Michigan Technology Law Review

Article 2

2021

How Can I Tell if My Algorithm Was Reasonable?

Karni A. Chagal-Feferkorn

University of Haifa Faculty of Law; University of Ottawa and the University of Ottawa Centre for Law, Technology and Society

Recommended Citation

Karni A. Chagal-Feferkorn, *How Can I Tell if My Algorithm Was Reasonable?*, 27 MICH. TECH. L. REV. 213 (2021).
Available at: https://repository.law.umich.edu/mtlr/vol27/iss2/2

HOW CAN I TELL IF MY ALGORITHM WAS REASONABLE?

*Karni A. Chagal-Feferkorn**

Abstract

Self-learning algorithms are gradually dominating more and more aspects of our lives. They do so by performing tasks and reaching decisions that were once reserved exclusively for human beings. And not only that—in certain contexts, their decision-making performance is shown to be superior to that of humans. However, as superior as they may be, self-learning algorithms (also referred to as artificial intelligence (AI) systems, "smart robots," or "autonomous machines") can still cause damage.

When determining the liability of a human tortfeasor causing damage, the applicable legal framework is generally that of negligence. To be found negligent, the tortfeasor must have acted in a manner not compliant with the standard of "the reasonable person." Given the growing similarity of self-learning algorithms to humans in the nature of decisions they make and the type of damages they may cause (for example, a human driver and a driverless vehicle causing similar car accidents), several scholars have proposed the development of a "reasonable algorithm" standard, to be applied to self-learning systems.

To date, however, academia has not attempted to address the practical question of how such a standard might be applied to

[*] Fellow, Haifa Center for Law & Technology, University of Haifa Faculty of Law; Scotiabank Postdoctoral Fellow at the AI + Society Initiative, University of Ottawa and the University of Ottawa Centre for Law, Technology and Society. I dedicate this paper to my mentor and role-model Niva Elkin-Koren for her ongoing support, which made this research possible. I would also like to thank Ido Feferkorn for his significant contribution to my research, and Florian Martin-Bariteau, Teresa Scassa and Michael Geist for their guidance and support. I also thank Yonatan Arbel, Woodrow Barfield, Michal Gal, Shlomi Hod, Dalit Ken-Dror, Bret Liney, Maayan Perel, Tito Rendas, Henrique Sousa Antunes, Jerrold Soh, Daniel Sokol, Steven van Uytsel, Erin Wall, the editors of the *Michigan Technology Law Review*, and the participants of the "Multidisciplinary Perspectives on Algorithms" conference at Kyushu University, "The Law of Artificial Intelligence" conference at Católica University, Lisbon, and the "Harmful Online Activity and Private Law" conference at Haifa University for their helpful comments. Any mistakes or omissions in this paper are the author's. This research was supported by the Center for Cyber Law & Policy (CCLP), established by the University of Haifa in collaboration with the Israeli National Cyber Bureau and by the Scotiabank Fund for AI and Society.

algorithms, and what the content of analysis ought to be in order to achieve the goals behind tort law of promoting safety and victims' compensation on the one hand, and achieving the right balance between these goals and encouraging the development of beneficial technologies on the other.

This Article analyzes the "reasonableness" standard used in tort law in the context of the unique qualities, weaknesses, and strengths that algorithms possess comparatively to human actors and also examines whether the reasonableness standard is at all compatible with self-learning algorithms. Concluding that it generally is, the Article's main contribution is its proposal of a concrete "reasonable algorithm" standard that could be practically applied by decision-makers. This standard accounts for the differences between human and algorithmic decision-making. The "reasonable algorithm" standard also allows the application of the reasonableness standard to algorithms in a manner that promotes the aims of tort law while avoiding a dampening effect on the development and usage of new, beneficial technologies.

Table of Contents

Introduction

The clock has struck midnight, and the tired physician on call at the hospital sees a patient suffering from meningitis. Having asked the patient whether she has any allergies and receiving a negative answer, the physician immediately administers penicillin-based antibiotics. Unfortunately, scribbled in the long medical record the patient brought with her was a note that she was indeed allergic to penicillin. If the patient then suffers significant physical damage as a result of the decision to administer penicillin-based antibiotics, is the physician legally liable for the resulting damage?

Damages caused by human tortfeasors are judged under the well-established framework of negligence. Under the negligence analysis, a wrongdoer is liable for damages if all four elements of establishing negligence exist.[1] One of these elements is the "breach of a duty of care," which is determined by scrutinizing whether a "reasonable person" would have made comparable decisions under similar circumstances. In cases dealing with professional negligence, such as medical malpractice, "reasonableness" is evaluated in comparison to the decisions expected from a reasonable professional in the same field, rather than from an ordinary person.[2]

The reasonableness analysis is advantageous in part because it allows courts to promote desired goals and social behaviors in a flexible manner.[3] In general, the two main aims of tort law are to compensate victims for harm suffered and to encourage potential wrongdoers to minimize risks.[4] At the same time, these rationales must be balanced with considerations of efficiency and the potential chilling effect a tort framework may have on the development and usage of socially-desirable practices or technologies.[5] In our meningitis example, modern tort analysis considers the desire to encourage physicians to be more careful in their work and granting the patient compensation for the injury she suffered, both of which favor holding the physician liable for administrating antibiotics without thoroughly reviewing the patient's medical record. But tort analysis would also be reluctant to cause inefficiency in the healthcare system by requiring physicians to read

[1] These consist of a duty of care by the wrongdoer to the injured; a breach of that duty; damage sustained by the injured; and a causal link between the breach of the duty and the damage suffered. *See* RESTATEMENT (SECOND) OF TORTS § 281 (AM. L. INST. 1965); W. PAGE KEETON ET AL., PROSSER AND KEETON ON THE LAW OF TORTS 164–65 (5th ed. 1984).

[2] For example, in the hypothetical presented above, the physician's decision to prescribe penicillin-based antibiotics despite the patient's allergy records would be evaluated in comparison to the decisions expected of reasonable physicians in her same field. *See* RESTATEMENT (SECOND) OF TORTS § 299A (AM. L. INST. 1965) ("Unless he represents that he has greater or less skill or knowledge, one who undertakes to render services in the practice of a profession or trade is required to exercise the skill and knowledge normally possessed by members of that profession or trade in good standing in similar communities."); Michael A. Froomkin et al., *When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning*, 61 ARIZ. L. REV. 33, 54–58 (2019).

[3] *See infra* Section I.B.ii for further discussion.

[4] *See infra* notes 29–34 and accompanying text.

[5] *See* Alberto Galasso & Hong Luo, *Tort Reform and Innovation*, 60 J.L. & ECON. 385, 386 (2017); Jennifer L. Phillips, *Information Liability: The Possible Chilling Effect of Tort Claims Against Producers of Geographic Information Systems Data*, 26 FLA. ST. U. L. REV. 743 (1999) (considering chilling effects on producers of GIS data); David A. Anderson, *First Amendment Limitations on Tort Law*, 69 BROOK. L. REV. 755 (2004) (describing general hesitation to enforce tort law where such enforcement would chill free speech); Laurel Witt, *Preventing the Rogue Bot Journalist: Protection from Non-Human Defamation*, 15 COLO. TECH. L.J. 517, 532 (2017) ("Congress intended Section 230 [of the Communications Decency Act] to prevent a chilling effect for platforms that could otherwise be liable for third-party information.").

every detail of lengthy medical records, and it would additionally be concerned with creating an undesirable discouraging effect on the physician profession due to fear of legal exposure.[6] The reasonableness analysis allows courts to strike tailor-made balances between these considerations, based on the current circumstances and objectives they wish to promote.[7]

The potential benefits and dangers of some newly emerging technologies warrant a very careful balance between the aforementioned considerations.[8] One particularly important question for torts is whether the reasonableness analysis ought to apply when the tortfeasor is not a person, but rather a "self-learning," "autonomous," or "artificially intelligent" system.[9] This Article will refer to such systems as "thinking algorithms" because their algorithms model various stages of human thinking and do so independently of a human.[10]

[6] Jeffrey O'Connell & Andrew S. Boutros, *Treating Medical Malpractice Claims Under a Variant of the Business Judgment Rule*, 77 NOTRE DAME L. REV. 373, 396–97 (2002) ("It is often asserted that fear of liability has also caused many physicians . . . to leave high-risk geographic areas, to abandon their particular area of medical practice in exchange for less litigious fields of medicine, or to even retire early. Moreover, fear of liability may also have an impact on the rate of students willing to matriculate in medical schools.").

[7] If our meningitis example were to occur during a crisis in the healthcare system, for example, courts may determine that it was reasonable of the physician not to look at the patient's medical record. They could also fine-tune said conclusion based on relevant parameters such as the number of other patients who were awaiting treatment, on the urgency of the patient's medical condition, or on the length of the medical record that was ignored. If, on the other hand, physicians' attention was not a scarce resource, then the court would likely set higher standards of reasonableness, favoring risk-avoidance and compensating the victim considerations over ones of efficiency and avoiding a chilling effect (thus, potentially, holding that physicians must always read the entire medical record regardless of the circumstances, or setting more strict standards of when it is reasonable not to do so in the case of overcrowded healthcare systems).

[8] *See generally* T. W. Small et al., *Designing an Accessible, Technology-Driven Justice System: An Exercise in Testing the Access to Justice Technology Bill of Rights*, 79 WASH. L. REV. 223, 250 (2004) (theorizing that designing a new system of laws would include attempts to optimize benefits and minimize risks of technology); Steven D. Seybold, *Somebody's Watching Me: Civilian Oversight of Data-Collection Technologies*, 93 TEX. L. REV. 1029, 1059 (2015) (discussing the importance of maintaining a proper balance as technology "continue[s] to evolve and present new dangers to the rights of citizens"); Shlomit Yanisky-Ravid & Sean K. Hallisey, *"Equality and Privacy by Design": A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes*, 46 FORDHAM URB. L.J. 428, 474–75 (2019) (discussing the need for "balanc[ing] the need for AI innovation against the dangers of discrimination").

[9] The "buzzwords" used to describe the type of systems we view as "smart" have various definitions and may have different embodiments. This paper is generally indifferent to the existence (or lack thereof) of any physical embodiment of the "smart" system. See, in that context, Professor Jack Balkin's suggestion that both algorithms and robots are similar members of the "Algorithmic Society" and might be treated alike. Jack M. Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1226 (2017).

[10] For a discussion of the difference between algorithmic and human decision-making, see *infra* Part II.

Technology is ever advancing, and thinking algorithms' self-learning abilities allow them to reach conclusions based on databases of previous cases.[11] This in turn enables humans to entrust machines to make complex decisions that until recently required human discretion, and even to replace professional human judgment in matters of expertise where there is no clear right or wrong answer. In the field of medicine, physicians increasingly rely on algorithms to diagnose medical conditions and select optimal treatments.[12] In the field of law, virtual attorneys are utilized by law firms to conduct legal research,[13] algorithmic online dispute resolution mechanisms solve disputes online,[14] and bail algorithms determine whether defendants awaiting trial may post bail to be released.[15] In everyday life, human drivers in various countries now share the road with autonomous vehicles,[16] while algorithms are used interchangeably with human professionals to provide tax advice, serve as company directors, and even offer religious services.[17]

Assume then that our physician prescribing penicillin is no longer a flesh-and-blood practitioner but is instead a sophisticated medical thinking

[11] For further discussion on how "self-learning" works, see *infra* Part II.

[12] *See, e.g.,* Vinod Khosla, *Technology Will Replace 80% of What Doctors Do*, FORTUNE (Dec. 4, 2012, 2:26 PM), http://fortune.com/2012/12/04/technology-will-replace-80-of-what-doctors-do; Alina Shrourou, *Deep Learning in Healthcare: A Move Towards Algorithmic Doctors*, NEWS MED. (Mar. 15, 2017), https://www.news-medical.net/news/20170315/Deep-learning-in-healthcare-a-move-towards-algorithmic-doctors.aspx.

[13] *See, e.g.*, John Mannes, *ROSS Intelligence Lands $8.7M Series A to Speed up Legal Research with AI*, TECHCRUNCH (Oct. 11, 2017, 2:34 PM), https://techcrunch.com/2017/10/11/ross-intelligence-lands-8-7m-series-a-to-speed-up-legal-research-with-ai; Anthony Sills, *ROSS and Watson Tackle the Law*, IBM: WATSON BLOG (Jan. 14, 2016), https://www.ibm.com/blogs/watson/2016/01/ross-and-watson-tackle-the-law.

[14] Michael Legg, *The Future of Dispute Resolution: Online ADR and Online Courts*, 27 AUSTRALASIAN DISP. RESOL. J. 227 (2016).

[15] A.J. Wang, Procedural Justice and Risk-Assessment Algorithms 1 (June 21, 2018) (unpublished manuscript) (https://ssrn.com/abstract=3170136); *cf.* Tom Simonite, *How to Upgrade Judges with Machine Learning,* MIT TECH. REV. (Mar. 6, 2017), https://www.technologyreview.com/s/603763/how-to-upgrade-judges-with-machine-learning (discussing how algorithms may assist judges to predict which defendants will fail to show to court).

[16] For the current state of autonomous vehicle development and deployment, see, e.g., Sean Bollman, *Autonomous Vehicles: A Future Fast Approaching with No One Behind the Wheel*, 20 PITT. J. TECH. L. & POL'Y 1, 1–4 (2020); Laura J. Grabouski, *On the Road: Driverless Cars Are Disrupting Norms and Legal Standards*, 82 TEX. BAR J. 232 (2019); Adam Kaslikowsi, *Everything You Need to Know About Autonomous Vehicles*, DIGIT. TRENDS (June 30, 2019), https://www.digitaltrends.com/cars/the-current-state-of-autonomous-vehicles.

[17] Richard Susskind & Daniel Susskind, *Technology Will Replace Many Doctors, Lawyers, and Other Professionals*, HARV. BUS. REV. (Oct. 11, 2016), https://hbr.org/2016/10/robots-will-replace-doctors-lawyers-and-other-professionals; Sasha A.Q. Scott, *Algorithmic Absolution: The Case of Catholic Confessional Apps*, 11 ONLINE - HEIDELBERG J. RELIGIONS ON INTERNET 254 (2016). For a forecast on the percentage of actions currently performed by human professionals that could be replaced by automation, see *Automation Potential and Wages for US Jobs*, MCKINSEY GLOB. INST. (Oct. 1, 2018), https://public.tableau.com/profile/mckinsey.analytics#!/vizhome/AutomationandUSjobs/Technicalpotentialforautomation.

algorithm. Should the reasonableness analysis be applied to the medical thinking algorithm, just as it would be in the case of a human doctor? Although case law has left the question largely unaddressed,[18] academia has to some extent discussed the general concept of applying a reasonableness standard to algorithms and robots. Setting the obvious concern of who would pay for any damages caused by algorithms aside,[19] academics have pointed to several advantages of subjecting algorithms to the reasonableness analysis. Law and health professor Ryan Abbott, for example, has argued that holding computerized tortfeasors to a negligence standard would accelerate the adoption of safer technology and thus result in fewer accidents.[20] I have previously argued that applying a reasonableness standard to algorithms would result in more equality among victims by minimizing anomalies that would otherwise occur under a scheme where human and algorithmic tortfeasors who caused identical harm were subject to completely different tort frameworks.[21]

Additionally, the growing improvement of algorithmic performance vis-à-vis humans could lead to a reality in which the general standard of care sufficient to avoid liability would no longer be that expected of a person. Rather, an elevated level of care—one that represents the generally higher level of care that algorithms can provide—might become the prevailing

[18] A very similar question might have been addressed in the case of *Nilsson v. General Motors LLC*, where a motorcyclist hit by an autonomous vehicle alleged that the vehicle itself drove in a negligent manner. The case, brought before the U.S. District Court for the Northern District of California, was settled before trial, leaving the question of algorithmic negligence, and thus of algorithmic reasonableness, undiscussed. Complaint for Damages, Nilsson v. General Motors LLC, No. 4:18-cv-00471-KAW (N.D. Cal. Jan. 22, 2018).

