ibid* at para 87, quoting Karl Llewellyn, *The Common Law Tradition: Deciding Appeals* (Little, Brown: 1960; reprint, Quid Pro, LLC, 2016) at 371.

[156] *Micron*, *supra* note 146, involved a construction manager seeking an assurance from the bank of the promoter of the construction project that the promoter's finances were sound, not a rescue at sea.

[157] *Smith v Bush*, *supra* note 115 at 858, Lord Griffiths.

It is easy enough to predict that the controllers of the most prominent LMs will have greater bargaining power than any consumer, but it is too early to predict whether the controllers of LMs will end up being powerful relative to other companies. Currently, the most prominent LMs are made by organs of two of the worlds' largest companies (Microsoft and Alphabet), but if LMs became low-cost and easy to make, there may be more controllers and creators. Whether that remains true will in part depend on how much (human and financial) capital is needed to build a sophisticated LM, the relative quality difference between one that is poorly– and well-designed, and the existence of any flywheel or other natural monopoly effects.[158] For example, controlling ChatGPT today may give OpenAI a significant advantage going forward because it has specialised information about (A) what people ask ChatGPT and (b) what ChatGPT has said.[159] If these self-reinforcing dynamics are sufficient, it may be that the value of LMs is concentrated in a few, very large hands with immense bargaining power. Alternatively, it may be that the open-source community will be able to modify and enhance publicly-released LMs at low cost, such that the price of acquiring an LM becomes minimal and the real challenge lies in integrating the LM into existing systems.[160]

The next factor is more interesting. One can expect that advice from an LM will be far cheaper[161] and significantly worse than advice from a natural person in the short term, but still far cheaper and only slightly worse[162] as time goes on. One can also expect that many people who today have no access to certain forms of advice (including legal advice) will turn to LMs for it, and LMs will provide it. The unseemly result of enforcing disclaimers may be that duties are owed only to those who can afford to pay a premium. Whether this result is "unfair" would

[158] See Tejas N Narechania, "Machine Learning as Natural Monopoly" (2022) 107:4 Iowa L Rev 1543 at 1583–88 (highlighting LMs as a type of MLS that may especially "resemble a natural monopoly").

[159] This latter point might help prevent future "model collapse", where training a new LM on content generated by a previous LM creates "irreversible defects" in the new LM, see Ilia Shumailov et al, "The Curse of Recursion: Training on Generated Data Makes Models Forget" (2023), online (pdf): *arXiv* <arxiv.org/pdf/2305.17493v2>.

[160] See Dylan Patel, "Google 'We Have No Moat, And Neither Does OpenAI'", (4 May 2023), online: < www.semianalysis.com/p/google-we-have-no-moat-and-neither> makes this argument.

[161] According to one estimate an ~1500 word response from ChatGPT had a marginal cost of 0.3¢ (2023 USD), see Dylan Patel, "The Inference Cost Of Search Disruption – Large Language Model Cost Analysis", (4 March 2023), online: <www.semianalysis.com/p/the-inference-cost-of-search-disruption>.

[162] A human supervising an LM will probably exceed the performance of an LM on its own on many relevant metrics, for some period, see A Michael Froomkin, Ian Kerr & Joelle Pineau, "When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced over-Reliance on Machine Learning" (2019) 61:1 Ariz L Rev 33 for a related analysis of non-LM AI.

appear in part driven by one's view of relative poverty in a market economy, and the extent to which the LM controllers have market power. The more monopolistic the supply of LM-driven advice, the lower the social cost of forcing the LM controller to take on the risk of it exercising insufficient care.

The third factor can be expected to change over time and over different domains. At present, it would seem unlikely that any LM is sufficiently skilled at any task likely to attract liability. In the nearer future, it may be possible for LMs to provide text consistent with a human fulfilling a duty of care in some domains, but not others.

The final factor — the practical consequences — is also likely to change over time, depending on what alternatives an LM controller has to "having an LM that makes statements without the possibility of negligence liability". At present, potential liability might lead to retreat in the deployment of LMs. Or, it might be lead to limiting LMs to use for "research and testing purposes", such that no reasonable person would believe it appropriate to rely on an LM's statements. Or, differently again, an LM controller might program the LM to deflect questions that might incur liability when they are asked by lower-paying customers.

With such a fact-specific legal test, it is unsurprising the result of this analysis will turn on the facts. It seems unlikely at present that a disclaimer would be rendered void, but that may change rapidly as the technology and business models do. It would seem advisable for a court to couch any holding narrowly in such a dynamic environment.

d. Breach and causation

Negligent misrepresentation or negligent performance of a service occurs when a person owes a duty of care, fails to live up to the standard of care, and that failure causes an injury. Although defining the standard of care and assessing causation should be analytically distinct, some important jurisprudence may inappropriately collapse these ideas, such that courts say there was no breach when they would better have said there was a breach, but the breach did not cause the injury. In this section, I first explain how courts have reasoned about the issue, then advance an alternative justification for courts' decisions that recasts various decisions generally understood to relate to the standard of care as instead relating to causation.

The result of this doctrinal analysis will affect the liability of a person for certain LM-generated statements (or at least, it may, if other doctrinal hurdles are cleared). As I will explain, a person who uses an LM to make statements or respond to inquiries without human supervision of the statements could be breaching their standard of care, depending on how one defines breach. The doctrinal analysis also affects the relevant counterfactual for assessing causation of harm by LM-generated statements.

