Canadian Journal of Law and Technology

Volume 19 | Number 1 Article 4

1-27-2023

Recognizing Operators' Duties to Properly Select and Supervise AI Agents – A (Better?) Tool for Algorithmic Accountability

Richard Zuroff

Recognizing Operators' Duties to Properly Select and Supervise AI Agents – A (Better?) Tool for Algorithmic Accountability

Richard Zuroff[*]

Keywords: artificial intelligence, explainability, algorithmic accountability, GDPR, Bill C-11, PIPEDA reform, Digital Charter Implementation Act

Abstract

In November of 2020, the Privacy Commissioner of Canada proposed creating GDPR-inspired rights for decision subjects and allowing financial penalties for violations of those rights. Shortly afterward, the proposal to create a right to an explanation for algorithmic decisions was incorporated into Bill C-11, the Digital Charter Implementation Act. This commentary proposes that creating duties for operators to properly select and supervise artificial agents would be a complementary, and potentially more effective, accountability mechanism than creating a right to an explanation. These duties would be a natural extension of employers' duties to properly select and retain human employees. Allowing victims to recover under theories of negligent hiring or supervision of AI-system-as-agents would reflect their increasing (but less than full) autonomy and avoid some of the challenges that victims face in proving the foreseeability elements of other liability theories.

INTRODUCTION

It is now well-recognized that the use of artificial intelligence (AI) and algorithmic decision-making technologies[1] can bring significant social and

---

[*] BCL (McGill), JD (McGill), MBA (McGill). Director of AI Success, DataRobot. The views expressed are personal and not necessarily the views of DataRobot.

[1] This commentary treats AI and algorithmic decision-making interchangeably, and henceforth will refer just to AI. These terms refer to software systems that are capable of creating heuristics from data to make predictions and take decisions with some degree of autonomy. In many cases, this mapping from inputs to outputs (such as predictions or decisions) is learned by using machine learning algorithms to train models such as (deep) neural networks. However, not all AI uses machine learning, and scientists are exploring hybrid approaches that combine machine learning with symbolic processing approaches. (See e.g. Jiayuan Mao et al, "The Neuro-symbolic Concept Learner: Interpreting Scenes, Words, And Sentences From Natural Supervision" (2019) arXiv 1904.12584 online: < https://arxiv.org/abs/1904.12584 > .) Nonetheless, for this commentary, the precise algorithms being implemented in AI systems matter less than the fact that they often operate with some degree of autonomy and are more complex and opaque than software that executes simple, fixed business rules. This piece also does not distinguish between pure software implementations and robots, as the main challenges posed by artificial agents for the law are independent of whether they are physically embodied or not. In

economic benefits, while also creating new challenges related to privacy, security, ethics, trust, and accountability.[2] In Canada, one reaction to these challenges is the *Digital Charter Implementation Act* (DCIA),[3] which would provide the subjects of automated decision-making with rights modelled on Europe's *General Data Protection Regulation* (GDPR).[4] This commentary argues that, while it is useful for the federal government to create explainability rights and other data protection reforms, it would be equally, if not more, beneficial for Canada's provinces to follow the example of European policymakers in creating a duty of care for the operators of AI systems to properly select and supervise these systems. This would allow victims to recover under theories of negligent selection or supervision of AI agents, which would hold the operators of such systems accountable for algorithmic harms and incentivize them to invest in testing and monitoring processes that could decrease the risk of harm.

The first section of this commentary reviews the limited evidence from Europe that explainability rights, even when accompanied by the possibility of administrative fines, actually increase accountability for AI operators, which suggests the need for complementary mechanisms. The second section outlines one such approach: extending employers' duties to properly hire and retain human employees into operators' duties to properly select and supervise AI systems in ways that maintain equivalent liability patterns between firms' use of human and artificial agents. The third section suggests two benefits of enabling claims for negligent selection and supervision of AI. Treating AI systems as economic agents of employers has the theoretical benefit of appropriately reflecting how the technology is operationalized — since operators do and will deploy them with more autonomy than simple tools — without needing to assign AI systems any legal personhood. Creating these duties can also provide the practical benefit of avoiding some of the challenges that victims face in proving the foreseeability elements of other types of AI liability claims.

---

support of ignoring this distinction, see Jack Balkin, "The Path of Robotics Law" (2015) 6:45 Calif. L. Rev. Circ. at 46 (SSRN).

2 See e.g. G20 Trade Ministers and Digital Economy Ministers, "G20 Ministerial Statement on Trade and Digital Economy" (June 9, 2019) online: < https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf > .

3 Bill C-11, Digital Charter Implementation Act, 2020, 2nd Session, 43rd Parliament, 2020 [DCIA].

4 General Data Protection Regulation, Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, [2016] OJ L 119 at art. 3 [GDPR].