[19] Rather, applying a reasonableness analysis would be a way of determining liability, which would then be undertaken by a legal or natural person, similar to how employees' reasonableness is evaluated in tort cases where employers are sued, even though it is the employer who would pay damages if the worker was negligent. For further discussion, see *infra* Part III.

[20] Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1, 22 (2018).

[21] Karni Chagal-Feferkorn, *The Reasonable Algorithm*, 1 U. ILL. J.L. TECH. & POL'Y 111, 116–17 (2018).

The concept of algorithmic reasonableness was also discussed by David Vladeck who referred to an elevated standard of care to be required of autonomous vehicles, by Jeffrey Gurney who suggested imputing a "reasonableness" standard to driverless vehicles, and by Justice Curtis E.A. Karnow who referred to the "reasonableness" of the actions of a robot as a means of dealing with the fact that its actions are unexpected and not fully pre-programmed by its manufacturers. David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, (2014); Jeffrey K. Gurney, *Imputing Driverhood: Applying a Reasonable Driver Standard to Accidents Caused by Autonomous Vehicles*, *in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 51 (Patrick Lin et al. eds., 2017); Curtis E.A. Karnow, *The Application of Traditional Tort Theory to Embodied Machine Intelligence*, *in* ROBOT LAW 51 (Ryan Calo et al. ed., 2016); Curtis E.A. Karnow, *The Opinion of Machines*, 19 COLUM. SCI. & TECH. L. REV. 137 (2017); *see also* Stephanie Pywell, *The Reasonable Robot*, 7700 NEW L.J. 19 (2016).

standard for both humans and algorithms.[22] Under such a scenario, determining whether an algorithmic standard of care has been met, such a standard must first be developed.

In that context and given the Sisyphean quest for the appropriate legal framework to be applied to damages caused by thinking algorithms,[23] a reasonableness analysis is clearly an option worth considering. However, two crucial questions have not yet been addressed: whether the reasonableness standard is compatible with the unique traits of thinking algorithms, and, if yes, what would analysis of algorithmic reasonableness look like? Other than a general, perhaps somewhat intuitive, notion that the standard of care expected from a thinking algorithm would be elevated compared to that required of a person,[24] existing literature has not yet delved into the practical question of *how* to assess whether an algorithm acted reasonably or not. As one scholar notes, "[w]e have a generally workable view of what it means for a person to act negligently or otherwise act in a legally culpable manner,

[22] Froomkin et al., *supra* note 2, at 50, 60–61 ("once ML diagnostics are statistically superior to humans, it will only be a short while before legal systems, including in the United States, treat machine diagnosis as the 'standard of care.' . . . Thus, a physician, hospital, or insurer relying on an ML diagnosis will, at least initially, be held to no higher standard than that of the ordinary physician. Once ML itself becomes the standard of care, ML will raise the bar."); *see also* Abbott, *supra* note 20; Gurney, *supra* note 21 ("[G]iven that autonomous vehicles do not suffer from human frailties, and given that these vehicles will supposedly have the ability to detect objects better than humans, one would expect that (eventually) autonomous vehicles will be held to a higher standard than human drivers.")

[23] "Precisely how revolutionary robot-driven accidents will be for our legal system is less clear . . . although opinions vary, a tentative consensus has emerged on at least one front. For most, conventional tort liability theories are essentially nonstarters." Bryan Casey, *Robot Ipsa Loquitur*, 108 GEO. L.J. 225, 230 (2019); *see also* Marta Infantino & Weiwei Wang, *Algorithmic Torts: A Prospective Comparative Overview*, 28 TRANSNAT'L L. & CONTEMP. PROBS. 309, 313–14 (2019) (stating that "[o]ne does not need a crystal ball to predict that, as algorithm-related activities multiply around us, so will accidents . . . [that] will cause people to suffer some kind of loss," and noting that tort law must develop to determine remedies to such loss); Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017) ("The rise of increasingly complex algorithms calls for critical thought about how best to prevent, deter, and compensate for the harms that they cause."); Alexander B. Lemann, *Autonomous Vehicles, Technological Progress, and the Scope Problem in Products Liability*, 12 J. TORT L. 157 (2019) ("How the legal system should attribute responsibility for the (hopefully few) crashes autonomous vehicles cause is an open and hotly debated question.").

[24] *See* Vladeck, *supra* note 21, at 131 (The court in *Arnold v. Reuther* ruled that a driver was not liable for hitting a pedestrian because he was only human and not a robot: "A human being, no matter how efficient, is not a mechanical robot and does not possess the ability of a radar machine to discover danger before it becomes manifest. Some allowances, however slight, must be made for human frailties and for reaction, and if any allowance whatever is made for the fact that a human being must require a fraction of a second for reaction and cannot respond with the mechanical speed and accuracy such as is found in modern mechanical devices, it must be realized that there was nothing that Reuther, a human being, could have done to have avoided the unfortunate result . . . .").

but we have no similarly well-defined conception of what it means for an algorithm to do so."[25]

What *specific criteria* ought to shape the "reasonable algorithm" analysis? The factors we use must allow the assessment itself to be practical while simultaneously both fulfilling tort law's goals of promoting safety and compensating the injured party and balancing the need to avoid a chilling effect on the development and usage of innovative technologies. Developing a reasonableness standard suited to thinking algorithms is therefore no easy task.

From a practical perspective, reasonableness is a reference point, measured against other alternative courses of action or decisions that could have been reached. When we assess the reasonableness of humans, we look to how similar persons or professionals might have behaved in similar situations, and this assessment is often aided by expert testimony.[26] In the case of thinking algorithms, however, there is no cohort of equivalent systems with which to compare. Experts cannot necessarily opine on what a different algorithm might have done because there is often no sufficient volume of other examples to point to. And we cannot solve this practical challenge by assessing the reasonableness of algorithms based on how *humans* would have acted, as doing so would preserve a static standard and would ignore the tremendous scope of technology to continuously improve and achieve better and safer results than those of humans.

Alternatively, we may presume prima facie that algorithms are absolutely superior to their human counterparts and set the bar at some elevated point beyond that of human reasonableness.[27] However, doing so would still be using human competence as a baseline, resulting in demanding *too little* of the algorithms whose capabilities in certain fields might be beyond human abilities or comprehension. At the same time, such a solution might also pose the threat of demanding *too much* of the algorithm, as it assumes their general superiority over humans, ignoring the fact that, even when algorithms deliver better results than humans on average, their decision-

[25] Tutt, *supra* note 23, at 105; *see also* Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. Rev. 1315, 1318 (2020) ("With most new technologies, we gain familiarity over time, eventually creating a sense of what constitutes reasonable care or a collective intuition on which negligence law can rely as it adapts. AI poses challenges for negligence law that may delay the common law's ability to adapt or even prevent adaptation outright.").

[26] Alan D. Miller & Ronen Perry, *The Reasonable Person*, 87 N.Y.U. L. REV. 323, 370 (2012). Said approach for assessing reasonableness is the *positive* one. For a discussion differentiating between the positive and normative approaches for determining reasonableness, see *infra* notes 35–46 and accompanying text.

[27] For instance, by granting a safe harbour to systems whose general performance is better than that of humans. *See* Mark M. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311, 1384 (2019). For further discussion of said proposal, see *infra* notes 161–62 and accompanying text.

making process is still characterized by certain weaknesses compared to that of a person.[28]

This Article analyzes aspects in which the reasonableness assessment is compatible with the unique qualities of thinking algorithms and examines which of their traits require creative adjustments of the analysis. The Article's main contribution lies in proposing a concrete "reasonable algorithm" standard that could be practically applied by decision-makers. This is done in a manner that accounts for the differences between humans and algorithms, and applies the reasonableness standard to algorithms in a way that promotes safety and victims' compensation while avoiding a dampening effect on the development and usage of new, beneficial technologies.

Part I reviews the concept of "the reasonable person" to provide contextual background to my proposed model of determining algorithmic reasonableness, while Part II focuses on thinking algorithms and analyzes how their decision-making process is different (for better or worse) from that of humans. Part III discusses the concept of applying a reasonableness standard to thinking algorithms, including its advantages and challenges, and Part IV proposes a concrete model for analyzing an algorithm's reasonableness in a manner that overcomes the challenges previously discussed.

Part I: The Reasonable Person

A. *Rationales, History, and Manner of Assessment*

Imposing legal liability on the wrongdoer for damages she caused advances two core goals. First, payment of damages by the wrongdoer assures that the injured party is compensated so that the wrong is theoretically "corrected," and principles of justice and fairness are met.[29] Second, imposing liability also serves as a deterrent—it encourages behavior modifications to increase overall safety. Potential tortfeasors faced with the risk of being found liable for damages will, the tort framework assumes, be incentivized to take measures to minimize the risk of causing damage.[30] Torts that do not require the demonstration of fault on the part of the tortfeasor might be considered superior in meeting these aims, such as product liability torts that impose liability only upon the existence of a defect, regardless of any wrongdoing by the manufacturer.[31]

[28] See *infra* Part II.

[29] *See generally* ERNEST WEINRIB, THE IDEA OF PRIVATE LAW (1995).

[30] *See, e.g.*, John C.P. Goldberg, *Twentieth-Century Tort Theory*, 91 GEO. L.J. 513 (2003).

[31] Such is the case of defects found in the manufacturing of a product, which expose (under the laws of most states) the manufactures and sellers to strict liability, which does not require proof of any negligence. *See* RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2(a) (AM. L. INST. 1998) ("A product . . . contains a manufacturing defect when the product

Under the negligence theory, however, the fulfillment of tort law's goals is more limited, as liability depends on the "unreasonableness" of the tortfeasor. To establish negligence, the following four elements must be proven: a duty of care exists; there has been a breach of that duty, either by an overt act or omission; the non-breaching party has suffered damage; and there is a causal link between the breach of the duty and the damage suffered.[32]

The reasonableness analysis comes into play when evaluating the breach element. A person breaches her duty of care if she does not adhere to the standard of reasonable care when carrying out actions that might foreseeably harm others.[33] "Reasonableness," therefore, limits the cases in which the injured party is entitled to compensation, and also limits the extent to which potential wrongdoers must take measures to protect against damage. This standard can prevent over-deterrence, and within the context of development and usage of innovative means, utilizing the appropriate reasonableness standard may result in encouraging technological improvements.[34] The precise balance between satisfying the objectives of tort law and avoiding a dampening effect on innovation heavily depends on the term "reasonableness" and how it is interpreted. Who is the "reasonable person" that serves as a reference point against which the behavior of wrongdoers is measured?

Since its origin, "the reasonable person" has been a vague term, open to various interpretations and methods of assessment.[35] Tracing back to the eighteenth century, English criminal liability cases used the standard of a "person of an ordinary capacity" to determine defendants' liability in fraud cases. If a person of ordinary capacity would not have been defrauded under similar circumstances, then the plaintiff was considered a "fool" rather than a victim of a criminal act of fraud.[36] According to scholar William Hawkins, when "common prudence and caution" were sufficient to protect one from being tricked or defrauded, the underlying false pretense would not be crim-

departs from its intended design even though all possible care was exercised in the preparation and marketing of the product"); JOHN C. P. GOLDBERG & BENJAMIN C. ZIPURSKY, THE OXFORD INTRODUCTIONS TO U.S. LAW: TORTS 284–88 (2010).

[32] *See supra* note 1.

[33] Benjamin C. Zipursky, *Foreseeability in Breach, Duty, and Proximate Cause*, 55 WAKE FOREST L. REV. 1247, 1249–50 (2009).

[34] For a discussion on the balance between encouraging technological advancements and assuring their safety, see *infra* notes 58–72 and accompanying text.

[35] *See infra* notes 36–50 and accompanying text.

[36] One of the early decisions adopting the "person of ordinary capacity" standard was *R. v. Jones*, where the court refused to convict the defendant who presented himself as a collector of debt owed by plaintiff and disappeared after collecting the money. The court held that "the deceit would be criminal only if it were 'such a Cheat as a Person of an ordinary Capacity can't discover.'" Simon Stern, *R v Jones (1703)*, *in* LANDMARK CASES IN CRIMINAL LAW 59, 60 (Ian Williams et al. eds., 2016).

inally punishable.[37] The use of "common" and "ordinary" found its way into American jurisprudence as well, where a standard of "ordinary caution" was adopted in criminal and civil cases.[38] Subsequent decisions used the standard of a person of "ordinary fitness" until 1856, when an English court introduced the familiar standard of reasonableness, holding that negligence is an act or omission from which the "prudent and reasonable" man would refrain.[39]

These origins of reasonableness seem to point to a normative approach: one that seeks to shape the behavior of members of society using legal means. In the context of criminal fraud and other historic cases, courts sought to encourage the public to take everyday measures such as using common sense against fraud. Later, this normative approach—which focused on economic efficiency and deterrence—was further developed by Judge Learned Hand and his famous "Hand Formula" for determining whether a breach has occurred. Under the Hand Formula, meeting the standard of care requires the adoption of relevant precautions, so long as the cost of doing so is not higher than the expected value of the gain in safety.[40] Therefore, under a normative economic efficiency analysis, reasonableness is determined based on the relative cost of safety measures and the expected improvement of safety achieved as a result.[41] To return to the example of a physician administrating penicillin without reading the patient's medical record, a court applying the Hand Formula would compare the cost of me-

[37] 1 William Hawkins, A Treatise of the Pleas of the Crown: Or a System of the Principal Matters Relating to That Subject, Digested Under Their Proper Heads 188 (Eliz. Nutt 1716).

[38] *People v. Conger*, for example, held that if a person of ordinary caution would not be deceived, then the argued act of deception is not a criminal act. The ruling also clarified that the assessment of ordinary caution may differ from case to case and "may depend on a thousand circumstances to be considered on trial." Stern, *supra* note 36, at 73.

[39] Blyth v. Co. of Proprietors of the Birmingham Waterworks (1856) 156 Eng. Rep. 1047, 1049.

[40] United States v. Carroll Towing Co., 159 F.2d 169 (2d Cir. 1947); Stephen G. Gilles, *United States v. Carroll Towing Co.: The Hand Formula's Home Port*, *in* Torts Stories 11 (Robert L. Rabin & Stephen D. Sugarman eds., 2003).