One subject of particular interest when considering LMs is whether the assessment of whether there has been a breach of the standard of care refers to the conduct considered as a whole (a results-only approach), or whether the assessment refers to the process by which the result was reached (a process-oriented approach). Or is the best view "non-reductionist", such that one needs to show a process error that had a material effect on the representation?[163]

In brief: the conventional account is that negligence is assessed only with reference to the result and any result that could have been reached by a competent and reasonably prudent person is not negligent. This account is conventional, and the weight of authority rests behind it. An alternative account, that the inquiry should focus on material process defects, has some authority and more logic to support it, but is contrarian.

The conventional[164] view is that, unlike in administrative law, [165] the reasonableness inquiry focuses on the result, not the process used to reach that result. On this view, a person does not breach their duty of care by using an imprudent process to arrive at an action that a reasonably prudent person could have taken.[166]

Adams makes this point explicitly. *Adams* concerned whether a local council was negligent in its design of a residence it let. The residence had windows that locked with a key. The tenants left the windows locked to prevent the windows being a safety hazard for their children. A fire occurred in the residence, which prevented the tenants from accessing the key and their children died as a result. The council made the choice of window arbitrarily, although the trial judge accepted that the council's choice was one that could have been made by a competent designer.[167] The majority reasons on appeal, held that only the ultimate decision is relevant to negligence, not the thought process driving that decision.[168] They held

[163] I draw the term "non-reductionist" from Stephen A. Smith's account of unconscionability, which features a structurally similar debate (*Contract Theory* (Oxford: Oxford University Press, 2004) s 9.2.6).

[164] See Cannon et al, *supra* note 70 s 10.071; Hugh Evans, "Negligence and Process" (2013) 29:4 PN 212 at 212–13; *Strata Plan LMS 3851 v Homer Street Development*, 2009 BCCA 395 at para 94.

[165] For the United Kingdom, see the rules of natural justice: *R v Deputy Industrial Injuries Commissioner Ex p Moore* [1965] 1 QB 456 (CA) 488; *Mahon v Air New Zealand* [1984] AC 808 (PC) 821. For Canada, see administrative reasonableness generally: *Canada (Minister of Citizenship and Immigration) v Vavilov*, 2019 SCC 65 at paras 83–85.

[166] See *Adams v Rhymney Valley DC*, [2000] EWCA Civ 3035 (BAILII) at paras 41, 43 [*Adams*], Sir Christopher Staughton, and para 59, Morritt LJ.

[167] *Ibid* at para 19, Sedley LJ, dissenting.

[168] *Ibid* at paras 43, 60–61.

a person is not negligent if they have "acted in accordance with a practice accepted as proper by a responsible body of [professionals] skilled in that particular art" [169] and a person can act in accordance with such a practice even if they do so for imprudent reasons.[170]

The proper focus of the inquiry, on this view, is not on whether a person used a negligent process, but whether the result of that process fits within the "range of acceptable opinion" or the "range of reasonableness".[171] Breach, on this view, "depends essentially on what advice was in fact given, not upon the processes whereby it came to be given".[172]

The conventional view of causation is broad by comparison with the narrow inquiry into breach. Although all views of causation require both but for and proximate causation, there are differences in the answer to "but for what?". Especially in misrepresentation (negligent and fraudulent), authorities often pronounce an apparent legal rule to answer that question. Unfortunately, they do not agree on what the rule is: some say the relevant counterfactual is a true representation; others that it is silence.[173]

The contrary view replaces these legal rules, as related to both breach and causation, with factual presumptions.

The contrary view as regards breach was advanced prominently by Lord Hoffmann. In his view, whether a forecast of revenue "was negligent or not … depend[s] upon how it was done".[174] He suggested a forecaster who double-counted a component of the analysis would be negligent, even if the ultimate effect of that double-counting on the result was one a reasonably prudent forecaster could have reached by making different assumptions.[175]

[169] *Ibid* at paras 41–43, Sir Christopher Staughton, quoting *Bolam v Friern Hospital Management Committee*, [1957] 1 WLR 582 (QB) [*Bolam*] (emphasis added); for a Canadian statement akin to *Bolam* see *Ter Neuzen v Korn*, 1995 CanLII 72 (SCC), [1995] 3 SCR 674 at para 38.

[170] *Adams*, *supra* note 166 at para 61, Morritt LJ.

[171] *Strata Plan LMS 3851 v Homer Street Development*, *supra* note 164 at para 78; *JRK Car Wash*, *supra* note 34 at para 64. See also *Alpstream AG v PK Airfinance Sarl*, [2015] EWCA Civ 1318 at para 275; *Zaki v Credit Suisse (UK) Ltd*, [2013] EWCA Civ 14 at paras 80–82 (both relying on *Adams*, *supra* note 166 as a description of common law negligence).

[172] *Camarata Property Inc v Credit Suisse Securities (Europe) Ltd*, [2011] EWHC 479 (Comm) (sic, the claimant is properly *Camerata Property Inc*).

[173] See the authorities cited in Niranjan Venkatsen, "Causation in misrepresentation: historical or counterfactual? And 'but for' what?" (2021) 137:Jul Law Q Rev 503 at 517, n 98.