1. THE LIMITATIONS OF TRANSPARENCY RIGHTS IN PROMOTING ACCOUNTABILITY FOR AI SYSTEMS

GDPR aims to promote accountability by providing the subjects of algorithmic decisions with a set of rights and allowing organizations to be fined for violations of those rights. These rights include the right to request human intervention or challenge an automated decision[5] and the right to receive meaningful information about the logic involved in a decision[6] (commonly referred to as the right to an explanation[7]). Canada does not have any general-purpose algorithmic processing laws that are as extensive as GDPR, but the *Personal Information Protection and Electronic Documents Act* (PIPEDA)[8] governs how private sector organizations collect, use, and disclose personal information. The Privacy Commissioner of Canada recommended that PIPEDA be reformed to provide individuals with "a right to a meaningful explanation of, and a right to contest, automated decision-making under PIPEDA."[9] The proposed right to an explanation was explicitly modelled on GDPR,[10] and it "provides an avenue of recourse and respects basic human dignity by ensuring that the organization is able to explain the reasoning for the particular decision in understandable terms."[11] This proposal was partially enacted in the DCIA (Bill C-11). The first reading of this bill does not include a right to object to or opt out of automated decision-making, but does provide the right to an explanation: "[i]f the organization has used an automated decision system to make a prediction, recommendation or decision about the individual, the organization must, on request by the individual, provide them with an explanation of the prediction, recommendation or decision and of how the personal information that was used to make the prediction, recommendation or

---

5 GDPR, art. 22(3).

6 GDPR, art. 15(1)(h).

7 The more expansive term "right to an explanation" is grounded in a combined reading of GDPR's Articles and Recitals. Recital 71 clarifies that suitable safeguards for automated processing should include "the right... to obtain an explanation of the decision reached after such assessment and to challenge the decision." While only the Articles have direct force of law, the Recitals are cited by courts as authoritative interpretations, and the Guidelines have been promulgated by a group representing the Data Protection Authorities who will actually enforce the law. See Margot E. Kaminski, "The Right to Explanation, Explained" (2019) 34:1 Berkeley Tech. L.J. 189 (SSRN).

8 *Personal Information Protection and Electronic Documents Act*, R.S.C. 2000, c. 5.

9 Office of the Privacy Commissioner of Canada, "A Regulatory Framework for AI: Recommendations for PIPEDA Reform" (Gatineau, Q.C.: Office of the Privacy Commissioner of Canada, November 2020) online: < https://priv.gc.ca/en/about-the-opc/what-we-do/consultations/completed-consultations/consultation-ai/reg-fw_202011/ > .

10 "The right would be similar to what is found in Article 15(1)(h) of the GDPR, which requires data controllers to provide individuals with 'meaningful information about the logic involved' in decisions." *Ibid.*

11 *Ibid.*

decision was obtained."[12] The legislation also proposes to give the Privacy Commissioner the ability to require an organization to modify its practices[13] but, unlike GDPR, does not provide for the possibility of administrative fines for contraventions of the right to an explanation.[14]

Creating the possibility of penalties for organizations that fail to enact appropriate safeguards (such as providing explanations) is one potential mechanism for creating accountability for algorithmic harms. However, critics have pointed out that GDPR's transparency rights alone are unlikely to redress many of the harms that might be caused by AI systems,[15] nor will they strongly incentivize organizations deploying AI to proactively avoid causing harm since decision subjects rarely have the time or expertise to meaningfully make use of their individual rights.[16] Individual rights place the burden on decision subjects to request explanations or challenge decisions, so the possibility of fines or penalties will only be as effective in shaping organizations' behaviour as individuals are active in making requests. Two years of data on GDPR enforcement suggests that the criticism that transparency rights are a weak mechanism for promoting accountability may be valid: as of December 2020, only 35 fines have ever been levied for failures to fulfill data subjects' right to an explanation.[17] This low level of fine activity could be explained either by high levels of compliance by data processors or limited exercise of these rights by decision subjects. Currently, limited rights exercise is a much more plausible explanation as surveys show that less than half of companies consider themselves to be "fully" or "very" GDPR compliant.[18]

---

[12] DCIA, *supra* note 3 at 63(3).

[13] *Ibid*. at 92(2).

[14] *Ibid*. at 93(1) lists several subsections under which the Commissioner may decide to impose a penalty, but section 63(3) is not one of them. However, an individual may bring an action for damages if the Commissioner has found a contravention of the act. See *ibid*. at 106(1).

[15] Lilian Edwards & Michael Veale, "Slave to the Algorithm? Why a 'right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16:1 Duke L & Tech Rev. 18 at 42 (SSRN).

[16] *Ibid*. at 67.

[17] CMS, "GDPR Enforcement Tracker" accessed 1 December 2020 online: <https://www.enforcementtracker.com/>.

[18] Example survey results are from International Association of Privacy Professionals & Ernst and Young, "IAPP-EY Annual Privacy Governance Report 2019" (2019) online: *IAPP* <https://iapp.org/store/books/a191P000003Qv5xQAC/>. Another survey found that, as of 2020, only 55 percent of respondents said they are now ready for GDPR. See Cisco, "Data Privacy Benchmark Study 2020: From Privacy to Profit: Achieving Positive Returns on Privacy Investments" (2020) at 11, online: <https://www.cisco.com/c/dam/global/en_uk/products/collateral/security/2020-data-privacy-cybersecurity-series-jan-2020.pdf>.