[41] The Hand Formula provides that when the damage expectancy (the probability of an accident or a damaging act multiplied by the gravity of damage) is higher than the cost of the safety measures required to prevent or minimise the damage, then the tortfeasor was not reasonable. *Carroll Towing Co.*, 159 F.2d at 173. Another primary normative consideration discussed in the context of the reasonable person is that of Kantian morals, focusing on the balance between one's freedom to act versus one's freedom not to be constrained by another's (damaging) choice. Under a Kantian perspective of reasonableness, imposing liability on a person only to deter others from similar conduct, for instance from not taking cost-effective precautionary measures, is ethically problematic, as it would be treating a person as a means rather than as an end in itself. Miller & Perry, *supra* note 26, at 328. While it is interesting to discuss this concern in the context of robots and algorithms (presumably, treating them as means would not raise Kantian resistance), in its normative considerations this paper focuses on those concerns associated with deterrence and economic efficiency.

ticulously reading the patient's medical record (likely translated into the precious time invested by the physician in said task rather than in examining other patients) with the likelihood of missing an important factor mentioned in the record, multiplied by the resulting damage such a scenario would entail. For instance, if the court quantifies the cost in additional time required by the physician to read the medical record in detail at $1,000, the likelihood of missing important information at 1%, and the potential resulting damage at $50,000, then the physician's decision to treat the patient without thoroughly reading her medical record first would be deemed reasonable under the Hand Formula (since $1,000 is greater than $500). As further explained later in this Part, courts often factor in additional considerations when quantifying both the cost of precautions and expected damages, which allows for greater flexibility in shaping desired behaviors by potential wrongdoers.

An alternative method for assessing the reasonableness of persons is the positive approach. Unlike the normative approach, which is designed to promote certain interests or values, the positive approach is more qualitative and assesses the behavior of the wrongdoer in comparison to how others, similarly situated, would have acted. In the past, such comparison to what the "reasonable person" would have done relied on a hypothetical "average man," whose attributes were conceived based on statistical empirical observations. Measurable variables including height, weight, and propensity for criminal behavior were taken into account to create a model of the "average man," whose theoretical behavior would constitute the reference point for reasonableness.[42] Pragmatically, positive reasonableness may have been based on a less statistical and more open-ended set of perceptions about the common nature of the public, based on the rationale that "if we know the individuals in society, we know the reasonable person."[43]

In practice, courts use both normative and positive assessments of reasonableness when determining the tort liability of human beings. In medical malpractice cases, for example, courts tend to focus on positive assessments of what other physicians in a given field would have done.[44] However, they are not always satisfied with only evaluating whether the wrongdoer acted as others would have. Courts also have discretion to weigh normative economic-efficiency considerations in negligence cases,[45] and in some instances

[42] Miller & Perry, *supra* note 26, at 370–71; STEPHEN M. STIGLER, THE HISTORY OF STATISTICS: THE MEASUREMENT OF UNCERTAINTY BEFORE 1900 201 (1986).

[43] Miller & Perry, *supra* note 26*,* at 377.

[44] For many years, the prevailing test for physicians' reasonableness has been whether they followed medical custom. Froomkin et al., *supra* note 2, at 51. As discussed below, though, in recent years several states have adopted a different approach.

[45] RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 3 (AM. L. INST. 2010) ("A person acts negligently if the person does not exercise reasonable care under all the circumstances. Primary factors to consider in ascertaining whether

have determined acts to be unreasonable even after acknowledging that comparable persons would have taken the same actions.[46]

Another dimension of the reasonableness analysis is that it accounts for subjectivity. The key question here is to what extent should a court take the individual attributes into consideration when determining whether the wrongdoer acted reasonably or not. Allowing a fully subjective standard of reasonableness would be meaningless; holding that a physician was not negligent only because, subjectively, she was tired or confused when making a damaging decision would not leave any room for assessing reasonableness on the basis of the underlying facts and circumstances. On the other hand, a fully objective standard would ignore crucial information in the case, such as physical constraints of the wrongdoer.[47]

In practice, reasonableness is assessed primarily objectively, allowing some subjectivity regarding the specific situation and state of knowledge the individual tortfeasor possessed at the time.[48] In some instances, however, the courts do develop individualized groups of wrongdoers whose reasonableness is then compared to the other members of the group, rather than to the general reasonable person. Children, for example, are not held to the reasonableness standards expected of adults but rather to those of other children,[49] and laypersons are not expected to employ the same level of care of professionals.[50]

The assessment of the reasonable person in the tort context is therefore an open-ended exercise that allows courts to weigh a variety of considerations on a case-by-case basis in order to determine whether a wrongdoer's act or omission was reasonable.

the person's conduct lacks reasonable care are the foreseeable likelihood that the person's conduct will result in harm, the foreseeable severity of any harm that may ensue, and the burden of precautions to eliminate or reduce the risk of harm.")

[46] *See, e.g.*, Helling v. Carey, 519 P.2d 981 (Wash. 1974) (where the court disregarded the standards of the ophthalmology profession which required no glaucoma tests, given its very low cost); *see also* Froomkin et al., *supra* note 2, at 47, 55–58 (describing a trend of abandoning the deference to medical custom in the case of physicians and holding physicians to a "reasonable physician" standard of care similar to the one applied in the case of other wrongdoers, in other words one that weighs normative considerations in addition to whether current medical practices have been followed).

[47] For a proposal of applying subjective, personalised negligence law, see Omri Ben-Shahar & Ariel Porat, *Personalizing Negligence Law*, 91 N.Y.U. L. REV. 627 (2016). *See also* Victoria Nourse, *After the Reasonable Man: Getting over the Subjectivity Objectivity Question*, 11 NEW CRIM. L. REV. 33, 33–50 (2008).

[48] Miller & Perry, *supra* note 26, at 378–79.

[49] *See* DAN B. DOBBS, THE LAW OF TORTS 293 (2001).

[50] *See* Hous. Auth. of Carrollton v. Ayers, 88 S.E.2d 368, 372 (Ga. 1955) ("The law imposes upon persons performing architectural, engineering, and other professional and skilled services the obligation to exercise a reasonable degree of care, skill, and ability, which generally is taken and considered to be such a degree of care and skill as, under similar conditions and like surrounding circumstances, is ordinarily employed by their respective professions.").

B. *The Application of the Reasonableness Assessment through the Eyes of the Court*

From a judicial point of view, the use of a reasonableness analysis to determine liability has different advantages and disadvantages compared to other legal methods for deciding tort cases (manifested both in the procedural and material aspects of the judicial process). This Part will briefly review certain aspects of the reasonableness analysis that are most relevant to the subsequent discussion in Part III of how reasonableness is affected when the tortfeasor is algorithmic and their effect on the legal procedure and outcome.

i. Allowance for Flexibility in Promoting Desired Goals

Reasonableness is a broad and vague standard rather than a concrete rule.[51] As a result, courts evaluating reasonableness have the flexibility to decide which interests and social goals to promote, even when these goals are sometimes contradictory. In the meningitis hypothetical, a court that is concerned about a shortage of physicians in the healthcare system could decide to emphasize efficiency considerations over safety ones. As a result, that court may determine that a physician's choice to rely on the information collected while taking the patient's history, instead of spending time reviewing the medical record in detail, was reasonable. By contrast, if there were no shortage of doctors in the healthcare system, a court might instead hold the opposite. The reasonableness standard allows for different rulings given different situations and concerns—not making a blanket statement that physicians are either always or never obliged to read medical records. Rather, a court may amplify or downplay certain considerations by shifting the exact balance point where reasonableness is set. For example, it could take into account the length of the medical record or the number of other patients awaiting the physician's attention and tweak the specific parameters it uses to establish reasonable physician behavior. A court may determine that ignoring the medical record is reasonable if the record is more than five pages long, or it may decide this behavior is only permissible if the record is over one hundred pages. It could also rule that ignoring the record is only deemed reasonable if a certain ratio between the length of the document and the number of patients awaiting treatment is exceeded, or impose liability based on the level of medical urgency of the case at hand. Each determination balances both the need to improve patient safety and the need for efficiency, albeit the exact balance of criteria differs in each case.

In the absence of a well-defined classification of which behavior is reasonable and which is not, potential wrongdoers may also constantly fine-

[51] Miller & Perry, *supra* note 26, at 325 ("The reasonable person is a legal concept that can be imbued with different content.").

tune their behavior and generally act in ways that take into account the various goals that courts are known to promote. Thus, even if operating under circumstances that have not yet been adjudicated, potential wrongdoers can have a general sense of what would likely be deemed reasonable and what would not.

As we shall see in Part III.B discussing the algorithmic context, the flexibility of the reasonableness analysis presents both advantages and challenges. On the positive side, flexibility by the courts to change the relative weights given to competing considerations such as safety promotion and avoiding a chilling effect on technology might be especially important when new, quickly evolving technologies are concerned. On the more problematic side is the fact that, unlike humans, thinking algorithms will likely find it almost impossible to fine-tune their behavior based on the broad and ever-changing concept of reasonableness expected of them.[52]

ii. Adaptable to the Dynamic Nature of Wrongdoers

Both normative and positive approaches to assessing reasonableness may be applied as is, regardless of the type of tortfeasor or the decision-making tools at their disposal. A normative analysis of reasonableness is based on damage-prevention utility calculations or on other considerations the court wishes to promote. Any classifications pertaining to the tortfeasor (such as education level, gender, political opinion, artistic taste, or zodiac sign) would generally not affect the normative analysis of reasonableness. Rather, the court would apply a relatively objective standard and would make its reasonableness assessment based on the social goals it wishes to promote. Granted, in some special and well-defined instances, the court *would* take into account the fact that the tortfeasor belongs to a certain group. In the case of children and professionals, for example, the reasonableness assessment is based on what is expected of the reasonableness of peers, rather than the general notion of the reasonable *person*.[53]

None of this, however, changes the basic concept of the normative assessment. Whether the court decides to compare the wrongdoer to their own distinct group or to the general reasonable person, the mechanism of balancing between competing social goals (be it economic efficiency considerations or others) could be equally applied regardless of the identity of the tortfeasor. If we take our meningitis example, then the fact that the wrongdoer is a physician, rather than an administrator in charge of reviewing medical records, might affect the normative calculus but not the normative mechanism itself. Presumably, the cost of reading patients' lengthy medical records is higher in the case of the doctor, whose time at the hospital is like-

[52] *See infra* notes 154–55 and accompanying text.

[53] *See supra* notes 49–50 and accompanying text.

ly much more costly than that of non-medical personnel. The differences in wrongdoers may therefore lead to opposite outcomes of reasonableness. If, for example, the damage expectancy of not reading the medical record is calculated at $200 and the value of time required by an administrative staff member to read the medical file is quantified at $150, then ignoring the file would be deemed unreasonable of the administrative staff member. If the wrongdoer is a physician whose time reading the file is valued at $500, on the other hand, the calculation would result in a finding of reasonableness. The outcome of the normative analysis may, of course, be dependent on the type of wrongdoer or the skills and abilities they possess. The essence of the normative analysis—whether conducting a Hand Formula analysis or applying other balances between desired social goals—is indifferent to the identity of players to which the assessment is applied. The same equation would be used both in the case of the physician and in the case of the administrative staff member, but only the numerical values would be changed.

The same is true for a positive analysis of reasonableness. A positive analysis would take into account the exact type of wrongdoer who caused damage. In the field of medicine, where positive analyses are common,[54] the reasonableness of physicians would be determined based on the standard of care adhered to by other physicians.[55] However, the basis for making the analysis—of using similarly situated wrongdoers as reference points—would be indifferent to the nature and type of the wrongdoer. In other words, the specific traits or capabilities of the wrongdoer would be used by the court, but only when deciding who to compare the wrongdoer to.

As will be further discussed in Part III.B, both normative and positive measurements of reasonableness could be easily applied to thinking algorithms, regardless of how different their decision-making is compared to that of humans and their ever-evolving capabilities.

iii. Understanding the Exact Reason for Reaching the Damaging Decision is Unnecessary

Although having more information on the damaging decision and how it was reached can assist in determining reasonableness, the assessment can be made without a precise understanding of the reasons behind the decision. To take an example of a driver hitting a pedestrian with his car, knowing that the driver swerved in an attempt to avoid hitting a different pedestrian

[54] Froomkin et al., *supra* note 2, at 51–58.

[55] *Id.* at 54 ("The standard of care is that established by the 'relevant community', which is now understood to be the national group of practitioners in that specialty."); *see also* Philip G. Peters, *The Quiet Demise of Deference to Custom: Malpractice Law at the Millennium*, 57 WASH. & LEE L. REV. 163, 180 (2000); John W. Ely et al., *Determining the Standard of Care in Medical Malpractice: The Physician's Perspective*, 37 WAKE FOREST L. REV. 861, 862 (2002).

would be relevant to the analysis. However, a court could still determine whether the driver's conduct was reasonable even if the driver couldn't provide any information about how or why he hit the pedestrian.[56]

In the medical malpractice context, understanding how or why a physician misdiagnosed a certain medical condition would be relevant in assessing the reasonableness of the misdiagnosis. If the doctor's decision was based on erroneous information given by the patient, this could lead to a finding of reasonableness, so long as it was reasonable of the physician to rely on such information without further investigation. If instead the physician confused one medical condition with another for no apparent reason,[57] the court can still determine whether the physician's action was reasonable. Rather than focusing on the source of the mistake, the court could use a positive analysis to determine whether the physician was in line with protocols and customs, or alternatively it could use a normative economic analysis and assume the values of the parameters. In other words, another dimension that renders the reasonableness analysis adaptive, flexible, and easy to apply is its capacity to be used even in the absence of a profound understanding of why exactly the wrongdoer acted as they did (a trait which characterizes self-learning algorithms).

Part II: Thinking Algorithms and their Decision-Making Process

A. *Background*

Automated systems have been utilized by humankind for centuries.[58] In addition to assisting humans with physical tasks, machines have long been used to assist or replace humans in processing data. The electronic calculator enables engineers and other professionals to provide faster, more accurate outputs,[59] autopilot controls in airplanes improve flight safety through an automated system capable of processing huge amounts of information in split seconds,[60] and cruise control and auto-parking devices assist everyday

[56] Driver v. Brooks, 10 S.E.2d 887, 892 (Va. 1940) (Virginia Supreme Court affirmed jury finding that defendant driver was liable for accident that he did not remember, stating "[l]iability for acts constituting negligence is not removed by a mere statement of the tortfeasor that he does not remember the circumstances.").

[57] For a discussion of human inherent weaknesses, see *infra* notes 93–99 and accompanying text.

[58] *See* IBN AL-RAZZAZ AL-JAZARI, THE BOOK OF KNOWLEDGE OF INGENIOUS MECHANICAL DEVICES (Donald R. Hill trans., Pakistan Hijra Council 1989) (1974).

[59] *See Electronic Calculators—Handheld*, NAT'L MUSEUM AM. HISTORY, http://americanhistory.si.edu/collections/object-groups/handheld-electronic-calculators (last visited Mar. 15, 2021); Nick Valentine, *The History of the Calculator*, CALCULATOR SITE, https://www.thecalculatorsite.com/articles/units/history-of-the-calculator.php (Mar. 11, 2019).

[60] Kyle Colonna, *Autonomous Cars and Tort Liability*, 4 CASE W. RES. J.L. TECH. & INTERNET 81, 93–97 (2012); M. C. Elish & Tim Hwang, *Praise the Machine! Punish the Hu-*

drivers.[61] Despite the increasing sophistication of these machines and devices, mankind has only fairly recently begun to develop machine-learning systems. Often referred to as deep learning, artificial intelligence, autonomous systems, or thinking algorithms,[62] technologies that are based on self-learning abilities are capable of solving tasks which involve more than one domain.[63] Scholars have deemed machine-learning a potential "game-changer" in both technological and legal spaces.[64]

But this technological advancement does not come without challenges and concerns.[65] Fast-paced thinking algorithms raise general fears of entrusting too much learning power to the hands of machines, to the point

man! The Contradicting History of Accountability in Automated Aviation (Data & Soc'y Rsch. Inst., Compar. Stud. Intelligent Sys., Working Paper No. 1 V2, 2015), http://dx.doi.org/10.2139/ssrn.2720477.