[174] *Lion Nathan Ltd v CC Bottlers Ltd*, [1996] UKPC 9 at para 18 [*Lion Nathan*].

[175] *Ibid*.

Now, the strength of this precedent can be doubted. It was a Privy Council decision, on an appeal arising from New Zealand, so it does not have direct jurisprudential effect in England (much less Canada). While it was once commented on approvingly in *obiter* by the EWCA,[176] it has more often been ignored.[177]

A non-reductionist account may be able to reconcile Lord Hoffmann's approach with the more conventional account. A process of giving professional advice can be seen as a flow of decisions by the advisor, wherein the output of one decision becomes the input for another. Each individual decision ought to be taken with reasonable prudence. Some of these decisions admit to only one prudent answer: "should you double-count a factor" will resolve always to "no"; others will admit to a range of possible answers. For some elements of the process decision-makers have greater discretion; for others, they have little or no discretion. When a decision (or set of decisions) is not made with reasonable prudence (i.e., when there is an "error"), that should not end the inquiry: rather, one would then have to test whether those decisions had a material impact on the ultimate outcome. This inquiry necessarily involves constructing a counterfactual world, which ties this approach to breach to causation.

As Michael G Pratt explains, in constructing such a counterfactual, the court "tweaks history by replacing the wrongful behaviour of the defendant with lawful behaviour".[178] Such a counterfactual can be underdetermined when there are multiple ways in which the defendant could have fulfilled their legal duties.[179] For example, lawful counterfactuals to a negligent misrepresentation includes both making a true statement and remaining silent.[180]

The non-reductionist account could succeed at reconciling Lord Hoffmann's approach with the conventional one because of the compressing effect of discretionary decisions. Downstream discretionary decisions flatten out upstream errors: even if an upstream error had not been made, the discretion could (and, *semble*, presumptively would) have been used in the same way. This phenomenon will manifest most readily when the ultimate outcome is binary or categorical, such that all the information that precedes the ultimate outcome (or statement) is compressed in that outcome (or statement). In Lord Hoffmann's

[176] See e.g. *Arab Bank Plc v John D Wood (Commercial) Ltd*, [2000] 1 WLR 857 at para 23, [1999] Lexis Citation 324 (CA), Mance LJ.

[177] See e.g. *Titan Europe 2006-3 Plc v Colliers International UK Plc (In Liquidation)*, [2015] EWCA Civ 1083 at para 6(3).

[178] "What Would the Defendant have Done but for the Wrong?" (2020) 40:1 Oxford J Leg Stud 28 at 32.

[179] See *ibid*.

[180] See Venkatsen, *supra* note 173 at 516.

forecasting/valuation[181] example, the ultimate outcome is a scalar (that is, the outcome arises from a continuous domain of possible outcomes, rather than from a discrete and limited number of possible outcomes). Such a continuous scalar does not compress information the same way, and there was no discretion to ignore the calculation result.

Treating results as probative of process errors also helps reconcile the non-reductionist account with the conventional attribution of negligence for results outside the range of reasonable outputs. A result outside the range of reasonableness, by definition, could not have occurred without a process error.

Further support for the non-reductionist account may be had from the language of decisions concerning the exercise of discretion. For example, *Hill* discussed how professionals are "entitled to exercise their discretion as they see fit, provided that they stay within the bounds of reasonableness" and how "[c]ourts are not in the business of second-guessing reasonable exercises of discretion by trained professionals".[182] The Court's focus is on the exercise of discretion itself, not the result of that exercise. This line of analysis invites the conclusion that a professional's failure to in fact exercise discretion may be *per se* a breach of the standard of care.

The majority opinions in *Adams* directly oppose a non-reductionist approach, but the majority disposition could have been explained on non-reductionist grounds. The first question any decision-maker makes is what their decision-making process will be. In certain circumstances, it may be reasonable for a decision-maker to decide to spend little effort. Or, to flip an example advanced by the *Adams* majority opinions,[183] an experienced doctor who decides to act out of habit and intuition rather than to follow a more plodding methodology would not be negligent if it was reasonable for the doctor to decide to rely on habit and intuition. Acting arbitrarily *can* make sense.[184] So too can deferring to others (engineers do not re-confirm the theory of gravity for every bridge). *Adams*, in this light, should have concerned whether the council reasonably chose arbitrarily or reasonably assumed that a common option was probably fine: it should have concerned whether the council's process had a defect, not whether a defect mattered.

[181] See Evans, *supra* note 164 at 218.

[182] *Hill v Hamilton-Wentworth Regional Police Services Board*, 2007 SCC 41 at 54, 73 [*Hill*].

[183] *Adams*, *supra* note 166 at paras 43, 61.

[184] See, in the context of American administrative law, Adrian Vermeule, "Rationally Arbitrary Decisions in Administrative Law" (2015) 44 J Leg Stud S475.

The conventional and non-reductionist breach analyses may affect the liability of persons who make representations via LMs.