2. DEFINING THE DUTIES TO SELECT AND SUPERVISE ARTIFICIAL AGENTS

A complement to reforming data protection is to create or recognize bases for claims under which a manufacturer, operator, or user of an AI system could be liable when it causes some damage or harm. There are three potential regimes for claims: negligence (i.e., fault-based liability), products liability, and strict (i.e., no-fault) liability.[19] Of these different regimes, negligence had generally received the least attention.[20] However, in 2020, the European Parliament's Committee on Legal Affairs, following the advice of its Expert Commission,[21] approved a series of recommendations on AI liability rules, including that the operator of non-high-risk[22] AI systems should be subject to fault-based liability for any harm or damage. Under this proposal, an operator could avoid liability if he or she can show that "due diligence was observed by performing all the following actions: selecting a suitable AI-system for the right task and skills, putting the AI-system duly into operation, monitoring the activities and maintaining the operational reliability by regularly installing all available updates."[23]

This proposal for creating fault-based liability follows from the principle of functional equivalence, which holds that "victims of harm caused by the operation of emerging digital technologies receive less or no compensation compared to victims in a functionally equivalent situation involving human conduct and conventional technology."[24] That is, since a victim could recover based on the theory that an employer's failure to select an appropriate human employee or a failure to properly supervise an employee's activities breached a duty to the victim and caused a harm, the substitution of an algorithm for the

---

[19] This list excludes the possibility of holding the AI system itself responsible, which would require that the system have some degree of legal personhood. For an early consideration of legal personhood for AI, see Lawrence B. Solum, "Legal Personhood for Artificial Intelligences" (1992) 70 NC. L. Rev. 1231 (SSRN). The European Parliament also considered the potential of creating "electronic personality"; however, the current consensus is that legal personhood is not necessary. See Expert Group On Liability And New Technologies — New Technologies Formation, "Liability for Artificial Intelligence and other Emerging Digital Technologies" (2019), online: < https://ec.europa.eu/transparency/regexpert/index.cfm?do = groupDetail.groupMeetingDoc&docid = 36608 > at 37 [EC Expert Group], which cites an open letter from technologists and scholars opposing personhood.

[20] Andrew D. Selbst, "Negligence And AI's Human Users" (2020) 100 B.U. L. Rev. 1315 at 1318 (SSRN) [Selbst].

[21] EC Expert Group, *supra* note 19.

[22] Strict liability is recommended for operators of high-risk AI systems such as AI-driven robots in public spaces. *Ibid.* at 40.

[23] European Parliament Committee on Legal Affairs, "Report with recommendations to the Commission on a civil liability regime for artificial intelligence" (2020), online: < https://ec.europa.eu/transparency/regexpert/index.cfm?do = groupDetail.groupMeetingDoc&docid = 36608 > .

[24] EC Expert Group *supra* note 19 at 5. at 5.

employee should not enable the employer to avoid liability. Employers' duties have thus been translated into a duty of care for AI operators to suitably select, monitor, and maintain the system.

A similar duty of care and liability regime could be defined in Canada by extending employers' potential scope of direct liability for harms caused by employees. Canadian courts and commentators have focused mostly on defining the limits of employers' vicarious liability for employees' actions.[25] Under a vicarious liability theory, the employee has breached a duty of care, and the question is whether there was a sufficiently close nexus between the employee's employment conditions (such as creating an elevated risk of harm) and the harm that occurred to justify holding the employer liable.[26] Claims for direct negligence in employment rely on a different legal theory: that the employer breached a duty of care by failing to properly hire, supervise, or retain its employees, and that this failure caused a harm. The general rule is that an employer can be held directly liable for negligent hiring if the employer knew or should have known of the employee's potential risk or if reasonable investigation before hiring would have uncovered such a risk.[27] The test for negligent supervision or retention is similar but focuses on activities occurring after the employee had been hired. An employer can be liable for negligent supervision if the employer knew, or through a reasonable investigation should have known, that "the acts or omissions of its employee would subject third parties to an unreasonable risk of harm."[28]

There have been successful claims for negligent hiring in Canada,[29] although these are rarer than in the United States, which provides a richer set of fact patterns to outline the limits of employers' duties.[30] In the United States, cases

---

[25] See e.g. *671122 Ontario Ltd. v. Sagaz Industries Canada Inc.*, 2001 SCC 59, 2001 CarswellOnt 3357, 2001 CarswellOnt 3358 (S.C.C.), reconsideration / rehearing refused 2001 CarswellOnt 4155, 2001 CarswellOnt 4156 (S.C.C.), and *Bazley v. Curry*, 1999 CarswellBC 1264, 1999 CarswellBC 1265, [1999] 2 S.C.R. 534 (S.C.C.).

[26] See *John Doe v. Bennett*, 2004 SCC 17, 2004 CarswellNfld 75, 2004 CarswellNfld 76 (S.C.C.) [*Bennet*]. In *Bennet*, a church was found vicariously liable for the actions of a priest who had sexually abused young boys in his care. Even though there was no evidence that the employer was negligent in supervising the priest, the court found that vicarious negligence was still appropriate because the nature of his duties and position heightened the risk.

[27] For the law in the United States, see Restatement (Third) Of Agency § 7.05.

[28] Frank J. Cavico et al. "The Tort of Negligence in Employment Hiring, Supervision and Retention" (2016) 1:4 Am. J. Bus. 205 at 213 (AIS) [Cavico].