[61] *Ralph Teetor and the History of Cruise Control*, AM. SAFETY COUNCIL, https://blog.americansafetycouncil.com/history-of-cruise-control-2 (last visited Mar. 15, 2021); *A Brief History of Cruise Control*, FRONTIER CHRYSLER (July 20, 2016), https://www.frontierchrysler.ca/history-cruise-control; Stephanie Levis, *Look, Ma, No Hands: Self-Parking Cars Take the Wheel*, GEICO, https://www.geico.com/more/driving/auto/car-safety-insurance/look-ma-no-hands-parking-technology-takes-the-wheel (last visited Mar. 15, 2021).

[62] These are all broad terms which I will very generally group together under systems that can "think", in the sense that they may learn from experience and reach results that were not predetermined in an injective manner. For a detailed discussion of the various definitions of the term "autonomy", see Karni A. Chagal-Feferkorn, *Am I an Algorithm or a Product? When Products Liability Should Apply to Algorithmic Decision-Makers*, 30 STAN. L. & POL'Y REV. 61 (2019).

[63] In general, artificial intelligence systems are divided into "weak AI," which is capable of focusing on a specific task domain, and "strong AI," which is capable of solving tasks associated with multiple domains. *See, e.g.*, Kathleen Walch, *Rethinking Weak vs. Strong AI*, FORBES (Oct. 4, 2019, 6:30 AM), https://www.forbes.com/sites/cognitiveworld/2019/10/04/rethinking-weak-vs-strong-ai/#7421c6d96da3.

[64] Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 513 (2015) ("Robotics will prove 'exceptional' in the sense of occasioning systematic changes to law, institutions, and the legal academy.").

[65] From the legal perspective, certain core principles are put to question when the decision-maker is no longer human but rather an algorithm. Criminal liability, for example, often requires *"intent"* by the offender, an element which naturally is difficult to reconcile with an algorithm. *See, e.g.*, Jeffrey K. Gurney, *Driving into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles*, 5 WAKE FOREST J.L. & POL'Y 393, 419–29 (2015); Jeffrey K. Gurney, *Crashing into the Unknown: An Examination of Crash-Optimization Algorithms Through the Two Lanes of Ethics and Law*, 79 ALB. L. REV. 183, 240–44 (2016) (discussing the difficulties of determining criminal liability with respect to driverless cars). In antitrust law, although intent of anticompetitive behavior may be imputed at the firm level, *see* D. Daniel Sokol, *Reinvigorating Criminal Antitrust?*, 60 WM. & MARY L. REV. 1545, 1590–93 (2019), the question of imputing intent to the algorithms themselves (for example, in coordinating prices) must be addressed, *see, e.g.*, Salil K. Mehra, *Antitrust and the Robo-Seller: Competition in the Time of Algorithms*, 100 MINN. L. REV. 1323, 1324–31 (2015). Contract law, to give another example, requires a "meeting of the minds" for a contract to form. Can such be achieved when obligations are undertaken by an algorithm? *See, e.g.*, Lauren Henry Scholz, *Algorithmic Contracts*, 2 STAN. TECH. L. REV. 128, 132–33 (2017).

where mankind will neither have control nor an intellectual advantage over the very machines it created.[66] Even if such a threshold were never reached,[67] the fact that thinking algorithms are starting to deliver better results than humans in various fields,[68] combined with the fact that their decisions are often inexplicable or non-transparent,[69] raises considerable concerns regarding a future where people are subject to significant decisions whose accuracy and fairness could not be questioned.[70] At the same time, the enormous potential of thinking algorithms to save lives, improve quality of life, and assist economic growth is incomparable with any other type of technology ever seen before.[71] Therefore, determining how to balance the

[66] *See, e.g.*, Bob Lambrechts, *May It Please the Algorithm*, J. KAN. BAR ASS'N, Jan. 2020, at 38, 38 ("According to Ray Kurtzweil, an American inventor, futurist and director of engineering at Google, by 2045, computers utilizing artificial intelligence will surpass human intelligence."); Ryan Dowell, *Fundamental Protections for Non-Biological Intelligences or: How We Learn to Stop Worrying and Love Our Robot Brethren*, 19 MINN. J.L. SCI. & TECH. 305, 306–07 (2018) (referring to a singularity, "a theoretical event in which humans create a technology that leads to a domino effect of rapidly escalating, self-improving intelligence"); Rory Cellan-Jones, *Stephen Hawking Warns Artificial Intelligence Could End Mankind*, BBC NEWS (Dec. 2, 2014), https://www.bbc.com/news/technology-30290540; Maureen Dowd, *Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse*, VANITY FAIR (Mar. 26, 2017), https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x (explaining that tech entrepreneur Elon Musk fears AI so much that he wants to colonize Mars as a "bolt-hole if A.I. goes rogue and turns on humanity").

[67] *See* Dowell, *supra* note 66, at 315 (identifying author John Searle and Google's head of AI research as two people who consider a singularity as fantasy).

[68] In the medical context, for example, algorithms have shown to outperform physicians in several medical tasks, including predicting heart attacks and diagnosing brain tumors. Yamei, *China Focus: AI Beats Human Doctors in Neuroimaging Recognition Contest*, XINHUANET (June 30, 2018, 10:48 PM), http://www.xinhuanet.com/english/2018-06/30/c_137292451.htm; Lulu Chang, *Machine Learning Algorithms Surpass Doctors at Predicting Heart Attacks*, DIGIT. TRENDS (Apr. 17, 2017), http://www.digitaltrends.com/health-fitness/ai-algorithm-heart-attack; Ian Steadman, *IBM's Watson Is Better at Diagnosing Cancer than Human Doctors*, WIRED (Feb. 11, 2013), https://www.wired.co.uk/article/ibm-watson-medical-doctor. For additional examples, see Froomkin et al., *supra* note 2, 39–44.

[69] *See infra* notes 79–85 and accompanying text.

[70] *See, e.g.*, Cade Metz & Adam Satariano, *An Algorithm That Grants Freedom, or Takes It Away*, N.Y. TIMES (Feb. 7, 2020), https://www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html ("The algorithm is one of many making decisions about people's lives in the United States and Europe. Local authorities use so-called predictive algorithms to set police patrols, prison sentences and probation rules. In the Netherlands, an algorithm flagged welfare fraud risks. A British city rates which teenagers are most likely to become criminals."); Natalie Ram, *Innovating Criminal Justice*, 112 NW. U. L. REV. 659, 665–66 (2018) (stating both that "[p]rivately-developed algorithms have come to occupy a key role in criminal justice processes" and that the assertion of trade secret protection by the creators of the algorithms "cripples courts and defense counsel—and sometimes prosecutors, as well—from ensuring accuracy in criminal justice").

[71] W. Nicholson Price II, *Regulating Black Box Medicine*, 116 MICH. L. REV. 421, 432 (2017) (with respect to algorithms in the field of medical treatment, stating that "[a]voiding the implementation of algorithms because we fear the problems that might arise means leaving in place a system of medical errors that we know already exist, and foregoing the potential benefits of innovative treatment options that can save lives"); Ric Simmons, *Big*

desire to quickly develop and deploy such life-saving systems with the need to maintain adequate safeguards against their potential harms is an extremely delicate yet critical task.[72]

What renders thinking algorithms so different from other types of automated and information-processing systems?[73] It is challenging for scholars to define and make a distinction between artificial intelligence (AI) systems and non-AI systems, given that "artificial intelligence" is a broad, nonbinary classification, and by nature is dynamic and ever-changing.[74] It is similarly challenging to define "thinking algorithms." One key characteristic of thinking algorithms, however, is their ability to be trained and to improve their capabilities as a result. Unlike systems that are pre-programmed to execute a certain task in a well-defined manner, thinking algorithms are trained to learn how to achieve a certain mission.[75] As computer science professor Suresh Venkatasubramanian describes, while "traditional" algorithms follow a recipe that a human has created and inputted, thinking algorithms learn on their own how to create the recipe based on the manner in which they were trained, the data on which they trained, and the subsequent conclusions they have drawn from that information.[76]

The training program of algorithms sometimes entails feeding enormous amounts of data to the algorithm, accompanied by right and wrong answers. Under this approach, referred to as "supervised learning", thinking algorithms can develop a model for predicting the right answer for similar datasets that were not included in the training. A greater degree of freedom to come up with their own conclusions or recognize their own patterns is given to algorithms that are "unsupervised." Here, algorithms are fed the same amount of data but are not provided any answers, leaving the system

Data, Machine Judges, and the Legitimacy of the Criminal Justice System, 52 U.C. DAVIS L. REV. 1067, 1070–71 (2018) (pointing out the following advantages of self-driving cars: "they will be safer; reduce traffic congestion; offer increased fuel efficiency; and make automobiles accessible to segments of the population that were unable to drive cars in the past"; suggesting that algorithms in the criminal justice system may be able to "produce decisions that are more fair, efficient, and accurate than human judgment").

[72] Price, *supra* note 71, at 474 ("And while medicine may be especially salient, machine-learning algorithms create risks and benefits that will need to be addressed, measured, and regulated in many contexts.").

[73] For an analysis of how to differentiate between "products" and "thinking algorithms", see Chagal-Feferkorn, *supra* note 62.

[74] Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 300, 404 (2017) ("There is no straightforward, consensus definition of artificial intelligence."); Bryan Casey & Mark A. Lemley, *You Might Be a Robot*, 105 CORNELL. L. REV. 287, 294 (2019) ("And, in the end, it might be impossible to come up with a satisfying definition that regulates only the robots or humans we really want to. This is particularly true because the nature of robots is changing fast, and legal definitions set with today's technology in mind will rapidly become obsolete.").

[75] Tutt, *supra* note 23, at 94–95.

[76] Suresh Venkat, *When an Algorithm Isn't. . .*, MEDIUM (Oct 2, 2015), https://medium.com/@geomblog/when-an-algorithm-isn-t-2b9fe01b9bb5.

free to decipher patterns in the data that may indicate the right answer.[77] Levels of freedom entrusted within the unsupervised system may require very little human involvement throughout the entire process:

> [W]e will soon no longer need (or wish) to provide algorithms with hard-coded hints about how to solve problems. Instead, algorithms will be provided with some basic tools for solving problems, and then left to construct for themselves tools to solve intermediate problems, on the way to achieving abstract goals.[78]

B. *Predictability and Explainability*

As thinking algorithms possess the freedom and unrivaled computational abilities to learn from past experiences and discover hidden patterns,[79] they often yield unpredictable outputs. In fact, not only are such algorithms designed to outsmart the limits of the human mind by drawing conclusions based on massive amounts of data,[80] they also often incorporate online databases into their decision-making processes and may update their prediction models after each decision they make. In short, unless the human programmer designs the algorithm to require authorization for each decision it makes, the algorithm is able to reach conclusions based on new information that the programmer never has the chance to consider.[81] Moreover, much of the information affecting the algorithm's decision-making may be dynamic and constantly changing.[82] In keeping with our inability to predict the choices of thinking algorithms, thinking algorithms are considered black boxes whose inputs and outputs are known but whose decision-making process

[77] *See* Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 93–95 (2014); Org. for Econ. Co-op. & Dev. [OECD], *It's a Feature, Not a Bug: On Learning Algorithms and What They Teach Us – Note by Avigdor Gal*, at 3, OECD Doc. DAF/COMP /WD(2017)50 (June 7, 2017), https://one.oecd.org/document/DAF/COMP/WD(2017)50/en /pdf.

[78] Tutt, *supra* note 23, at 100–01.

[79] *See infra* notes 113–15 and accompanying text.

[80] *See* P.J.G. Lisboa, *Interpretability in Machine Learning – Principles and Practice*, *in* FUZZY LOGIC AND APPLICATIONS 15 (Francesco Masulli et al. eds., 2013); RODNEY A. BROOKS, FLESH AND MACHINES: HOW ROBOTS WILL CHANGE US (2002); Calo, *supra* note 64, at 550–558; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 679 (2017).

[81] Kroll et al*., supra* note 80, at 660 ("'Online' machine learning systems can update their model for predictions after each decision, incorporating each new observation as part of their training data. Even knowing the source code and data for such systems is not enough to replicate or predict their behavior—we also must know precisely how and when they interacted or will interact with their environment.").

[82] For a more detailed discussion of algorithms' unpredictability, see, e.g., Chagal-Feferkorn, *supra* note 21, at 133–135.

remains mysterious and inexplicable.[83] Although several technological measures could be used in order to better understand choices made by the algorithm (for example, annotating code with assertions that signal if the program crashes, or having the program log certain actions in a file before or after they have taken place),[84] many believe that the sheer volume of data at the basis of algorithms' functionality would prevent humans from tracing a particular decision back to a specific, tangible reason.[85]

C. *Strengths and Weaknesses of Algorithms Versus Humans in Decision-Making*

In a growing number of fields, algorithms are achieving better results, on average, than their human counterparts.[86] Society therefore has an interest in relinquishing more decision-making powers to thinking algorithms.[87] However, while thinking algorithms may be superior to human cognitive abilities, they are certainly far from perfect and are likely to cause damage.[88]

[83] Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 184–85 (2017) ("algorithmic decision-making is essentially concealed behind a veil of code, which is often protected under trade secrecy law, and even when it is not, its mathematical complexity and learning capacities make it impenetrable"); Michael Luca et al., *Algorithms Need Managers, Too*, HARV. BUS. REV., Jan.–Feb. 2016, at 96 ("Algorithms are black boxes. . . . [They] often can predict the future with great accuracy but tell you neither what will cause an event nor why.").

[84] Additional approaches can include: organizing the code into modules that can be evaluated separately; providing a specification of the program's behavior as well as proof that could be automatically evaluated to check whether the code satisfies this specification. Kroll et al., *supra* note 80, at 663–65.

[85] *See id.*; *see* Cesare Bartolini et al., *Critical Features of Autonomous Road Transport from the Perspective of Technological Regulation and Law*, 27 TRANSP. RSCH. PROCEDIA 791, 796–798 (2017) ("[I]n AI-based systems, it is hard to identify the exact reason that led to a certain decision. As the training of the AI involves configuring millions of connections by means of a training phase, backtracing the decision to the exact training set of inputs that stimulated those connections is considered an almost impossible task.").

[86] *See* Yamei, *supra* note 68 for examples related to the medical field. In the field of transportation, as well, driverless vehicles are already demonstrating improved safety measured by number of casualties as well as injuries resulting from car accidents. *See, e.g.*, Bryant Walker Smith, *Automated Driving and Product Liability*, 1 MICH. ST. L. REV. 7–20 (2017); Chris Isidore, *Self-driving Cars Are Already Really Safe,* CNN BUS. (Mar. 21, 2018), https://money.cnn.com/2018/03/21/technology/self-driving-car-safety.

[87] "[A]ccording to evidence-based practice, if there is good evidence to suggest that a particular action produces the most favorable outcomes, then that action is the most justifiable one. . . . Once there are expert robots, it will be easier to argue in some instances that they *ought* to be used to their full potential, because the evidence will suggest that in those instances they will, on average, deliver better results than human experts." Jason Millar & Ian Kerr, *Delegation, Relinquishment and Responsibility: The Prospect of Expert Robots*, *in* ROBOT LAW 116–117 (Ryan Calo et al. eds., 2016).