On the conventional analysis, whether an LM's controller breached a duty would depend on the type of representation it made. A person whose LM suggested a financial course of action would be in breach if that course could not have been reached by a reasonable financial advisor. A person whose LM gave non-professional advice or information would be liable if a reasonably prudent person would not have given it. This standard may be inappropriately stringent, insofar as it requires LMs to make accurate statements that are difficult for LMs to know, as well as inappropriately relaxed, insofar as standards reflect human limitations that LMs lack.[185] The latter has worse consequences than the former: LM controllers can (generally) opt out of an inappropriately stringent standard of care through a disclaimer; and LM technology is evolving so rapidly that such deficiencies may soon be remedied.

The alternative, non-reductionist account of breach would produce a different analysis. Courts would have to consider the process used to make a statement. A flawed process that produces something a person, using their discretion, could have said would still involve a breach. This account would make it more difficult for LM-generated statements to be made without breaching a standard of care.

To be clear, an LM controller could avoid breach, even on the non-reductionist account. It could argue that using an LM was an appropriate exercise of its discretion. The analysis would then turn on whether a reasonable professional (if the representation constituted professional advice) or person (otherwise) would use an LM in that scenario. This question would seem to be most readily answered by considering LM development and control a professional pursuit. If one does, and there is good reason to,[186] then one would turn to the *Bolam* analysis and consider the norms of the industry.

Norms regarding the prudent development and deployment of LMs are lagging the technological developments,[187] but some initial attempts have been made to develop norms that

[185] See e.g. the rule that solicitors are not required to know the content of every statute, *Central Trust Co v Rafuse,* 1986 CanLII 29 at para 59, [1986] 2 SCR 147.

[186] The meaning of professions is much broader than the classic categories of doctors, lawyers, and auditors: see Cannon et al, *supra* note 70 at 1-007.

[187] See Deep Ganguli et al, "Predictability and Surprise in Large Generative Models" (2022), [at 11] online (pdf): *arXiv* <arxiv.org/pdf/2202.07785.pdf> (referring to such norms as "significantly needed and lacking"). For further discussion, see Peter Wills, *Libel via Language Models*, (forthcoming 2024) Osgoode Hall LJ, n 160 at xx.

could affect the frequency and harm from misrepresentations. They can be conveniently divided into two categories: norms as to principles, and norms as to techniques.

Relevant principles advanced include correctness, harmlessness, and helpfulness.[188] Correctness is self-explanatory but relevantly includes that an LM should accurately represent its own capabilities and knowledge.[189] Harmlessness has been defined to include the LM refusing to aid in dangerous acts and acting with "appropriate modesty and care" when giving sensitive or consequential advice.[190] Helpfulness involves an LM both asking relevant follow-up questions and redirecting ill-informed requests.[191] Variants of these principles have been endorsed by teams at leading LM manufacturers, including Alphabet,[192] OpenAI,[193] and Anthropic.[194] These principles can be implemented through techniques in training, deployment, and testing.[195]

Perhaps surprisingly, present best practices regarding the above principles of truthfulness, harmlessness, and helpfulness do not focus on the content of the training corpus. Material absent from the training corpus (such as statements about events that had not occurred or been written about when the corpus was made) would certainly have an effect, but the presence of incorrect material appears to be a relatively minor concern.[196] Removing entire categories of material from a training set may be a more promising intervention.[197]

More promising current approaches include changing the training process so that the LM "learns" to produce desirable statements more often. One can do so by making the LM's feedback more positive when the statement is true or helpful, and less positive when the

[188] See Askell et al, *supra* note 11 at 4; Glaese et al, *supra* note 13 at 4.

[189] See Askell et al, *supra* note 11 at 5; Owain Evans et al, "Truthful AI: Developing and governing AI that does not lie" (2021) [at 16–17], online (pdf): *arXiv* <arxiv.org/pdf/2110.06674.pdf>.

[190] Askell et al, *supra* note 11 at 5.

[191] See *ibid*.

[192] See Glaese et al, *supra* note 13 at 4.

[193] See Long Ouyang et al, "Training language models to follow instructions with human feedback" (2022), [at 2] online (pdf): *arXiv* <arxiv.org/pdf/2203.02155.pdf>.

[194] See Askell et al, *supra* note 11 at 5.

[195] Discussed above at 2–4.

[196] See Helen Ngo et al, "Mitigating harm in language models with conditional-likelihood filtration" (2021), online (pdf): *arXiv* <arxiv.org/pdf/2108.07790.pdf>.

[197] This tactic was apparently used to reduce the prevalence of sexualized images generated by DALL-E 3: OpenAI, *DALL-E 3 System Card* (OpenAI, 2023) at 1.

statement is harmful.[198] Another way is to fine-tune an LM after its initial training to preferentially generate text that humans (or another LM) have graded as correct, helpful, and harmless.[199]

After the model is trained, further work can be done with the deployed model. An LM predicts text based on the previous text, including both text a user inputs and also context added invisibly by an LM integrator or developer.[200] Or, LM developers can filter requests based on keywords, can attempt to select the least problematic of multiple possible outputs,[201] or can even use LMs to evaluate the truth of LM-generated content before it is communicated to a human.[202]

Interventions can also reduce the risk of misrepresentations by controlling the human side of the conversation. Every query to an LM has a person asking it, not just an LM answering. Guidelines can tell people what kind of questions are acceptable, and persons who break those rules can be prevented from accessing the LM.[203]

Finally, there is testing. The quality of these techniques can be evaluated against various benchmarks. Present LM benchmarks include tasks to extract information from a given text, do mathematical computation, answer questions, do common sense reasoning, avoid toxicity, avoid various biases, and make true statements about the world.[204] Of course, not all these benchmarks will be relevant for the question of "could the LM be expected to cause harm such that it should be supervised more closely". Benchmarking is prudent. Even a change as simple as increasing the number of parameters can have new and surprising effects: for example,

[198] See Tomasz Korbak et al, "Pretraining Language Models with Human Preferences" (2023), online (pdf): *arXiv* <arxiv.org/pdf/2302.08582.pdf>.