[29] See *Wilson v. Clarica Life Insurance Co.*, 2002 BCCA 502, 2002 CarswellBC 2078 (B.C. C.A.). In this case the employer (Clarica) had been advised by Wilson's former employer that they suspected he had stolen from them and a customer. Clarica was held to have been negligent in hiring Wilson after he stole funds from a customer of Clarica's.

[30] Negligent hiring is a tort claim recognized in more than half of the states in the United States. See Timothy L. Creed, "Negligent Hiring and Criminal Rehabilitation: Employing Ex-Convicts, Yet Avoiding Liability" (2017) 20 St. Thomas L. Rev. (SSRN).

tend to succeed where employers had notice (such as the results of background checks or criminal records) or should have known of the potential employees' dangerous nature or propensity to cause harm or damages.[31] In some domains, this potential for liability extends to a failure to ensure that employees have the appropriate skills. For instance, doctors must be credentialed by verifying their education, board certifications, and complaint history, and American plaintiffs have successfully sued hospitals under theories of negligent credentialing when doctors performed procedures with which they were unfamiliar.[32] Plaintiffs are often successful in negligent supervision or retention claims where there were previous reports of harassment,[33] but fail when the human employee's misconduct is unexpected and unpredictable based on their past behaviour. For instance, in a pregnancy discrimination case, "the plaintiff lost her claim for negligent retention because she presented no evidence that the employer knew or should have known of her supervisor's tendency to discriminate against pregnant women."[34]

There are thus two threshold questions in defining an operator's duty to properly select and supervise AI systems. First, can we define an appropriate investigation (such as an equivalent to credentialing) which should be done before deploying an AI system? Second, can we define what could constitute a pattern of bad behaviour by an AI system that would provide notice to a company that continuing to use the system puts third parties at unreasonable risk?

As to the first question, what constitutes an appropriate investigation to select an AI system will likely depend on the domain of application, but parallels with the screening of human actors have already been suggested in medicine. Building on hospitals' duty to provide well-functioning equipment for patient

---

[31] Cavico, *supra* note 28 at 209. However, blanket policies requiring all current employees to participate in a mandatory criminal background check have been struck down by Canadian labor arbitrators. See *Rouge Valley Health System and ONA (13-40), Re*, 2015 CarswellOnt 16480 (Ont. Arb.).

[32] Negligent certification may be based on a failure to have board certifications at all or on a lack of more specific qualifications. In one case, the on-call physician had not treated a fracture in three years but was called into the hospital to perform the procedure. Complications led to the plaintiff's leg being amputated and a jury awarded damages, of which 80 percent were to come from the hospital. See *Darling v. Charleston Community Hospital*, 211 N.E.2d 253 (1965), discussed in Morgan Haefner, "How hospital and physician leaders can prevent negligent credentialing lawsuits" (10 October 2019), online: *Becker's Hospital Review* <https://www.beckershospitalreview.com/legal-regulatory-issues/how-hospital-and-physician-leaders-can-prevent-negligent-credentialing-lawsuits.html>. In Canada, plaintiffs have generally used vicarious liability theories rather than direct liability for hospitals' negligent certification. Vicarious liability claims generally fail as doctors are considered independent contractors. *See Yepremian v. Scarborough General Hospital*, 1980 CarswellOnt 612, 110 D.L.R. (3d) 513 (Ont. C.A.), leave to appeal allowed (1980), 120 D.L.R. (3d) 337 (note) (Ont. C.A.)

[33] Cavico, *supra* note 28 at 216.

[34] Cavico, *supra* note 28 at 214.

care and doctors' responsibilities to treat patients with due expertise and care, it has been suggested that hospitals or doctors might be liable for negligently choosing, implementing, and using black-box medical systems.[35] Under this standard of care, facilities and clinicians would have a duty to validate systems before using them, either through procedural mechanisms (such as assessing the qualifications of the developers) or computationally (such as replicating results on parallel data[36]) for higher risk applications. The technical community has also begun to focus on various ways to make it easier to verify AI developers' claims, with a focus on providing evidence about "the safety, security, fairness, and privacy protection of AI systems."[37] As the methods to produce independent evidence about AI systems' capabilities and limitations mature, it would be increasingly reasonable to expect firms to use these investigative tools before deploying AI systems, and to hold them negligent for failures to do so.

As to the second question, the case law on negligent supervision of human employees suggests that an equivalent test could be developed for whether a firm had notice of the risk posed by an AI system, despite issues with the opacity and complexity of the underlying algorithms. Many commentators have raised the difficulty of proving the foreseeability of harms from AI due to these systems' abilities to identify patterns in data beyond human recognition[38] and to learn continuously from unpredictable environments.[39] Crucially, the foreseeability challenge (discussed more in section III) applies especially to the *first* victim harmed by an AI system. Even if the first instance of harm by a system was unpredictable, once it occurred an organization would be on notice of the issue. If it failed to address the issue, future harms would be foreseeable, so future victims could point to past harms as providing notice to the employer. This could provide a coherent theory under which plaintiffs could be allowed to recover for negligent supervision or negligent retention.