[88] For a list of notorious AI failures in 2018 (some involving tasks where algorithms are assumed to achieve better results than humans), see *2018 in Review: 10 AI Failures*, SYNCED REV. (Dec. 10, 2018), https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983.

In certain situations, a thinking algorithm may be incapable of avoiding harm.[89] In other circumstances, harm may be avoidable yet would involve no fault if it occurred.[90] However, for a significant proportion of harmful decisions, the harm itself is both avoidable and unjustifiable. The underlying legal question of whether these damages *should* have been avoided would then constitute the basis of a reasonableness or other type of legal analysis. But before we consider the applicability of the reasonableness analysis to algorithms, it is important to first examine the differences in the decision-making process of humans versus that of algorithms and their respective advantages and weaknesses. Analyzing these differences will illuminate whether it is a good idea to apply reasonableness analysis to algorithms and will also influence how best to do so.

There are many similarities between the decision-making process of humans and algorithms. First, both humans and algorithms make decisions using the same four-stage cycle, referred to as the "OODA Loop cycle," comprising of: "observe," which refers to information-acquisition; "orient," which refers to information analysis; "decide," where the information gathered and analyzed in the preceding two steps is used to select a decision; and "act," which is the final stage of implementing the chosen decision.[91]

Second, many of the skills and capabilities that make humans good or expert decision-makers are ones characteristic of algorithms as well, but to a greater degree: the existence of a vast and well-organized knowledge base, the ability to handle large amounts of data and to automate certain sequences of actions, and the ability to derive insights from information—even if the information is of low quality or followed by irrelevant "noises."[92]

[89] To take the example of a driverless vehicle deciding whether to swerve to the right and hit another car or hit the car in front of it, either scenario will result in damage, and the question would only be which of the two choices is less damaging. For a detailed discussion of unavoidable harms and 'least-cost' harms caused by robots see Lemley & Casey, *supra* note 27, at 1327–31.

[90] Sophisticated systems, in particular self-learning algorithms, rely on probability-based predictions, and probabilities by nature inevitably "get it wrong" some of the time. Focusing on the damage caused due to a patient or user being on the "bad side of the statistics" does not mean the thinking algorithm erred. For a more detailed discussion, see Chagal-Feferkorn, *supra* note 62, at 63, 84–85.

[91] *See* FRANS P.B. OSINGA, SCIENCE, STRATEGY AND WAR: THE STRATEGIC THEORY OF JOHN BOYD 1–3 (1st ed. 2006). Naturally, many decisions require a constant flow within the loop. In medicine, for example, a physician will gather and analyze information on the patient, and then often their decision will be to send the patient for additional tests, so that the results could then be "put back" into the physician's OODA loop decision-making process when deciding whether to request any further exams, prescribe a certain medicine, hospitalize the patient, and so on. Once a medical diagnosis or medical treatment decision has been made, it is naturally always subject to further decision-making processes comprising the OODA loop stages, verifying that the diagnosis was correct, that the treatment is efficient, that the patient still requires it, and so on.

[92] *See generally* Itiel E. Dror, *The Paradox of Human Expertise: Why Experts Get It Wrong*, in THE PARADOXICAL BRAIN 177, 179 (Narinder Kapur ed., 2011); Vimla L. Patel et

Nevertheless, there are also very significant differences in the decision-making process of humans and algorithms, which will naturally affect the standard of reasonableness to which we can, and should, hold them. Algorithms boast various strengths compared to humans, which often translate into better cognitive performance. First, algorithms have the ability to compute a huge amount of data, so vast that the human mind cannot grasp it, at an incomparable speed.[93] In addition to their elevated computational abilities, thinking algorithms do not have self-interests affecting their judgment,[94] they do not omit any of the decision-making stages,[95] and they are not subject to human physical or mental limitations such as exhaustion, stress, or emotionality.[96] Though algorithms are often accused of being biased,[97] there are at least certain types of cognitive biases that affect humans but not algorithms.[98] Lastly, an algorithm that is put to commercial or mass

al., *Expertise and Tacit Knowledge in Medicine*, in TACIT KNOWLEDGE IN PROFESSIONAL PRACTICE: RESEARCHER AND PRACTITIONER PERSPECTIVES 75, 75 (Robert J. Sternberg & Joseph A. Horvath eds., 1999); *see also* Beverly P. Wood, *Visual Expertise*, RADIOLOGY, Apr. 1, 1999, at 1–3.

[93] *See infra* notes 113–115 and accompanying text.

[94] A human physician conducting research on a certain type of cancer, for instance, might unconsciously select one diagnosis over another because the former would make the patient a candidate for their research. Unless programmed to take such considerations into account, an algorithm would not.

[95] This may thus guarantee a meticulous analysis in each and every case, and at the same time avoid biases based on routine (such as an attorney always citing the same precedent without examining alternatives that may be better suited to particular circumstances, or a physician prescribing a certain medicine even though a different drug might be more efficient for a certain patient). *See* Michal S. Gal & Niva Elkin-Koren, *Algorithmic Consumers*, 30 HARV. J. L. & TECH. 309, 321 (2017).

[96] "[I]n my opinion . . ., one's behavior cannot be controlled at all times by reason or logic, and emotional stress will have a great influence on one's conduct." Petersen v. Honolulu, 462 P.2d 1007, 1010 (Haw. 1969).

[97] *See, e.g.*, United States v. Maclin, No. 19-cr-122-pp, 2019 WL 3240745 at *3 (E.D. Wis. 2019) (defendant in a criminal case arguing that an algorithm used in Milwaukee was "biased against black people"); Arthur Rizer & Caleb Watney, *Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just*, 23 TEX. REV. L. & POL. 181, 210 (2018) ("The core critique here is that algorithms are not completely neutral and objective tools, as they can be biased through the improper curation of data and the selection of variables the algorithms will seek to optimize towards.").

[98] An algorithm, for instance, would not base its decisions on the "availability heuristic." The availability heuristic refers to people's tendency to base their estimations of the likelihood of certain events on prior knowledge that is easily retrievable. The more dramatic, emotional, or unusual an event is, the better people tend to recall it, and inaccurately base their estimations on it. *See* Amos Tversky & Daniel Kahneman, *Judgement Under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124, 1128 (1974). A human lawyer asked to predict a judge's reactions to a certain argument, for example, may recall the judge reprimanding them for raising a similar argument, and will therefore overestimate the likelihood of the judge reacting negatively to their argument. An algorithm, on the other hand, will be affected by the database it "trained" on (and therefore if the data was not representative it would yield poor results), but it would systematically analyse all prior incidents and give them an equal weight, without relying on a particular emotional event. For more discussion on law, heuristics, and

use will likely have been extensively tested and selected over other inferior algorithms. This is in contrast to human decision-makers, as both poor and high-quality decision-makers may still share the same road or profession.[99]

Such advantages render thinking algorithms superior to human decision-makers in a growing number of fields. But algorithms are not better than humans in every aspect. From a technical perspective, although algorithms are not subject to human frailties such as fatigue or stress, they are at risk of suffering from technical malfunctions or being vulnerable to cyber-attacks and gaming attempts by users.[100] Even when operating smoothly, certain algorithmic characteristics can cause the algorithm to make basic or clumsy mistakes that a human is highly unlikely to make, such as classifying photos of people as gorillas[101] or naming Toronto as a U.S. city during the *Jeopardy!* championship.[102]

Although algorithms' potential for knowledge of facts (referred to as "declarative" knowledge, or "how-to" knowledge)[103] is enormous, they are still lacking when it comes to tacit knowledge, or the "things you just know how to do without being able to explain the rules for how you do them."[104] Algorithms also lack both creativity and flexibility, which are often im-

biases see, e.g., Cass R. Sunstein, *Moral Heuristics* 4 (Chicago John M. Olin L. & Econ., Working Paper No. 180, 2003); Cass R. Sunstein, *Hazardous Heuristics* 1 (Chicago John M. Olin L. & Econ., Working Paper No. 165); Russel Korobkin, *The Problems with Heuristics for Law* 3–4 (UCLA SCHOOL OF LAW, L. & ECON. RSCH. PAPER SERIES, 2004). It should be noted, however, that even if algorithms do not base their decisions directly on heuristics or biases, their decisions might nevertheless be affected by "hidden" biases (as is argued, for example, in the context of algorithms in the service of the criminal justice system). *See* Matthias Leese, *The New Profiling: Algorithms, Black Boxes, and the Failure of Anti-Discriminatory Safeguards in the European Union*, 45 SEC. DIALOGUE 494, 494–95 (2014); Toon Calders & Sicco Verwer, *Three Naïve Bayes Approaches for Discrimination-Free Classification*, 21 DATA MINING & KNOWLEDGE DISCOVERY 277, 277–78 (2010). Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016).

[99] Khosla, *supra* note 12.

[100] Importantly, the negative outcome of an algorithm making erroneous decisions (due to malfunction of some sort, or in general) could far exceed that of a human. This is because an algorithmic error might be duplicated to all other algorithms making that same decision, unlike a human whose decisions could vary from the ones made by other decision-makers. Eliav Lieblich & Eyal Benvenisti, *The Obligation to Exercise Discretion in Warfare: Why Autonomous Weapon Systems Are Unlawful, in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 245, 272 (Cambridge Univ. Press, 2016).

[101] *See* Alistair Barr, *Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms*, DOW JONES INSTITUTIONAL NEWS (July 1, 2015), http://www.usatoday.com/story/tech/2015/07/01/google-apologizesafter-photos-identify-black-people-as-gorillas/29567465.

[102] *See* Steve Hamm, *Watson on Jeopardy! Day Two: The Confusion Over an Airport Clue*, BUILDING A SMARTER PLANET (Feb. 15, 2011, 7:30 PM), http://web.archive.org/web/20160422135346/http://asmarterplanet.com/blog/2011/02/watson-onjeopardy-day-two-the-confusion-over-an-airport-clue.html.

[103] Dror, *supra* note 92, at 178.

[104] HARRY COLLINS & ROBERT EVANS, RETHINKING EXPERTISE 13 (2008).

portant factors in becoming an expert in a professional field.[105] Similarly, algorithms are inferior to humans in adjusting their decisions when they encounter new parameters that were not part of their training process, or when such adjustments are not in line with their programmed limitation.[106] Another aspect of algorithms' unhelpful rigidity is that, unlike human decision-makers, they do not discuss their decisions with colleagues and peers. Thinking algorithms therefore have fewer opportunities to identify errors or tweak the decision based on other perspectives.[107] In addition, although algorithms' abilities outperform humans' in many aspects, they face an inherent disadvantage when making decisions calling for certain human traits that they are not yet capable of copying, such as intuition or any other trait related to tacit versus formal knowledge.[108] Lastly, though algorithms may boast unparalleled abilities in analyzing enormous volumes of data and identifying patterns—at a level beyond the capacity of the human brain—their continued improvement in certain fields might very well depend on additional feedback received from humans. Fields characterized by having no clear-cut determinations or golden standards agreed on by humans will likely replicate a given problem to algorithmic decision-makers as well.[109] Moreover, if decision-making authority is exclusively relinquished to algorithms, then the lack of additional training data produced by humans (who would no longer possess the necessary expertise in the fields taken over by algorithms) will prevent detection of algorithmic mistakes and set the quality of their performance at a stagnant level.[110]

How do these general characteristics, as well as more specific technical abilities and limitations, come into play in the decision-making process of an algorithm, and thus shape the potential reasonableness we would be able

[105] "Automation" is, in fact, a source of concern when assessing the performances of human experts, given that flexibility and creativity are essential for their functioning. Dror, *supra* note 92, at 182.

[106] Lieblich & Benvenisti, *supra* note 100, at 29; Thomas J. Barth & Eddy F. Arnold, *Artificial Intelligence and Administrative Discretion*, 29 AM. REV. PUB. ADMIN. 332, 338 (1999).

[107] Lieblich & Benvenisti, *supra* note 100, at 1, 29.

[108] "For instance, [algorithms] might be less accurate than humans in detecting nuances or human body language or gestures indicating that a person is deliberately feeding them inaccurate information." Chagal-Feferkorn, *supra* note 21, at 145.

[109] Adewole S. Adamson & H. Gilbert Welch, *Machine Learning and Cancer-Diagnosis Problem: No Gold Standard*, 381 N. ENGL. J. MED. 2285, 2286 (2019).

[110] Froomkin et. al., *supra* note 2, at 36: "Many ML systems are not easily audited or understood by human physicians, and if this remains true, it will be harder to detect sub-par performance, jeopardizing the system's efficacy, accuracy, and reliability. Once ML systems displace doctors in a specialty, the demand for such doctors will shrink, as will training opportunities for human experts. Because we will continue to need humans to generate much of the training data for future ML systems, this reduction in human competence may create roadblocks to the continuing improvement of ML systems, especially once new diagnostic sensors are available."

to demand of it? To deconstruct the decision-making process of both humans and algorithms we shall use John Boyd's "OODA Loop cycle" mentioned earlier, which characterizes the decision-making process of both humans and algorithms.[111]

i. "Observe" (Information-Acquisition)

The OODA loop's first stage consists of the decision-maker observing the environment in order to collect information that will later be used to make a decision. While human decision-makers make use of their senses in order to collect information, thinking algorithms may collect information using sensors, video cameras, etc.[112] The capabilities of a decision-maker (be it human or algorithmic) in the stage of information-acquisition depend on the decision-maker's ability to locate the information (at times, it will in turn depend on their ability to know that the information exists in the first place), and then to access it. It also depends on their ability to collect the information and store it in a manner that will allow the second stage of information analysis.

As far as the information-acquisition stage is concerned, algorithmic decision-makers boast abilities that are far beyond those of humans. Both knowledge of potentially relevant information and the ability to access and collect information are classic examples of where humans are very limited, but for algorithms, the sky is the limit. First, the dimension of time grants algorithms an incomparable advantage over humans. Not only do algorithms have a much better starting point in terms of free time (having no obligations or need to eat or sleep), but they are also able to run multiple simultaneous actions when searching for information, whereas humans' multitasking abilities are much more limited.[113] Moreover, the processing power of algorithms is incomparably greater than that of humans,[114] allowing them in

[111] *See generally* Osinga, *supra* note 91.

[112] RICHARD A. POISEL, INFORMATION WARFARE AND ELECTRONIC WARFARE SYSTEMS 29 (Artech House, 2013).

[113] Lauren A. Newell, *Redefining Attention (and Revamping the Legal Profession?) for the Digital Generation*, 15 NEV. L.J. 754, 766 (2015) (pointing out that "[a]t any given time, we face severe limits in the number of choices that we can select, the number of tasks that we can execute, and the number of responses that we can generate, along with the limits in the number of items that can be maintained in working memory" and explaining that multitasking comes with significant efficiency costs).