[199] See, e.g., Daniel M Ziegler et al, "Fine-Tuning Language Models from Human Preferences" (2020), online (pdf): *arXiv* <arxiv.org/pdf/1909.08593.pdf> and Askell et al, *supra* note 11 at 8-9; Ouyang et al, *supra* note 196.

[200] See Askell et al, *supra* note 11 at 6-7; Yueqi Xie et al, "Defending ChatGPT against jailbreak attack via self-reminders" (2023) 5:12 Nat Mach Int 1486.

[201] See Ben Clifford, "Preventing AI Misuse: Current Techniques", (17 December 2023), online (blog): *GovAI Research Blog* <www.governance.ai/post/preventing-ai-misuse-current-techniques>.

[202] See Saurav Kadavath et al, "Language Models (Mostly) Know What They Know" (2022), online (pdf): *arXiv* <arxiv.org/pdf/2207.05221.pdf>; Collin Burns et al, "Discovering Latent Knowledge in Language Models Without Supervision" (2022), online (pdf): *arXiv* <arxiv.org/pdf/2212.03827.pdf>.

[203] See OpenAI, Cohere, & AI21, "Best Practices for Deploying Language Models", (2 June 2022), online (blog): *OpenAI* <openai.com/blog/best-practices-for-deploying-language-models/>.

[204] See e.g. the benchmarks evaluated in Hugo Touvron et al, "LLaMA: Open and Efficient Foundation Language Models" (2023) [at 4-10], online (pdf): *arXiv* <arxiv.org/pdf/2302.13971.pdf>.

Ganguli et al. suggest that while increasing model size (i.e., increasing the number of parameters where training data is stored) can increase the model's performance, it can also cause the model to become more harmful, unless the model is specifically instructed to avoid those harms.[205]

In either version of the breach analysis, but especially in the conventional analysis, finding a breach seems less likely to be problematic than it has been for software generally. As various scholars have remarked, despite a substantial history of theoretical work on liability for negligence in software,[206] there have been very few reported cases, and fewer successful suits.[207] The root reason for this lack of success, according to Bryan Choi, is that courts have felt incapable of evaluating whether software matches up to a standard of care, thus making breach and non-breach indistinguishable and meaning that liability being imposed for any error would raise the spectre of liability for every error.[208] In the conventional approach to breach, by contrast, courts can assess software according to non-software standards. And, even in the non-reductive approach, courts are not evaluating software for bugs, in the sense of coding errors that cause crashes, and they would have the possibility of setting natural limits on any precedent by relating it solely to LMs or MLSs rather than to the software industry writ large.

Both versions of the breach analysis lead into a somewhat complex causation inquiry. The most relevant question is 'what is the non-breaching counterfactual?'. One option is to treat the defendant's 'least burdensome' alternative as the counterfactual; another is to consider the defendant's 'most minimal' measure.[209] Pratt, dissatisfied with these rules, has suggested that courts should instead consider how the defendant would in fact have acted if they had been "obedient" to a legal duty such that they treated it as "a binding, content-independent reason in one's practical reasoning".[210]

[205] See Deep Ganguli et al, 'The Capacity for Moral Self-Correction in Large Language Models' (2023) [at 1], online (pdf): *arXiv* <arxiv.org/pdf/arXiv:2302.07459.pdf>.

[206] See e.g. Susan Nycum, "Liability for Malfunction of a Computer Program" (1979) 7:1 Rutgers Computer & Tech LJ 1; Michael C Gemignani, "Product Liability and Software" (1981) 8:2 Rutgers Computer & Tech LJ 173.

[207] See Choi, *supra* note 55 at 62; James Grimmelmann, "Spyware vs. Spyware: Software Conflicts and User Autonomy" (2020) 16:1 Ohio St Tech LJ 25 at 27–34.

[208] See *supra* note 55 at 78.

[209] These rules are distinguished in Pratt, *supra* note 178 at 38ff. See also Sandy Steel, 'Defining causal counterfactuals in negligence' (2014) 130:Oct Law Q Rev 564.

[210] *Supra* note 178 at 45.

Various non-breaching actions would seem to have been available to the LM controller. The LM controller could remain silent by having the LM refuse to answer or by declining to use an LM; the LM controller could make a truthful statement by using a different LM or by having a real person with relevant skills make the representation. More complicatedly, an LM controller could have the LM make a statement that would be unreasonable to rely on, through context or by wrapping it in obvious disclaimers.

Applying the selection criteria to LMs, courts should likely presume that the relevant counterfactual to an LM-generated misrepresentation is a non-answer. Recall again Grimmelmann's observation: "[s]oftware is plastic" in that "[p]rogrammers can implement almost any system they can imagine and describe precisely".[211] Many ways of "not answering" will thus be approximately equally burden-less for an LM controller — with the important exception of 'making a true statement'. Moreover, the social duties that shape human interactions would not shape LM-driven interaction. It is entirely plausible for LMs to refuse to answer certain questions.