This would create a pattern of potential recovery similar to human cases, where the foreseeability of conduct is sometimes narrowly construed. For example, a plaintiff who was the victim of racial harassment in the United States lost their negligent retention case even though the employer was aware that the harasser had a physical altercation with another employee. The court held that

---

[35] William Nicholson Price II, "Medical Malpractice and Black-Box Medicine" in Glenn Cohen et al., eds., *Big Data, Health Law, and Bioethic*s (Cambridge: Cambridge University Press, 2018).

[36] William Nicholson Price II, "Big Data, Patents, and the Future of Medicine" (2017) 37 Cardozo L. Rev. 1401 at 1417 (SSRN).

[37] Miles Brundage et al, "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims" (2020) arXiv 2004.07213 at 2, online: <https://arxiv.org/pdf/2004.07213.pdf>.

[38] Selbst, *supra* note 20 at 1318.

[39] Peter M. Asaro, "The Liability Problem for Autonomous Artificial Agents" (2016) AAAI Spring Symposium Series at 4, online: <https://icps.gwu.edu/sites/g/files/zaxdzs1736/f/downloads/Asaro%201.pdf>.

because the previous incident was with an employee of the same race as the harasser, the later racially motivated harassment was not foreseeable to the employer.[40] Focusing a test for negligent supervision of AI agents on whether there were similar past incidents would thus uphold the contrapositive of the functional equivalence principle introduced above. Just as victims should not be prevented from recovering just because a firm uses an algorithmic system, a victim who would *not* have recovered in a functionally equivalent situation involving a human employee should presumably not be entitled to *more* compensation simply because of the substitution of AI technologies. (Of course, other remedies than liability would then be needed to help the first victims of AI systems.)

This brief review suggests that a duty of care to properly select and supervise algorithmic agents could reasonably be defined by courts or policymakers even as technical capabilities in the field continue to evolve quickly. Creating new grounds for liability would expand victims' ability to recover, but might also disincentivize the developers of AI systems, which in turn could delay the social and economic benefits they are expected to deliver.[41] Therefore, a natural next question is what benefits Canadians could expect if these duties are recognized, which is addressed in section III.

3. THE BENEFITS OF RECOGNIZING THESE DUTIES

Allowing direct negligence claims for failures to properly select and supervise algorithmic agents would: (a) provide redress for harms from AI systems that have more autonomy than simple decision-support tools by recognizing them as agents without legal personhood; (b) reduce the foreseeability barriers to recovery and the uncertainty facing producers by focusing on whether reasonable tests and validations have been performed.

(a) Placing Semi-Autonomous AI Systems in the Appropriate Legal Category - Agents

Scholars have suggested that AI systems, depending on their autonomy, unpredictability, and social capabilities, should be treated like entities in the most similar legal category, which span from tools, to wild animals, to domesticated animals, to children, to adult people.[42] Most scholarship focuses on one extreme

---

[40] Cavico, *supra* note 28 at 214.

[41] An EU analysis suggests that clear liability rules which promote investment and dissuade risky behaviour could generate €498.3 billion in added value for the EU economy by 2030 if "broader impacts are also considered, including reduced numbers of accidents, health and environmental impacts and user impacts." Tatjana Evas, "Civil liability regime for artificial intelligence European added value assessment" (2020) at I, online: *European Parliament* < https://www.europarl.europa.eu/RegData/etudes/STUD/2020/654178/EPRS_STU(2020)654178_EN.pdf > .

[42] Ignacio N. Cofone, "Servers and Waiters: What Matters in the Law of A.I." (2018) 21 Stan. Tech. L. Rev. 167 (SSRN).

of this continuum or the other. Some authors consider the benefits and challenges of providing AI systems with legal personhood,[43] while others focus on products liability because they assume there is no "person" to be negligent when autonomous systems cause injuries.[44] A third approach argues that, because the majority of AI systems deployed today only make recommendations to human agents who remain in control, the key challenge is adapting negligence rules to humans' use of these decision-assistance tools.[45] However, this type of strategy creates the risk that users are held responsible (as "moral crumple zones"[46]) for accidents when they have limited control over semi-autonomous systems. Moreover, even if some AI tools are used by human agents who remain in control of decisions, the trend is toward increasing autonomy.[47] Today, some algorithms already perform tasks with minimal decision-by-decision intervention from people, such as analyzing and trading securities.[48] Furthermore, the economic impetus for using AI is highest precisely where there are benefits from taking decisions or actions at superhuman speed, volume, or precision, with the goal of maximizing some objective defined by the deployer of AI.[49]

A fourth approach to AI liability, advocated for here, is to focus on the intersection of negligence and agency law. AI systems fulfill many of the canonical characteristics of economic agents, with the operator- or user-organization as their principals.[50] In an agency relationship, the principal engages the agent to perform some service on their behalf which involves

---

[43] See Solum, *supra* note 19.

[44] Omri Rachum-Twaig, "Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots" (2020) Univ. Ill. Law Rev. (forthcoming) (SSRN).

[45] See Selbst, *supra* note 20 at 1319.

[46] Madeleine Clare Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction" (2019) 5 Engaging Science, Technology, & Society (ESTS), DOI: < 10.17351/ESTS2019.260 > .