[114] Gal & Elkin-Koren, *supra* note 95, at 318. ("[t]he most basic advantage of algorithms is that they enable a speedier decision. Given any number of decisional parameters and data sources, computers can apply the relevant algorithm far more quickly than the human brain . . . . [In an example of trying to find the best deal for a product] [a]n algorithm may be able to compare a vastly greater number of offers in the same time.").

mere seconds to complete tasks that might take a human hours or even days to finish.[115]

Second, thinking algorithms have an advantage of scale. If a human wants to improve their skill in searching, identifying, and collecting relevant sources of information, that human would work alone to improve their own skill, and that human alone would enjoy the fruits of their efforts. However, in the case of thinking algorithms, large teams of coders all work together to enhance the algorithm's abilities. Any improvements made to one thinking algorithm can be directly used to improve other algorithms. Take for example learning how to read Japanese. In order for a human to become highly proficient in the language, that individual would likely have to invest hundreds, if not thousands of hours before they could easily read Japanese text. From an economic point of view, investing such efforts to be able to understand text is likely not justifiable if the human is a physician who wants to learn Japanese in order to be able to read Japanese-language medical studies. However, the same does not hold true for an algorithm. Although teaching an algorithm how to read Japanese might require a significant effort, once the algorithm has learned how to read Japanese, all units of the system (potentially millions of them) would have that capability. Developing algorithms' ability to reach and access even the most difficult to access information, therefore, might very well be economically justifiable when developing a human's ability would not.

As a result of these two advantages, thinking algorithms' ability to identify and access sources of various languages, of various media platforms*,* of various content types*,* and of enormous volume reaches far beyond the capabilities of human decision-makers. A medical algorithm, for example, might be able to look for sources of information in all existing languages, while a human physician could not.[116] A driverless car could listen to all radio stations within a vicinity to collect information about nearby hazards, while simultaneously collecting information from other diversified platforms such as navigation applications, consumer blogs, or online magazines. Moreover, unlike humans, thinking algorithms would not need to limit themselves to only searching for information directly about the situation at hand, but would be free to look for sources of information from completely different fields and potentially discover surprising correlations. While a human physician would likely collect information from medical journals, professional protocols, and guidelines, a medical algorithm may look for addi-

[115] *Artificial Intelligence Singles Out Neurons Faster Than a Human Can*, SCIENCE DAILY (Apr. 12, 2019), https://www.sciencedaily.com/releases/2019/04/190412150628.htm (describing an algorithm that can identify and segment neuros in minutes, a task that could take a trained human researcher up to 24 hours, assuming that the researcher is "fully focused for the duration and [doesn't] take breaks to sleep, eat or use the bathroom.").

[116] In the field of medicine this could be meaningful—for example, in the context of viruses or parasites that are endemic for specific regions.

tional data pertaining to the weather, levels of pollution, and levels of radiation. The medical algorithm might then determine that such data is relevant in making a medical diagnosis or in determining the optimal treatment for a condition.[117] Lastly, the sheer volume of information that algorithms may access makes its information-acquisition stage look vastly different than a human's information-acquisition stage. While a human driver is able to collect information on the speed and distance of their own vehicle and identify the presence of other vehicles and pedestrians around them, a driverless vehicle could theoretically identify all cars in a thirty mile radius and collect information on how these vehicles have been driving for the past minute, day, or week. Such information might alert the driverless car to drunk drivers, generally dangerous drivers, or autonomous vehicles that seem to have malfunctioned. It could also access data on the mechanical condition of all other cars on the road, potentially informing on the extent to which each car might be damaged in case of an accident, in a manner that might be relevant to the notorious dilemma of which car to hit when a crash is unavoidable.[118] In the medical context, a human physician would be able to rely on many sources of information pertaining to the patient's current and previous medical condition, along with potential treatments and their risks, but a medical application would also be able to access medical records of billions of patients worldwide. This could allow the medical application to identify epidemic patterns in real time[119] or learn how effective different treatments are among populations that are less common at the patient's current hospital.

The accuracy of the parameters collected is also very different where thinking algorithms are concerned. While a human driver might make a good estimation of their own speed even without consulting the speedometer, the human is unlikely to accurately estimate the speed or distance of other objects. Estimations as to current wind speed would likely not be available to a human driver at all. A driverless car, on the other hand, would have an accurate numerical value for each of these variables.

As far as identifying and accessing relevant data is concerned, thinking algorithms boast abilities that are light-years ahead of those of humans. This does not mean, however, that algorithms will always be better at infor-

[117] *See, e.g.*, Isobel Braithwaite et al., *Air Pollution (Particulate Matter) Exposure and Associations with Depression, Anxiety, Bipolar, Psychosis and Suicide Risk: A Systematic Review and Meta-Analysis*, 127 ENV'T HEALTH PERSPS. 126002-1, 126002-8 (2019) (using meta-analysis accumulated over forty years and identifying correlation between pollution and depression). A thinking algorithm able to access relevant sources of information on pollution may reach such findings well before they are published in a paper.

[118] For an ethical analysis of said dilemma and related ones, see Gurney, *supra* note 65.

[119] In 2009, Google's algorithm was able to identify where the H1N1 virus was in real time using its users' search terms. In contrast, the Centers for Disease Control and Prevention (CDC) relied on reports by physicians and identified where the virus had spread two weeks later than Google. *See* Tutt, *supra* note 23, at 97.

mation-acquisition than humans, as humans possess both creativity and tacit knowledge. For instance, a patient mentioning that he visited a foreign country recently might prompt a human physician to look for relevant sources of information relating to said country and thus correctly diagnosing the patient with an endemic disease, whereas an algorithm might not necessarily have this conversation to begin with or might ignore this information as it is not directly related to the task of diagnosis.[120]

As for storing collected information, algorithms clearly outpace humans in terms of memory capacity.[121] However, this does not necessarily mean that algorithms collect information in a meaningful way. Although algorithms can be programmed to "understand" various languages and therefore would be able to extract information from text or potentially audio in any language, algorithms would be limited by their declarative knowledge[122] and would only be able to extract information that resembles that with which they are already familiar. Once exposed to a different layout, either older than the one the algorithm trained on or a new technological advancement, then information that would be obvious to a human would be incomprehensible to an algorithm. For example, if a medical algorithm was only trained to decipher printed text, then the algorithm would completely miss any handwritten notes on the patient's medical record. Similarly, if the patient had been sent for a series of scans and the hospital had just moved to a more advanced scanning technology providing higher-resolution images, a human physician might not detect the difference while an algorithmic doctor might fail to interpret the scan altogether.[123]

The information-acquisition stage is therefore characterized by algorithmic performance that is oftentimes vastly superior to that of humans.

[120] Even in cases where the algorithm is programmed, or decides on its own, to ask the patient about recent travels, it may still miss information that a person would have picked up. If, for example, the patient says he did not visit a foreign country, a human physician might be more likely than an algorithm to detect other indications revealing that the patient did indeed visit the foreign country. For instance, a patient may have visited several countries and forgot he also visited the one in question, perhaps because his illness caused him confusion, or because he may not wish to reveal his visit to that country. Also, if the patient mentions that he cut his knee when climbing Mount Fuji, a physician will conclude the patient travelled to Japan, where the algorithm might not. In such cases, only the human physician would know to look for sources of information relevant to endemic diseases in Japan.

[121] Kris Sharma, *How Does the Human Brain Compare to a Computer?*, CRUCIAL (Aug. 28, 2019), https://www.crucial.com/usa/en/how-does-the-human-brain-compare-to-a-computer (pointing out how humans can forget information, while computers never will).

[122] *See supra* notes 103–05 and accompanying text.

[123] Froomkin et al., *supra* note 2, at 74 ("Imagine, for example, that someone invents a higher-resolution scanner that takes sharper images than its predecessor. Human beings who could recognise tumours on the old photos might have little or no difficulty recognising the same tumours on the new, sharper images; ideally, humans might also be able to see new things they had not been able to discern or become able to better distinguish previously ambiguous results. Unfortunately, ML systems do not work like that. To an ML system, the new, higher-resolution image is a completely new thing.").

Concerns we might have over whether it is economically feasible for a human physician to read long medical records when treating patients with urgent needs would likely not be applicable to a medical algorithm capable of collecting enormous amount of information in a split second. On the other hand, if a patient's medical record included handwritten notes about allergies, and the algorithm is not capable of deciphering those notes, then human information-acquisition would be clearly superior.

ii. "Orient" (Information-Analysis)

Similar to the "observe" stage of the OODA loop, the "orient" or information-analysis stage is greatly affected by technical abilities. In that sense, algorithms' superior computational abilities are similarly advantageous in this stage. Algorithms are able to analyze significantly more information and in a more methodical manner than humans; once a certain threshold of parameters to consider is crossed, humans will no longer be able to assimilate additional parameters in their analysis.[124] A driverless car will therefore not only be able to collect data on all driving records of all vehicles surrounding it, but it will also be able to actually process this input in an insightful manner. Moreover, while humans weigh different parameters based on their general knowledge and gut feelings, algorithms attach concrete numerical values to the different parameters based on statistics from past experiences.

Unlike the "observe" stage, however, the "orient" stage involves much more than technical-capability issues or the algorithm's ability to complete a certain task. First, the information-analysis stage involves questions that are not exclusively a matter of ability: deciding which elements of the information collected to disregard and which to take into account, along with addressing legal and ethical concerns.[125] A classic example is that of the self-driving car just about to crash—the algorithm has to decide how much weight to attach to its driver's chances of survival versus of those of other road users. Autonomous weapons similarly raise crucial legal and ethical questions with respect to how an algorithm decides how much weight it attaches to the presence of civilians versus the importance of the task at hand.

Second, an algorithm's material decision of how much weight to attach to each parameter will, for better or worse, not be affected by emotions, un-

[124] *See* Gal & Elkin-Koren, *supra* note 95, at 318 ("Given any number of decisional parameters and data sources, computers can apply the relevant algorithm far more quickly than the human brain.").

[125] Among the tasks involved in the orient stage is those of updating the decision maker's perception of the world, based on the information acquired in the previous stage, and then determine if and how desired goals are met under that new perception. Poisel, supra *note* 112, at 30. The analysis of the collected information therefore involves an input on what the system aims to achieve, in a manner that may involve nontrivial questions that require more than computation abilities.

like a human decision-maker. A human physician treating a dying child may, out of compassion or optimism, attach a higher weight than an algorithm would to a certain treatment that may only have a small chance of saving the child. Likewise, a human driver in danger of hitting a child may not, in the moment, be able to properly consider all the relevant factors and the actual chance of hitting the child, whereas an algorithm would likely take into account that the potentially injured party would be a child, but would at the same time impassively attach significant weight to other factors. In addition, algorithms and humans try to maximize different utility functions. For example, an algorithm might be taught to attach significant weight to cost-efficiency considerations. A medical algorithm therefore might attach more weight to the price of a drug than the average human physician would, as the algorithm's main goal might not be to cure patients but rather to balance budget considerations.

Another concern in deciding how to weigh parameters is bias. Human decision-makers may be susceptible to bias or to subconsciously focusing on their own welfare. A human physician might underrate parameters that indicate that a certain treatment is inferior to another treatment that she is also conducting research on, or when fatigued at the end of a long shift, she may overrate the significance of certain vital signs that indicate no emergency treatment is needed when other signs indicate the opposite. Attaching significant weight to parameters based on self-interest or bias is also very prevalent in the context of autonomous vehicles. While a human driver would naturally attach significant weight to their own well-being (leading potentially to an instinctive decision to protect oneself at the expense of hitting others), an algorithm might, depending on the choices of its manufacturer, seek to maximize "greater good" considerations and potentially give significant weight to the number of potential victims or the severity of potential injury in order to choose the least-damaging action.

The "orient" stage, therefore, is characterized by both inherent technical advantages of algorithms over humans, but it also reflects material differences in their decision-making rationales and goals.

iii. "Decide" (Decision-Selection)

Having attached different weights to the myriad of parameters gathered during the information-acquisition process, the human or algorithmic decision-maker now has to make an actual decision based on their analysis and desired outcomes.[126] This stage involves more than might meet the eye, given that decisions are almost never a binary choice between two alternatives of right or wrong, but rather comprise of choices whose relative benefits and

[126] *Id.*

risks can only be estimated by an iterative process of prediction.[127] Moreover, any choice of a certain treatment over others would not necessarily allow Pareto efficiency.[128] If one alternative has higher rates of success, lower chances of damage, and a lower cost, then choosing it would naturally be an easy decision to make as there is no need for either normative analysis or prioritization of different goals. However, if a treatment that has high recovery rates is also associated with more severe side effects, then the decision-maker would be required to factor in the level of risk aversion they feel comfortable with.

The stage of "deciding" therefore reflects similar differences between humans and thinking algorithms to those discussed in the prior stage. From a technical perspective, an algorithm has the capacity to decide between a very large number of alternatives, whereas the human brain is limited to truly evaluating only a few.[129] Furthermore, as with the action of data analysis, when deciding between the relevant options, a human would base its decision on a fuzzy estimation of why a certain alternative is favorable to another, rather than on a numerical calculation.

Another difference between humans and thinking algorithms is the presence of instinct. When a human driver is faced with an impending accident and has to decide which direction to swerve the car, the human will make a choice, presumably the self-preserving one, between all possible options based on instinct.[130] Similarly, a human physician deciding between two different treatment approaches in an urgent life or death matter will likely also make her choice instinctively. The thinking algorithm, on the other hand, would make instant choices like humans, but these choices would be made after fully weighing the available options and not on instinct. Certainly, human instinct may at times be sharper than the algorithmic process, as humans are able to factor in circumstances the algorithm may not be familiar with and would not know how to consider (this was very dramatically demonstrated in the U.S. Airways crash on the Hudson in 2009).[131] But overall, thinking algorithms have the general advantage of be-

[127] *See, e.g.*, Barocas, *supra* note 98, at 681. For instance, the selection of a given medicine among many alternatives for a patient might involve an 80% likelihood that a certain tumour is indeed the type of tumour assumed, a 50% likelihood that it will respond well to a certain medicine, a 20% likelihood that its side effects would be acute, and so on.

[128] Pareto efficiency refers to a reallocation of resources resulting in at least one component is better off compared to the previous allocation, while all other components remain unharmed. A DICTIONARY OF FINANCE AND BANKING (Jonathan Law & John Smullen eds., Oxford Univ. Press 2008).

[129] *See supra* notes 113–115 and accompanying text.

[130] *See, e.g.*, Giuseppe Contissa et al., *The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law*, ARTIFICIAL INTEL. & L. 365, 368 (2017).

[131] The famous "Miracle on the Hudson," where U.S. Airways pilot Chesley "Sully" Sullenberger safely performed an emergency landing after two of the airplane's engines suddenly failed, is an example of human's superiority in that context, given that pilot Sullen-

ing able to make rational and well-thought-out decisions rather than instinctive ones, even in times of emergency.[132]

iv. "Act" (Action-Implementation)

The last stage—the actual execution of the decision made in the preceding stage—is one generally characterized by algorithmic superiority for those decisions that require precision, such as administrating exact dosages by a radiation therapy machine, enabling very small corrections in the movement of a surgeon's hands, or launching an air defense rocket in the split second that would allow its trajectory to collide with that of an aerial threat.[133] However, humans are better at making certain kinds of decisions that require an understanding of empathy and social nuances, such as a doctor explaining to a patient that she has a terminal illness with no treatment options, or a priest engaging in religious confession with a parishioner. This stage emphasizes the difference between executing automated and non-automated actions.[134] Whether an algorithm or a human will be more successful at executing a particular action in this final stage will heavily depend on the type of action to be executed and whether that action requires precision or a form of tacit knowledge.