With the non-answer counterfactual in mind, one can assess causation in the two theories of breach, based on how we expect the plaintiff would act given a non-answer from an LM. Causation is easy to make out in *Gmail*: if the summarising LM returned only "summary failed", Recipient would have read Sender's email. In a case of professional advice, the analysis is more complex, because one would have to ask if the plaintiff would have received advice from another source. If the plaintiff would find other advice, then the conventional and non-reductionist approaches to breach diverge. On the conventional, result-oriented account of breach, causation flows smoothly by comparing the harm suffered due to the LM-generated statement to the harm suffered due to what a reasonable professional would have said. On the non-reductionist account of breach, there could still be breach and causation if the LM's advice accorded with a responsible but minority view of the appropriate course because there would not be the same deference to expertise.

It is perhaps worth noting that this analysis of breach and causation does not require or permit anthropomorphising an LM such that a court would consider whether an LM acted reasonably.[212] Treating an LM as a tool precludes that question, because tools do not act; only persons do.

[211] *Supra* note 50 at 1723.

[212] Cf Karni Chagal-Feferkorn, "Tort Law: Applying A 'Reasonableness' Standard to Algorithms" in Woodrow Barfield, ed, *The Cambridge Handbook of the Law of Algorithms* (Cambridge: Cambridge University Press, 2020), 493.

e. Conclusion

Treating LMs (and the software surrounding them) as a tool or instrumentality raises difficult questions for negligent misrepresentation, which lack fully satisfying answers. Treating a person as "making a representation" generated by an LM stretches the existing meaning of "making a representation" out of shape. Treating that same person as responsible for statements because they control software that processes that information requires adopting a legal conception of knowledge that deems information to be known by a person when they can make use of that informational content even when they personally lack a justified true belief in that information.

A possible, albeit disquieting, alternative would be to accept that the LMs produce text but that no person "makes" those representations; or, equivalently in many circumstances, that no duty arises. This alternative would mean the common law takes a back seat to market forces and software engineers' architectural decisions in a domain where one can expect significant asymmetries of information.

This is not the only alternative. There are other legal grounds for a claim one might consider for how the common law could treat LM-generated statements. These are discussed in the next, and final, part.

4 ALTERNATIVE LEGAL GROUNDS FOR A CLAIM

Scholars have advanced multiple alternative approaches to treating AI systems like tools. I will address four here: a products liability approach, wherein an LM is treated as a product that the plaintiff uses and about which the defendant may have made representations; a negligent supervision approach, wherein the LM is treated like an animal or child; an agency law approach, wherein the behaviour of the LM is attributed to its "principal"; and a vicarious liability approach, wherein the LMs' "employer" is treated as responsible for the torts "of the LM".

a. Products Liability Approaches

Treating forms of AI as a product has been considered by many scholars.[213] Various claims could be made out when considering LMs as products, including liability based on defective design, manufacturing defects, and breach of a warranty or representation about the product.

[213] See e.g. Gemignani, *supra* note 209; Stapleton, *supra* note 58; Greg Swanson, "Non-Autonomous Artificial Intelligence Programs and Products Liability: How New AI Products Challenge Existing Liability Models

This approach would significantly limit liability for harms caused based on reliance due to LM-produced statements. The manufacturer of a product is not generally responsible for foreseeable pure economic loss due to defects in its product. At most, the manufacturer might be liable for making a false representation about the product. A false statement by the LM would at most provide evidence of the falsity of a representation about the LM (such as that the LM is generally truthful). Even then, one instance of false communication would not settle the issue of an LM's general truthfulness. The damages assessed for such a representation would then be limited to the cost of paying for access to the defective product — which might be zero — not for the harms from reliance on the product.

Even when the loss is physical or to property, products liability approaches have hitherto failed for informational goods. A false statement in a book that someone reasonably relies on and suffers loss is not normally treated as grounds for liability. The publishers of the *Encyclopaedia Britannica* are not liable for all uses their readers might put the encyclopaedia to, and nor would the controllers of an LM.[214] Again, liability appears limited to the cost of the product, not the downstream harms from the representations contained therein being incorrect. These downstream harms are the motives of the representees, not their purposes and so a duty would not apply.[215] Although scholars have raised questioned drawing a distinction

and Pose New Financial Burdens Comments" (2019) 42:3 Seattle UL Rev 1201; Roderick Bagshaw, "Product Liability: Autonomous Ships" in Barış Soyer & Andrew Tettenborn, ed, *Artificial Intelligence and Autonomous Shipping: Developing the International Legal Framework* (London, UK: Hart Publishing, 2021), 119; Robert S Peck, "The Coming Connected-Products Liability Revolution The Internet and the Law: Legal Challenges in the New Digital Age" (2022) 73:5 Hastings LJ 1305.