[47] For a contrary argument that autonomy is not a desirable differentiator between products and "thinking algorithms," see Karni A. Chagal-Feferkorn, "Am I An Algorithm Or A Product? When Products Liability Should Apply To Algorithmic Decision-Makers" (2019) 30 Stan. L. & Pol'y Rev. 61 (SLS).

[48] For instance, research has compared the performance of AI systems for portfolio management to human advisors and found that robo-analysts produced a more balanced distribution of buy, hold, and sell recommendations. "Portfolios formed based on the buy recommendations of Robo-Analysts appear to outperform those of human analysts." Braiden Coleman, Kenneth J. Merkley & Joseph Pacelli, "Man versus Machine: A Comparison of Robo-Analyst and Traditional Research Analyst Investment Recommendations" (2020) [unpublished], online: *SSRN* < http://dx.doi.org/10.2139/ssrn.3514879 > .

[49] See e.g. Ajay Agrawal, Joshua S. Gans, & Avi Goldfarb, "Prediction, Judgment, and Complexity: A Theory of Decision Making and Artificial Intelligence" (2018) NBER Working Papers 24243, online: < https://www.nber.org/papers/w24243 > .

[50] For an early proposal to treat AI systems as agents, see Samir Chopra & Laurence F. White, *A Legal Theory for Autonomous Artificial Agents* (Ann Arbor: University of Michigan Press, 2011).

delegating some decision-making authority to the agent who therefore has more information about the way the task will be carried out. [51] For human agents, this creates monitoring costs for the principal, and many analyses explore ways to incentivize the agent to align with the principal's preferences.[52] Firms set the goals and incentive structures for human agents but often have limited control over how these objectives are achieved. This is functionally equivalent to the situation with AI systems, which are typically given an objective, learn the best way to achieve it within some constraints,[53] and require ongoing monitoring during operations to ensure they continue to perform well. Developers and operators can sometimes fail to fully align the behaviour and incentives of an AI system with their own, just as employers can encourage bad behaviour from their employees.[54] Thus, treating AI systems in the legal category of "agents" — by creating an equivalence with human employees — appropriately reflects that their degrees of independence and unpredictability are often higher than tools but lower than unconstrained adult human actors.

---

[51] Michael C. Jensen & William H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure" (1976) 3:4 J. Financial Economics (Science Direct), DOI: < https://doi.org/10.1016/0304-405X(76)90026-X > .

[52] See e.g. Peter-Jürgen Jost, "Monitoring in Principal-Agent Relationships" (1991) 147:3 J Institutional & Theoretical Economics (JSTOR).

[53] Artificial agents built with reinforcement learning are given an objective (such as maximizing the number of points scored or completing a game as quickly as possible) and learn the best actions to take from interacting with the environment. Systems that use supervised learning paradigms are trained on input-output pairs and have the objective of minimizing the prediction error on future inputs.

[54] "Specification" refers to defining an AI system's goal in a way that ensures its behaviour aligns with the human operator's intentions. See Victoria Krakovna et al, "Specification gaming: the flip side of AI ingenuity" (21 April 2020), online: *DeepMind Research* < https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity > . Specification failures can cause unintended consequences. For instance, because YouTube's algorithm was designed to optimize for engagement and maximize viewing time, the recommendation algorithm sometimes steered users toward increasingly extremist content. See Mark Bergen, "YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant" (2 April 2019), online: *Bloomberg* < https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant?sref = jmcj2Pta > . Specification failures occur between corporations and human employees as well. For example, Wells Fargo's aggressive sales and cross-sales targets led its employees to open as many as 2 million accounts without customer authorization. These accounts only generated roughly $2 million in additional revenue for the bank but led to approximately $3 billion in fines. See Brian Tayan, "The Wells Fargo Cross-Selling Scandal" (6 February 2019), online: *Harvard Law School Forum on Corporate Governance* < https://corpgov.law.harvard.edu/2019/02/06/the-wells-fargo-cross-selling-scandal-2/ > .

(b) Avoiding the Challenges with Proving Foreseeability Faced by Other Liability Theories

As discussed in section II, under a direct liability theory for negligent selection or supervision of AI agents all but the first victim harmed by an AI system should be able to point to past errors by the system as evidence that the operator did or could foresee the risk of continuing to operate the AI system. In contrast, other liability theories — strict liability (especially vicarious liability), negligent use of AI as a tool, and products liability — may raise significant challenges for victims in proving the foreseeability elements of claims.

As set out by the Supreme Court's ruling in *Bazley v. Curry*,[55] vicarious liability is a type of strict liability that courts should apply where there is a significant connection between the creation or enhancement of risk and the wrong that flows from the risk. A plaintiff looking to hold the operator of an AI system vicariously liable could argue under this theory that in placing decision-making authority with an AI system, the operator created or enhanced the risk of a harm (such as physical injury or discrimination[56]). Just as employers cannot point to the fact that an employee's actions contravened their ethical principles to escape liability,[57] the operator of an AI system could not point to internal responsible development principles or ethics codes to fully absolve themselves of liability.[58] However, human workers performing manual processes have their own biases, so it is not obvious that the mere introduction of an algorithmic system increases the risk of harm. The ability of a victim to recover under this theory might depend on the specific facts of whether the system's performance was actually worse than, equal to, or better than the human equivalent. For instance, one study found that lenders in the United States charged otherwise-

---

[55] *Supra* note 25.