Although humans and thinking algorithms use the same four stages of the decision-making cycle, the two differ in how they operate within each stage. Humans undoubtedly reach their decisions in a different manner, which is reflected in all four stages of the decision-making process. Although algorithms show superiority when making certain types of decisions, especially those related to the technical ability to collect and process enormous amounts of information, algorithms also suffer from certain weaknesses compared to humans and may make certain mistakes that a human would not. The expectations we hold of each decision-maker, therefore, would vary in each of the decision-making stages. Our expectations will likely differ not only from the technical perspective of ability (which may

berger made the "right decision," in contrast to flight algorithms, which would have seemingly led to a catastrophic crash. Clint Eastwood's 2016 movie *Sully* focuses on that exact point. SULLY (Flashlight Films 2016). *See also* Adam Smith, *The Miracle on The Hudson: How It Happened*, TELEGRAPH (Nov. 22, 2016), https://www.telegraph.co.uk/films/sully/miracle-on-the-hudson-how-it-happened.

[132] *See supra* notes 68 and 86 (discussing examples where algorithms already outperform human decision makers, including in the context of autonomous vehicles and a reduced rate of car accidents).

[133] For example, the Da Vinci is a minimally invasive robotic surgery system that translates a human surgeon's hand movements into smaller, more precise, movements. *See generally* DA VINCI SURGERY, www.davincisurgery.com (last visited Jan. 22, 2020).

[134] For a discussion on the differences between automation and autonomy, *see* William C. Marra & Sonia K. McNeil, *Understanding "The Loop": Regulating the Next Generation of War Machines*, 36 HARV. J.L. & PUB. POL'Y 1140, 1150 (2013); Chagal-Feferkorn, *supra* note 62, at 70.

not always be characterized by algorithmic superiority), but also from the open-ended normative perspective of the different goals by which each decision-maker is guided.

Part III: Is The Reasonableness Analysis Compatible with Algorithmic Tortfeasors?

A. *Can the Algorithm Itself Be "Reasonable," and What Would That Mean?*

Before delving into the compatibility of reasonableness analyses with algorithms and proposing a concrete model for shaping what this analysis might look like, there is a basic conceptual issue that must be addressed first. Although thinking algorithms do not yet have legal personality,[135] each algorithm was programmed by an entity that does. Therefore, we must ask: does the exercise of assessing the reasonableness of an algorithm's behavior have any meaning, if performed separately from the reasonableness of its manufacturer? This conundrum is divided into two distinct concerns. First, a major objection to the development and application of a "reasonable algorithm" standard might be the notion that the reasonableness of an algorithm is actually one and the same as the reasonableness of its programmer.[136] Those who hold this objection deem that the reasonableness of the human programmer—and not that of the algorithm—is what needs attention.[137] As discussed at length in a previous paper,[138] however, there are several factors that render the reasonableness of the algorithm non-equivalent to that of its programmer. Among them is the dissimilarity in the time between when a programmer decides how to program a thinking algorithm and when that algorithm later makes a damaging decision (a time lapse that, in the case of thinking algorithms in comparison to less sophisticated systems, might have

[135] For a discussion on granting AI systems legal personality see, e.g., Ben Allgrove, Legal Personality for Artificial Intellects: Pragmatic Solution or Science Fiction? (June 2004) (MPhil dissertation, Oxford Univ.) (https://ssrn.com/abstract=926015). *See also* European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics, EUR. PARL. DOC. P8_TA(2017)0051 (suggesting to grant autonomous robots an independent legal status of "electronic persons": "[C]reating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently.").

[136] Balkin's "homunculus Fallacy" argument refers precisely to this point. According to Balkin, "there is no little person inside the program." Instead, algorithms act as they are programmed to act—no more, no less. *See* Balkin, *supra* note 9, at 1234.

[137] Spoiler alert: the reasonableness of the programmer of the algorithm, or its manufacturer, will indeed be relevant in our model for assessing algorithmic reasonableness.

[138] Chagal-Feferkorn, *supra* note 21.

extreme significance).[139] Another differentiating factor to consider is that reasonableness will be measured based on the programmer's professional field, and not the algorithm's application itself. For example, if a medical application makes a harmful decision, a positive analysis will compare the programmer's actions to how other human programmers might have programmed a medical application, rather than comparing the algorithm's decision against how a human doctor would have acted. This might lead to occasions where an algorithm's behavior would be deemed reasonable, while its programmer's behavior would not, and vice versa.[140]

Second, it is arguable that a determination that an algorithm acted unreasonably would be meaningless in terms of tort law policy, as an algorithm itself cannot pay damages from its own pocket—nor can it be financially deterred from adopting insufficient safety measures in future.[141] A counterargument is that thinking algorithms do not need to bear the consequences of their unreasonableness themselves for the aims of tort law to be met. Rather, a "reasonable algorithm" analysis could be used as a tool for determining the liability of the manufacturer of the system or the user of the system,[142] such that they be required to pay for damages caused by their algorithm's unreasonable decision.[143] This would be similar to analyzing the behavior of an employee when determining whether their employer is vicariously liable for their actions,[144] or how the behavior of dogs in dog-attack cases affects whether their owners are found liable.[145] With respect to deterrence, although algorithms themselves would not be affected by a finding that they acted unreasonably, this still does not mean that it is impossible to deter an algorithm from tortious conduct. A thinking algorithm may be programmed to consider the potential consequences of a finding of unreasonableness as part of the parameters it weighs before reaching its decision, at least to a certain extent—as discussed next.

[139] *Id.* at 136.

[140] *Id.* at 132–139.

[141] *See, e.g.*, Colonna, *supra* note 60, at 102–04.

[142] While manufacturers are probably the first to come to mind in the context of algorithmic liability, users' actions too may be subject of liability. A user can cause damage, for example, by intentionally (or unintentionally) feeding the algorithm misleading information. Users can also install "patches" of different sources that would alter the choices of the algorithm in a damaging manner. *See, e.g.*, Lothar Determann & Bruce Perens, *Open Cars*, 32 BERKELEY TECH. L.J. 915, 935 (2017).

[143] *See* Lemley & Casey, *supra* note 27, at 1351–53; Chagal-Feferkorn, *supra* note 21, at 139–43.

[144] *See generally* Catherine M. Sharkey, *Institutional Liability for Employees' Intentional Torts: Vicarious Liability as a Quasi-Substitute for Punitive Damages,* 53 VAL. U. L. REV. 1 (2019).

[145] In general, liability is not imposed if the dog reacted "proportionally" in response to a provocative act. *See* Jay M. Zitter, Annotation, *Intentional Provocation, Contributory or Comparative Negligence, or Assumption of Risk as Defense to Action for Injury by Dog*, 11 A.L.R. 5TH 127 (2010).

B. *Does the Standard Fit?*

i. Yes! Reasonableness is a Flexible Standard for a Dynamic Decision-Maker

Assuming we can develop a suitable method for assessing the reasonableness of algorithms, a tort framework would indeed be compatible with the nature of thinking algorithms. First, it may often be impossible to understand exactly why an algorithm made a certain decision.[146] Given that thinking algorithms no longer need to follow a pre-defined recipe but rather are capable of creating the recipe on their own, it may be impossible to understand the reason behind the algorithm's damaging choice.[147] As discussed earlier, a reasonableness analysis can be successfully applied in situations where it is not known why a tortfeasor acted as it did. Rather, both normative and positive approaches focus on known parameters, such as reviewing the cost of adding precautions versus the expected damage or examining whether other similarly situated peers would have achieved the same end result.

Second, the reasonableness assessment is indifferent to the type and nature of the tortfeasor whose actions are to be analyzed.[148] At a time where experts are endeavoring to find a legal solution for these unique systems that do not resemble any other tortfeasor (being non-human, on the one hand, but independent and unpredictable, on the other), the good old reasonableness standard may be easily applied. A legal solution specific to algorithms is not needed, as the reasonableness assessment can already adjust to different types of tortfeasors. In the case of the meningitis patient, if the medical record included a handwritten note in Spanish about the patient's allergy, then the normative economic efficiency analysis would be different for a human physician who is not fluent in Spanish compared to a medical algorithm. The exorbitant cost of having each physician study all potential languages that another doctor might use is significantly higher than the probability that important information will be expressed in an unfamiliar language multiplied by the magnitude of damage. But the result would likely be the opposite for an algorithm, if teaching languages to algorithms can be done quickly and would only require a one-time investment to replicate the skill throughout the system. None of these considerations, however, changes the essence of the normative assessment itself, which could be applied to think-

[146] *See supra* notes 79–85.

[147] By looking at the code, "what we would see is 'a mysterious alchemy in which each individual step might be comprehensible,' but any 'explanation' of why the code does what it does requires understanding how it evolved and what 'experiences' it had along the way." Perel & Elkin-Koren, *supra* note 83, at 189–90 (citing Suresh Venkatasubramanian, *supra* note 76).

[148] *See supra* notes 53–55 and accompanying text.

ing algorithms, despite the many inherent differences between them and humans.

In addition to accommodating different types of tortfeasors, including ones that are new kids on the block, the adaptability of the proposed reasonableness assessment would also be compatible with the dynamic nature of thinking algorithms' abilities. The powers of thinking algorithms grow at a dazzling speed, due to both the rapid pace of technological advancements as well as the self-learning abilities of the algorithms themselves. The strengths that algorithms currently possess in the decision-making process may only be the tip of the iceberg of future algorithms' capabilities. But the ever-evolving nature of thinking algorithms would only alter the potential results, not the normative or positive mechanisms of assessing reasonableness. For example, a driverless car hitting a very small animal because technological boundaries render it almost impossible to detect objects of such a small size while driving will likely be held reasonable under both normative and positive analyses.[149] As soon as the driverless car learns how to overcome the technological challenge, however, the same reasonableness analysis would yield the opposite result. But the exact same mechanism for determining reasonableness will be applied in both situations. In the context of rapidly evolving technology, the neutrality of the reasonableness assessment has great value. When law is said to chase technological improvements,[150] the neutral reasonableness assessment may be counted on without the expensive—and often futile—need to constantly reshape the legal framework when the technology advances.[151]

[149] Assuming that developing technology that would allow such detection would be very expensive and exceed the cost of expected damage (from a normative perspective) and that all other driverless vehicles too are unable to detect small objects (from a positive perspective).

[150] *E.g.,* Karen Eltis, *Breaking Through the "Tower of Babel": A "Right to be Forgotten" and How Trans-Systemic Thinking Can Help Re-Conceptualize Privacy Harm in the Age of Analytics*, 22 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 69, 82 (2011) ("It bears repeating: the law cannot keep chasing after technology; it will inevitably (by its very nature) be outpaced, often before the proverbial ink dries.").

[151] Larry Downes, *America Can't Lead the World in Innovation if the FAA Keeps Dragging Its Feet on Drone Rules*, WASH. POST (Oct. 8, 2014), https://www.washingtonpost.com/news/innovations/wp/2014/10/08/america-cant-lead-the-world-in-innovation-if-the-faa-keeps-dragging-its-feet-on-drone-rules (identifying how the FAA has failed to meet several self-imposed deadlines, as well as Congress's deadline, to promulgate rules relating to personal drone usage). Despite Congress starting to address the issue in 2012 (as stated in the Washington Post article), the FAA is still developing its drone rules, as evidenced by the Agency not proposing a rule requiring registration until Dec. 26, 2019. *See* Press Release, Fed. Aviation Admin., U.S. Department of Transportation Issues Proposed Rule on Remote ID for Drones (Dec. 26, 2019), https://www.faa.gov/news/press_releases/news_story.cfm?newsId=24534; Tam Herbert, *Can the Government Keep Up with the Pace of Tech?,* TECHNOMY (Nov. 11, 2018), https://techonomy.com/2018/11/can-government-keep-pace-tech (pointing to U.S. Senator Orrin Hatch being unaware that Facebook makes money by selling advertising and several internet pioneers claiming that the FCC

In addition to the technical compatibility of the reasonableness assessment to thinking algorithms, the assessment is also compatible with striking the right balance between the desire to promote safety and the need to avoid a chilling effect on technology. As discussed in Part I,[152] a reasonableness method for determining liability allows courts to shape social behavior in a multi-layered manner that may factor in various and sometimes competing interests and goals. Holding that a physician was unreasonable because they did not read the medical record despite the facts that the record was short, that no other patients were waiting to see the physician, or that the medical condition of the meningitis patient was not critical and could have allowed the additional time required to read the report would together send a multifaceted message as to the various considerations to be taken into account when striving to act reasonably. The ability to shape a flexible standard of behavior that is capable of factoring in several, potentially contradictory, considerations is perhaps of even greater importance in the case of thinking algorithms than of humans. As we saw earlier,[153] self-learning systems provide greater opportunities for mankind, but also greater dangers, than any previous technology. The ability to delicately balance the need to apply as many safeguards as possible with the desire to encourage the development of such systems is therefore of great relevance in our era. Moreover, with the rapid changes in self-learning systems' abilities and the pace at which new potential hazards as well as benefits are being discovered, the flexible, dynamic reasonableness assessment proposed here will grant the courts leeway to adapt both the considerations they take into account as well as the weight given to each in order to advance the most optimal policy.

ii. But Flexibility is Also the Greatest Challenge

The great advantage of the reasonableness assessment's flexibility is also its most significant obstacle in the case of thinking algorithms. To fulfill the court's aim of influencing the choices of potential wrongdoers, the entity whose behavior is being influenced must understand, to a certain extent, what is expected of them and how to comply with those expectations. When expectations are phrased as concrete rules (e.g., do not cross an intersection when there is a red light), both humans and algorithms can easily comply. In fact, an algorithm will show perfect rates of compliance, for as soon as they are programmed to do something, they will not question it.[154] However, when the desired behavior is not set as a concrete rule but instead is an open-ended standard, algorithms simply lack the common sense that enables

lacks a fundament understanding of how the internet works as two examples of how out of touch Congress is with technological advancement).

[152] *See supra* notes 51–52 and accompanying text.

[153] *See supra* notes 58–72 and accompanying text.

[154] Lemley & Casey, *supra* note 27, at 1370.

humans to "read" a given situation and understand, without concrete guidance, what is reasonable and what is not:

> [P]eople understand that rules and injunctions come with the implied catchall "unless you have sufficient justification for departing from the rule" exception. Try telling that to a robot, though. Machines, unlike at least some humans, lack common sense. They operate according to their instructions—no more, no less. If you mean "don't cross the double yellow line unless you need to swerve out of the lane to avoid running over a kid" you need to say that.[155]

In other words, clearly defined rules should pose no compliance problem for algorithms. But the reasonableness assessment often deals with scenarios that are not black and white, where several rules and interests may contradict one another ("do not run a red light" and "do not hit a pedestrian" for example, will conflict when the only way of avoiding a pedestrian is to cross a red light). Without specific guidance as to the exact circumstances where one rule or one consideration trumps the other, the thinking algorithm will likely not be able to use the vague concept of reasonableness to adjust its choices accordingly. If an algorithm is armed with sufficient data on what was deemed reasonable and what was not by courts, it could theoretically learn how to act reasonably from experience. Self-learning is, after all, the entire notion behind using such algorithms in the first place. But given the enormous volume of previous cases required in order for an algorithm to self-learn, it seems very unlikely that sufficient data could be accumulated to enable thinking algorithms to understand on their own what reasonableness is (a conclusion that would, of course, be individually required in each sector, with respect to each type of damage). From a normative perspective, the reasonableness analysis may be flexible and dynamic enough to account for new types of tortfeasors such as thinking algorithms, but directing the behavior of said tortfeasors might prove impossible.