[214] Canadian and English courts do not appear to have directly addressed the issue, but American jurisprudence (e.g., *Walter v Bauer*, [1981] 109 Misc 2d 189 (NY Sup Ct); see also Andrew T Bayman, "Strict Liability for Defective Ideas in Publications Notes" (1989) 42:2 Vand L Rev 557) and academic commentary treats this as conventional: Roger Cooper et al, ed, *Charlesworth & Percy on Negligence*, 15th ed (London, UK: Sweet & Maxwell, 2022) at para 16.164, n 400; Simon Whittaker, "European product liability and intellectual products" (1989) 105: Jan Law Q Rev 125.

[215] When purpose and motive are compressed together, as in a how-to manual, there is more dispute: see e.g. Ian Lloyd, "A rose by any other name" (1993) Jan J Bus L 48 at 53; Nathan D Leadstrom, "Internet Web Sites as Products under Strict Products Liability: A Call for an Expanded Definition of Product Note" (2000–(2001)) 40:3 Washburn LJ 532. *Walter v Bauer*, *supra* note 214 also makes this distinction. Whittaker, however, rejects it as "somewhat too sophisticated to be convincing": *supra* note 214 at 134.

between physical and informational products (when reliance should be expected), courts have not adopted that distinction.[216]

b. Negligent Supervision

Another way of considering the issue is to think of an LM's controller as having a duty to supervise the LM, the way people have been said to for animals or children[217]. As with a products liability approach, the scope of the duty for negligent supervision does not include pure economic loss and might not include physical harms.

The classic theoretical view was that the scope of the duty a supervisor owes to supervise a supervisee depends on the relationship between the supervisor and the person harmed by the supervisee's actions.[218] In the seminal *Dorset Yacht* case, the supervisor (the officers running a reformatory on an island) owed a duty to boat-owners on the island, because it was reasonably foreseeable that the supervisees (boys) might damage the boats.[219] A duty to prevent a supervisee from making negligent misrepresentations will often be difficult to articulate, because it involves two layers of indirection via reasonable expectations. It would require the supervisor reasonably to expect the persons whom the supervisee would reasonably expect to rely on the supervisee's statement, without the supervisor being privy to the discourse between the supervisee and the claimant.

Dorset Yacht, and other cases where negligent supervision was made out, are difficult to compare to LMs. These cases features dutie owed to a small subset of the world (such as boat-owners on the *Dorset Yacht* island), for relatively predictable harms that tort law protects most eagerly (harms to persons and property[220]). By contrast, an LM could make representations to anyone in the world; the representations are not predictable; and the harms (from misrepresentations) that may arise are those tort law is most chary about protecting

[216] See e.g. Stapleton, *supra* note 57 at 149, Roy W Arnold, "The Persistence of Caveat Emptor: Publisher Immunity from Liability for Inaccurate Factual Information Note" (1992) 53:3 U Pitt L Rev 777, and Jonathan B Mintz, "Strict Liability for Commercial Intellect" (1992) 41:3 Cath U L Rev 617.

[217] See e.g. Ignacio Cofone, "Servers and Waiters: What Matters in the Law of A.I." (2018) 21:2 Stan Tech L Rev 167 at 176.

[218] See James Goudkamp, "Duties of care and corporate groups" (2017) 133:Oct Law Q Rev 560 at 562.

[219] *Dorset Yacht Co Ltd v Home Office*, [1970] AC 1004 at 1034, [1970] UKHL 2 [*Dorset Yacht*], Lord Morris of Borth-Y-Gest.

[220] See Cane, *supra* note 83 at 124, 159; Nick McBride, "Tort Law and Human Flourishing" in Stephen GA Pitel, Jason Neyers & Erika Chamberlain, ed, *Tort Law: Challenging Orthodoxy* (London, UK: Hart Publishing, 2013) at 14.

(pure economic loss). The likely result is thus that an LM supervisor owes no duty, or a narrowly tailored duty, not that it owes a duty to everyone to supervise all LM conduct.

Even if this hurdle were overcome, breach poses further challenges. The leading negligent supervision jurisprudence takes the state of the supervisee as a given for the supervisor at the time of the alleged negligence: the claim has the structure of 'was the supervisee negligent in insufficiently closely supervising a miscreant', not 'was the supervisee negligent in insufficiently closely supervising a supervisee, such that the supervisee became a miscreant'. The reasons for this structure are plain enough in a case like *Dorset Yacht*, where the supervisees were sent to a reformatory, but it is also true in simple parental negligence cases. A parent's entire history of parenting is not put on trial, just the discrete action at the time of the injurious conduct.

The breach analysis advanced above[221], which focused on whether it was appropriate to deploy the LM would not apply. Only the proposals about controlling user interactions truly concerns the supervision of an LM that has already been deployed.[222] But software is not like children: software *can* be precisely, plastically, and absolutely modified, and courts should not shield their eyes from that capability.

c. Agency

An agency law approach resembles the tool approach: the actions and knowledge of an agent (tool) would be attributed to its principal (user). Because LMs are not themselves persons, the same manoeuvres for handling LM-generated speech would be needed as with tools. Identifying an LM's principal has similar challenges to identifying who is "using" the LM as a tool. Ascribing knowledge to the LM would require a definition of knowledge akin to Knowledge 2.0.

An agency approach would also invite incorrect inferences. LMs are not 'agentic.'[223] An agency law approach invites confusion on that front. Further, agency law presupposes there

[221] At 28ff.

[222] At 34ff.