[56] Discrimination claims in particular may be challenging because of the lack of a freestanding tort of discrimination as exists in other jurisdictions (see *Bhadauria v. Seneca College of Applied Arts & Technology*, 1981 CarswellOnt 117, 1981 CarswellOnt 616 (S.C.C.)). Plaintiffs in Canada have had to rely on the human rights framework, which placed jurisdiction with the Canadian Human Rights Tribunal rather than the courts. However, in *Lewis v. WestJet Airlines Ltd.*, 2019 BCCA 63, 2019 CarswellBC 318 (B.C. C.A.), leave to appeal refused *WestJet Airlines Ltd. v. Mandalena Lewis*, 2019 CarswellBC 2092, 2019 CarswellBC 2093 (S.C.C.), the B.C. Court of Appeal found that civil claims for workplace discrimination could at least be properly considered by the courts when framed as a breach of an employment contract, and the Supreme Court of Canada dismissed the employer's leave to appeal in *WestJet Airlines Ltd. v. Mandalena Lewis*, 2019 CarswellBC 2092, 2019 CarswellBC 2093 (S.C.C.).

[57] See *Bennett*, *supra* note 26, where the court negatively cited a previous decision in which an episcopal corporation was held not vicariously liable for sexual assaults committed by one of its priests because the acts were contrary to its religious tenets.

[58] For one of many proposals that corporations develop in-house AI ethics codes, see Darrell M. West, "The role of corporations in addressing AI's ethical dilemmas" (2018), online: *Brookings Institute* < https://www.brookings.edu/research/how-to-address-ai-ethical-dilemmas/ > .

equivalent Latinx/African-American borrowers higher rates for purchase mortgages, and that, while the use of algorithms did not eliminate discrimination, "algorithmic lenders do reduce rate disparities by more than a third and show no discrimination in rejection rates."[59] In this type of case, even where a loan applicant suffered a harm such as paying a higher rate, it might be difficult to prove that the introduction of an algorithmic system created or enhanced that risk to succeed in a vicarious liability claim. A benefit of allowing direct negligence claims for hiring and supervising AI agents is that it avoids this comparative question of whether the AI agent was worse than the human equivalent[60] and instead focuses on past behaviour of the system.

Proving the foreseeability or causation elements of claims can be difficult because AI algorithms are often complex, opaque, and potentially protected by trade secrets from outside disclosure.[61] That is, AI systems can be "black boxes" to both their victims and users,[62] which can make it difficult to prove negligence under a theory that treats AI systems as tools which operators should take appropriate care in using. When AI systems are designed to use large amounts of data, operate at high speeds, or otherwise exceed human capabilities, "users may often be unable to determine in real time whether the AI is making an error. In those cases, it will often be unclear how a user can satisfy any duty of care in the operation of the AI."[63] If the risk of harm is not responsive to a user's level of care, it becomes questionable whether applying negligence for breaching that duty of care is a coherent policy.[64] However, the same weakness does not apply to a theory of negligent selection and monitoring. *Before* the AI is put into operation, the risk of error can be reduced (although likely not eliminated) by rigorous testing and validation. Expert testimony could be provided about the reasonably attainable error elimination at the time of design and sale of an AI system to help define a standard of care.[65] Similarly, the ability to monitor the performance of AI systems *after deployment* is still maturing, but tools already

---

[59] Robert Bartlett et al, "Consumer-Lending Discrimination in the FinTech Era" (2019) NBER Working Papers 25943, online: < https://www.nber.org/papers/w25943 > .

[60] In this regard it also avoids relatively arbitrary performance cut-offs for finding a breach of duty. For instance, in the products liability context, one proposal was that autonomous vehicles should be considered not defective where aggregate data shows that a car is at least twice as safe as human drivers. Mark A. Geistfeld, "A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation" (2017) 105 Calif. L. Rev. 1611 at 1653 (SSRN).

[61] See e.g. Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation" (2018) 31:2 Harv. J.L. & Tech 889 (JOLT).

[62] See e.g. Davide Castelvecchi, "Can We Open the Black Box of AI?" (5 October 2016), online: *Nature* < https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731 > .

[63] See Selbst, *supra* note 20 at 1331.

[64] See Selbst, *supra* note 20.

[65] William D. Smart, Cindy M. Grimm, & Woodrow Hartzog, "An Education Theory of Fault for Autonomous Systems" (2017) [unpublished], online: < http://people.oregon-

exist,[66] and many categories of threats to reliable operation are already known.[67]

To further help plaintiffs overcome the evidentiary barriers of assessing the performance of AI systems, Canadian policymakers could adopt another proposal from the European Expert Commission, that AI "should come with logging features, where appropriate in the circumstances, and failure to log, or to provide reasonable access to logged data, should result in a reversal of the burden of proof in order not be to the detriment of the victim."[68] Expert testimony could identify reasonably attainable risk mitigation practices at the time of operation even if it would be impossible or impractical for humans to search through every line of code to infer every potential harm. As in the medical domain, the standard of care for selecting and supervising AI agents could be modified based on the operator's resources. "[A] physician practicing in a small rural hospital will not be required to use the same specialized equipment available to the most well-resourced urban medical centers,"[69] and, equivalently, larger organizations might be expected to apply more advanced investigations of their AI systems.

The developer of an AI system might also be a target for claims either under a negligence theory or products liability theory. Some authors argue that neither the inherent unpredictability of AI systems which continue to learn after they are sold, nor the potential intervening negligence by operators or other parties would be sufficient for developers to completely escape liability by arguing that the harm caused by an AI system was unforeseeable.[70] However, other scholars see

---

state.edu/~smartw/papers.php?q=papers& display=detail&tag=werobot2017> referenced in Selbst, *supra* note 20.

66 For example, see a description of the functionality in Amazon SageMaker in Julien Simon, "Amazon SageMaker Model Monitor — Fully Managed Automatic Monitoring For Your Machine Learning Models" (3 December 2019), online: *AWS News Blog* <https://aws.amazon.com/blogs/aws/amazon-sagemaker-model-monitor-fully-managed-automatic-monitoring-for-your-machine-learning-models/>. Other providers such as Microsoft and Google have similar tools.

67 See e.g. Andrew Marshall et al, "Threat Modeling AI/ML Systems and Dependencies, Microsoft AETHER Engineering Practices for AI Working Group" (11 November 2019), online: *Microsoft Security* <https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml>.

68 Selbst, *supra* note 20 at 4. System logging might include metrics such as "drift," which measures divergence between the distributions in training data compared to operational input data that is being used as the basis for predictions. Divergence can be an early warning sign that system performance is degrading, and thus might constitute notice that continued operation of the system without an intervention is increasing the risk exposure of others. Logging might also include relative performance metrics over time for different segments defined by sensitive variables such as gender, age, or race if this information is available.

69 William Nicholson Price II, *supra* note 35 at 7.

70 Weston Kowert, "The Foreseeability of Human—Artificial Intelligence Interactions" (2017) 96 Tex. L. Rev. 181, online: *Texas Law Review* <https://texaslawreview.org/foreseeability-human-artificial-intelligence-interactions/>.

foreseeability as a major challenge for plaintiffs seeking to recover from developers.[71] Even in a strict products liability regime the scope of liability is limited to where a defect is the proximate cause of the harm.[72] The proximate cause doctrine imposes several types of restrictions, including that the harm comes about in a foreseeable manner,[73] such as that the AI made an error which should have been anticipated and tested for. However, the foreseeability of a specific error may be difficult to prove,[74] whereas the analysis under a duty to supervise theory could focus on the operator's procedural rigour (or lack thereof) in testing and monitoring. Rigour might be easier to assess in terms of common industry practices, which would reduce the evidentiary burden for victims. Of course, imposing the duties to properly select and supervise AI agents would be just one of many potential enhancements to liability schemes[75] that can help address the practical challenges of enabling victims to recover from the harms created by AI systems.

CONCLUSION

Reforming PIPEDA to ensure Canada's privacy legislation keeps pace with the increasing volume and variety of data collection and usage, including by AI systems, is to be welcomed. Nonetheless, evidence from Europe where decision subjects have for several years enjoyed the rights to challenge algorithmic decisions and have their logic explained suggests that additional forms of protection are required to avoid creating accountability gaps for the harms that might be caused by AI technologies. Canadian policymakers should consider recognizing a duty for operators to properly select and supervise AI systems as a natural extension of employers' duties to properly hire and retain workers. Enabling victims to recover for failures to properly select and supervise AI systems as the economic agents of their employer (i.e., the operator) would provide a legal theory that is fit for the nature of AI systems that are more autonomous and unpredictable than mere tools, even if few systems today are fully autonomous. As a basis for direct liability, these duties would avoid the foreseeability challenges facing some other legal mechanisms for promoting

---

[71] See Matthew U. Scherer, "Regulating Artificial intelligence Systems: Risks, Challenges, Competencies, and Strategies" (2016) 29:2 Harv. J.L. & Tech. stating that it is "all but certain that issues pertaining to unforeseeable AI behavior will crop up with increasing frequency," cited in *ibid* at n. 8.

[72] David A. Fischer, "Products Liability— Proximate Cause, Intervening Cause, and Duty" (1987) 52 Mo. L. Rev. 547 at 559-60, online: *University of Missouri School of Law* < https://scholarship.law.missouri.edu/facpubs/186/ > .

[73] *Ibid.*

[74] See Kyle Graham, "Of Frightened Horses and Autonomous Vehicles: Tort Law and Its Assimilation of Innovations" (2012) 52 Santa Clara L. Rev. 1241, cited in Selbst, *supra* note 20.

[75] See Selbst, *supra* note 20 at 1326 for a discussion of potential enhancements to products liability specifically in the context of autonomous vehicles.

accountability and would create recovery patterns that align with the principle of functional equivalence. While no one regime will address all the conceptual and practical challenges raised by AI technologies, direct negligence has a role to play alongside other tools (such as administrative fines for violations of transparency rights, strict liability, and product liability) in striking the right balance between incentivizing innovation and fairly distributing the risk from AI systems.