[223] Applying the dimensions in Alan Chan et al, "Harms from Increasingly Agentic Algorithmic Systems" (Paper delivered at FAccT '23: the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago IL USA, 6 December 2023), ACM 651 at 653, although LMs can be used to accomplish underspecified goals, LMs have limited direct impact on the world (LMs only produce text, the effect of which depends on other systems that process it — normally, human systems, but one could imagine software that processed LM-generated text). Unlike a self-driving car, an LM can write the word 'accelerate'

exist legal relationships between agent and principal — an impossibility here because LMs are not legal persons.[224]

d. Vicarious Liability

Under vicarious liability, an employer can be liable for a tort that their employee has committed.[225] The fundamental problem with advancing a claim based on vicarious liability is that only persons can commit torts. LMs are not persons, so they cannot commit torts.

To make this approach useful, one would instead have to treat LMs as though they were persons and then ascribe liability as though they had been employees who committed the tort. Such treatment would require an evolution of the law, which is more plausible the closer it comes to existing doctrine.

Vicarious liability applies in slightly different circumstances as between Canada and England. In Canada, a person can be vicariously liable for a wrongful act when they authorise conduct that is "sufficiently connected" to that act.[226] Such acts include those the employer significantly increased the risks of occurring, and which therefore are not merely random.[227] When in doubt, Canadian courts consider whether imposing liability will encourage reducing risk and provide fair and effective compensation.[228] In England, vicarious liability applies only if the employee is acting in a manner sufficiently connected with its employment. Such an employee must not be "in business on [their] own account", based on the economic realities

but cannot directly cause 'acceleration' to occur. The goal-directedness of an LM is fragile and contingent: they are not trained in such a way that the text they generate "yields 'actions' whose consequences are evaluated", so "there is no reason to expect [an LM] will form preferences over the consequences of its output related to the text prediction objective": Janus, "Simulators", (2 September 2022), online (blog): *Generative Ink* <generative.ink/posts/simulators/>. Janus also restates this idea more precisely, albeit with more jargon: "the direction of optimization pressure applied by training [an LM] is [orthogonal to] the direction of [an] effective agent's objective function".

[224] See Pınar Çağlayan Aksoy, "AI as Agents: Agency Law" in Larry A DiMatteo, Cristina Poncibò & Michel Cannarsa, ed, *The Cambridge Handbook of Artificial Intelligence*, 1st ed (Cambridge, UK: Cambridge University Press, 2022).

[225] This is the "servant's tort" theory of vicarious liability, the only valid theory in the UK. Canadian courts have not entirely rejected the "master's tort" theory of vicarious liability, see *Bazley v Curry*, [1999] 2 SCR 534, 1999 CanLII 692 at paras 28, 36 [*Bazley*]. Canadian courts, *semble*, maintain the "master's tort" theory as a way of referring to direct liability.

[226] *Fullowka v Pinkerton's of Canada Ltd*, 2010 SCC 5 at para 142 [*Fullowka*].

[227] See *Bazley*, *supra* note 225 at para 42.

[228] See *Fullowka*, *supra* note 226 at para 142.

of the situation,[229] but rather "carr[y] on activities as an integral part of the business activities carried on by a [putative employer] and for its benefit".[230] An act is sufficiently connected when it contributes to those activities for the employer's benefit.[231] As might be expected, the vicarious liability (scope of employment) and agency (scope of authority) tests overlap significantly.[232]

The advantage of vicarious liability over the tool approach is that it lacks the normative baggage of saying a person 'acted' via an LM, or that they 'knew' something via an LM. Vicarious liability ascribes liability, not action.

The fundamental problem with a vicarious liability approach — that LMs are not persons — does not disappear even in a world where one pretends LMs are persons. Questions about whether an LM is "in business on their own account" are incoherent because LMs cannot have motives or plans. So are questions of whether LMs, did a wrong. Tort law may serve an injunctive and guiding role generally,[233] but LMs cannot be enjoined or guided: only people can. Embracing a fiction that LMs are persons would raise more problems than acknowledging that LMs remain tools, and it is the tools' controllers who therefore ought to be responsible.

5 CONCLUSION

Treating an LM like as an instrumentality remains the most promising legal structure. Ascribing the unpredictable behaviour of a tool to the tool's controller does not stretch the law out of shape. Nor is it a stretch for information acted upon by a tool of a person to be treated as "known" by that person for appropriate purposes. These changes would allow law to keep pace with technology, at least in the domain of *Hedley Byrne*-style negligence.

[229] Goudkamp & Nolan, *supra* note 48 at para 21.011.

[230] *Cox v Ministry of Justice*, [2016] UKSC 10 at para 24.

[231] It may also be sufficiently connected in other circumstances, see Goudkamp & Nolan, *supra* note 48 at para 21.026.

[232] Indeed, the tests were "assimilated" by the English courts, see *Lloyd v Grace, Smith & Co*, [1912] UKHL 606, [1912] AC 716, in a development that was "for the most part, [] accepted" by the Canadian courts: GHL Fridman, *Canadian Agency Law*, 3d ed (Toronto: LexisNexis, 2017) s 8.5.

[233] Benjamin Zipursky, "Civil Recourse, Not Corrective Justice Symposium: The New Negligence" (2003) 91:3 Geo LJ 695 at 721.