The reasonableness assessment presents a practical challenge from the positive perspective as well. Comparison of reasonableness with algorithms' peers is not yet available. Whether a positive comparison would be against an average algorithm, against the majority of algorithms, or against a general sense of what other algorithms would do, we simply do not yet have a compilation of comparable cases that would consistently enable a comparative analysis to the behavior expected from other algorithms. The "reasonable person" and "reasonable professional" standards have become possible because there are countless reference points with which to compare the tortfeasor's behavior. The reasonableness expected from potential wrongdoers would be a standard that is formed, and evolves, over a long period of time, allowing for the accumulation of countless relevant comparable cases, such

[155] *Id.* at 1371.

that even with respect to the specific circumstances of each damaging act, there could be relatively similar occurrences that would offer a relevant comparison.

While thousands of companies and startups worldwide are developing thinking algorithms,[156] the accumulated knowledge as to what other algorithms would have done in a given situation is limited. First, the development and usage of thinking algorithm systems is a relatively new phenomenon.[157] Second, barriers to entry such as high development costs or heavy regulation might lead to cases where very few brands dominate the entire market. Third, the technology itself is dynamic, such that certain systems may very quickly become obsolete and other systems may be one of a kind.[158]

Although judges and juries may compensate for the lack of equivalent reference points for human tortfeasors by drawing on their own personal perspectives or intuitions of what a reasonable person or professional would do, judges and juries will likely lack personal understanding of how thinking algorithms could or should act in similar circumstances.[159]

While one practical solution allowing for a positive analysis of reasonableness would be to compare thinking algorithms to reasonable humans, such comparison is not necessarily desirable. As demonstrated previously, thinking algorithms oftentimes boast significant advantages over humans in their decision-making process. Basing algorithmic reasonableness on that of humans would be counterproductive as far as the safety rationale of tort law is concerned, as it would not incentivize manufacturers to adhere to an elevated standard of care even if such a standard were available:

[156] For details on the number of AI start-ups per country, see *Number of Artificial Intelligence (AI) Startups Worldwide in 2018, by Country*, STATISTA (Mar. 2, 2020), https://www.statista.com/statistics/942657/global-ai-startups-by-country.

[157] Autonomous vehicles, for example, have been envisioned for many years, but have only been put to actual use on a commercial scale in the past decade. *See, e.g.*, Lemann, *supra* note 23, at 160.

[158] For example, the case of Open AI's language generator "GPT-3" which is, at least for the moment "an entirely new type of technology. . . . There's no precedent for it. . . ." Ben Dickson, *The implications of Microsoft's Exclusive GPT-3 License*, TECHTALKS (Sept. 24, 2020), https://bdtechtalks.com/2020/09/24/microsoft-openai-gpt-3-license.

[159] "[O]n the other hand, asking what a 'reasonable autonomous vehicle' would have done in a similar situation may make the lawsuit far more complicated, at least during the introduction of these vehicles. Jurors can relate with a human driver and understand what a 'reasonable human driver' could have done in a similar situation. Jurors may not, without expert testimony, be able to determine what a reasonable autonomous vehicle could have done under similar circumstances." Gurney, *supra* note 21, at 11. "[B]ut an important difference between machine learning algorithms and humans is that humans have a built-in advantage when trying to predict and explain human behaviour. Namely, we evolved to understand each other. Humans are social creatures whose brains have evolved the capacity to develop theories of mind about other human brains. There is no similar natural edge to intuiting how algorithms will behave." Tutt, *supra* note 23, at 103.

> Indeed, where once "custom"—what most people in the trade or profession do and have generally done—was the starting point for measuring the appropriate standard of care, US courts today are somewhat suspicious of custom-based arguments on the theory that these arguments provide too little incentive to modernize and may favor entrenched modes of service provision at the expense of the victim.[160]

The concern of holding algorithms to ill-suited standards will not be solved by applying a human-based elevated standard (such as granting safe harbors to algorithms that, on average, perform better than humans).[161] First, although applying a human-based elevated standard would account for the algorithm's general superiority, such a proposal ignores the continuous improvement of thinking algorithms' abilities resulting both from manufacturers' technological advancements and from the algorithms' own self-learning abilities. A tort analysis that is satisfied with a constant level of superiority of algorithms compared to humans does not incentivize continuous improvement.

One may offer to periodically raise the bar required in order to enjoy the safe-harbor protection: for instance, allowing a gap of 10% in favor of algorithms in the first year, but requiring 20% in the following one. However, at least one inherent problem of comparing algorithmic reasonability to that of humans would still remain unresolved: the fact that the two types of decision-makers are characterized by different strengths as well as weaknesses.[162] For several types of causes of damage algorithms may be infinitely preferable to humans (for example, failure to collect relevant information, failure to tag it with the right numerical value, and failure to precisely execute a certain movement). In other instances, however, algorithms' superiority may be less dramatic. Applying a "one size fits all" gap required between algorithms' abilities and those of humans would under-incentivize improvement in those areas where algorithms are already performing better than the dictated gap: requiring them to be only 10% better than humans, when, in practice, they can be 200% better would not create a sufficient incentive for improvement. At the same time, such comparison might cause a chilling effect with respect to areas where algorithms' advantage is less significant. The same is true with respect to algorithms' inherent weaknesses—perhaps even more so. Certain damaging decisions might be driven by algorithms' intrinsic disadvantages compared to humans, for example when tacit knowledge or human compassion is required. Holding algorithms to human

[160] Froomkin et. al., *supra* note 2, at 51–56.

[161] In other words, thinking algorithms (or, more accurately, their manufacturers) will be granted immunity from legal liability so long as they can show that on average their error rates are lower (or significantly lower) than those of humans making similar decisions.

[162] *See supra* notes 86–110 and accompanying text.

standards in these cases will indeed incentivize their improvement, but may also discourage manufacturers from developing beneficial technologies, depending on various factors (including the feasibility of enhancing the algorithm's ability and its cost).

Lastly, algorithms' and humans' decision-making processes don't just differ with respect to their respective levels of technical abilities. While humans typically maximize their own personal utility function when making a decision, algorithms optimize their decisions based on other types of programmed functions related to general welfare or profit maximization of their manufacturers. Moreover, when attempting to follow the two inherently different functions of utility, humans are affected by stress, emotion, and instinct—all of which algorithms are immune to. While human instinct might very well warrant a "reasonable" finding in a case of a human driver hitting a pedestrian to save herself, the same outcome may not necessarily be legally or normatively desired where algorithms are concerned.[163] In that sense, comparing the reasonableness of an algorithm to that of humans would not be comparing like with like, and might also yield unwarranted or absurd results.

How, then, might we take the flexible, dynamic assessment of reasonableness that is, from several points of view, very well-suited to thinking algorithms and apply it to them despite the lack of equivalent algorithms against which to compare their reasonableness?

Part IV: Proposed Model for Assessing Algorithmic Reasonableness

The discussion so far has shown us some of the primary shortcomings of basing an algorithmic standard of reasonableness on the reasonableness of humans. The discussion has also revealed that applying a normative analysis of reasonableness to thinking algorithms will likely do a poor job at advancing the desired goals, given that the algorithms themselves will find it difficult to understand the reasonableness expected of them and act accordingly. To overcome these shortcomings, the proposed model is based on a blend of human and algorithmic reasonableness.

According to this model, determining whether a thinking algorithm has acted reasonably would follow a two-pronged approach. First, a court would consider the reasonableness of a human engaging in similar decision-

[163] The defense of necessity, for example, may be invoked by a human driver but will likely not be applicable to driverless vehicles: "[c]hoices to stay on course or to swerve could be justified by invoking the state of necessity. . . . The described scenario is a particular case of a state of necessity in which the perpetrator (the manufacturer/programmer) does not directly face danger to his life but rather intervenes to save one or more persons, causing harm to someone else involved in the same dangerous situation. When the perpetrator is not directly in danger and does not act out of self-preservation (or kin-preservation), the applicability of the general state-of-necessity defence is controversial." Contissa et al., *supra* note 130, at 4–5.

making. For example, the reasonableness of an algorithm administrating penicillin to an allergic patient would be compared to that of a human physician. However, this analysis would ignore the inherent differences between humans and algorithms, and this prong alone would not allow courts to distinguish between cases where the algorithm could have been held to higher standard than its human equivalent or where the algorithm is at an inherent disadvantage and could not have performed as well as its human equivalent.

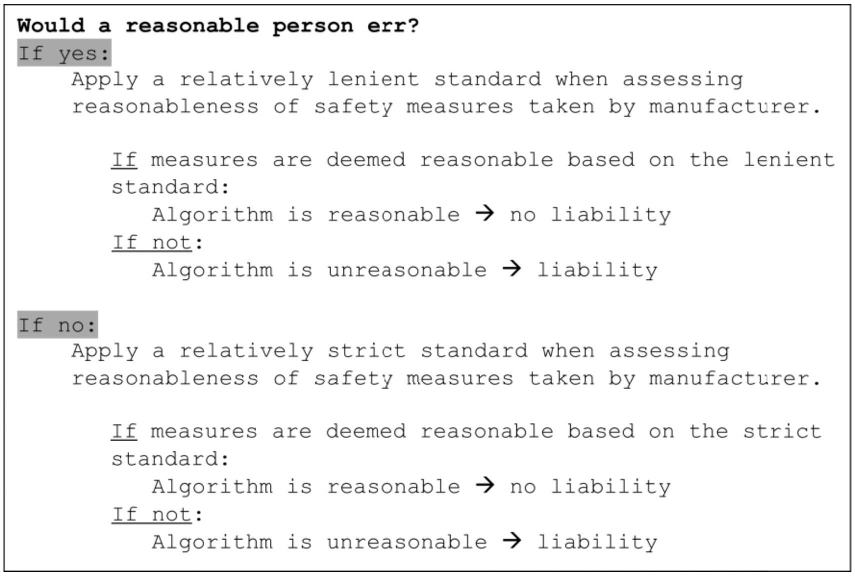
To solve this problem and ensure that the normative considerations advanced by the courts actually affect potential wrongdoers, the second prong of the model will rely on the reasonableness of the programmers (or manufacturers).[164] Having decided whether the algorithm itself was reasonable based on the reasonableness of a person, the second prong of analysis would look at the means undertaken by the manufacturers to minimize damage and determine whether these precautions were reasonable. When looking to the manufacturer of a medical algorithm that administers medications, for example, we would check what safety measures were in place to minimize the risk that the algorithm might cause damage to patients with allergies. The exact type and scope of what safety measures would be sufficient to meet the standard of reasonableness will likely be case-sensitive and in any event are a topic for a separate paper.[165] What is important with respect to the proposed model is that the two complementary reasonableness assessments operate on a sliding scale: the less reasonable an algorithm's behavior is compared to a human, the more we will demand of the manufacturer in order for the decision to be deemed reasonable.

[164] An important question this paper is not attempting to address, is the manner of allocating liability among various entities related to the manufacturing or designing of the system. In the context of cars, for example, the manufacturer of the car itself may be a completely different entity than the one designing the thinking algorithms installed in it, rendering it autonomous. A separate entity may install "patches" altering the technology in a manner that might lead to damaging decisions. The users of the car and of the road are naturally additional types of players to be considered when determining who is liable. *See generally*, Determann & Perens, *supra* note 142. For the purposes of the paper, however, the entities involved in the manufacturing of the thinking algorithms are perceived as one.

[165] They could, for example, include some means of constantly monitoring the system in order to signal when anything went wrong; of installing "emergency brakes," allowing manufacturers or users to shut down the system in certain circumstances; or of providing ongoing support and patching services. *See* Omri Rachum-Twaig, *Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots,* 2020 U. ILL. J.L. TECH. & POL'Y 1141 (2020). Other mechanisms for assuring safety could include developing more than one layer of decision-making for each system, such that decisions are independently reached by each layer, and those that are found not to be unanimous would then undergo additional scrutiny. *Cf.* Niva Elkin-Koren & Maayan Perel, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, 24 LEWIS & CLARK L. REV 857 (2020) (discussion of the multilayer, independent deciders in the context of AI generated content moderation. Alternatively, whether a system was subject to external auditing by a private firm or governmental entity or received quality certification could be factors in the analysis).

A "reasonable algorithm" analysis would therefore include both algorithmic reasonableness (focusing on the specific decision and specific damage caused by the algorithm, as is the case with the "reasonable person" analysis) and also manufacturer's reasonableness (focusing on the safety measures implemented by the manufacturer to avoid risk in general).[166] To be found reasonable, however, it is not necessary that both assessments result in a positive answer. Rather, the reasonableness outcome of the first prong will determine the level of reasonableness required in the second prong for the reasonable algorithm test to result in reasonableness. In other words, when the reasonable person analysis applied to the specific decision of the algorithm results in algorithmic reasonableness (meaning that a human would have also caused the same damage), there will consequently be more leeway in our reasonableness analysis of the second prong. If on the other hand the algorithm caused damage when a human would not have, then, to be considered reasonable and escape liability, the manufacturers will be held to a much higher standard of reasonableness with respect to the safety means they needed to have undertaken. Figure 1 illustrates the reasonable algorithm analysis.

[166] Similar to the analysis used in products liability cases concerning design defects. A design defect occurs when there is a flaw in the design of a product (and not in its manufacturing), which gives rise to a products liability claim. The majority of states apply a risk utility test on design defects, where liability is found when the foreseeable risks associated with the product could have been minimized by using a feasible safer alternative. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2(b) (AM. LAW INST. 1998).



```
Would a reasonable person err?
If yes:
    Apply a relatively lenient standard when assessing
    reasonableness of safety measures taken by manufacturer.

       If measures are deemed reasonable based on the lenient
       standard:
          Algorithm is reasonable → no liability
       If not:
          Algorithm is unreasonable → liability

If no:
    Apply a relatively strict standard when assessing
    reasonableness of safety measures taken by manufacturer.

       If measures are deemed reasonable based on the strict
       standard:
          Algorithm is reasonable → no liability
       If not:
          Algorithm is unreasonable → liability
```

Figure 1: Model for Assessing the Reasonableness of Thinking Algorithms

Linking the reasonableness of the algorithm itself, measured against the reasonableness of humans, and that of the safety measures implemented by the manufacturer serves several purposes. First, it enables a practical positive comparison of reasonableness that is based on the reasonable person assessment that is familiar to courts. Adding the reasonableness of the means taken by the manufacturers renders the normative approach meaningful as well, as the assessment would then be directed to the party capable of understanding reasonableness and modifying its behavior accordingly. Second, with respect to the sliding scale operation of the two prongs, this balance avoids both demanding too little and too much of thinking algorithms, fulfilling the underlying aims of tort law while avoiding a chilling effect.

Let us return to our first meningitis example, and further assume that, on average, a medical algorithm delivers safer results than a human physician. What would the two-pronged assessment look like if the algorithm prescribed penicillin to the allergic patient, and how does the approach reach a desired balance between promoting safety and avoiding a dampening effect on technology?

In one scenario, the first prong of the reasonableness assessment results in a finding of reasonableness—that is, a case in which the algorithm's decision to administer penicillin would have also been reached by a human doctor. Generally speaking, such scenarios are of less concern with respect to the desire to promote safety: not only is the algorithm superior to humans on average, but for this specific case, technology is no less safe than the human alternative. Because the minimum safety threshold of the human-level is upheld, society would have an interest in further developing the algorithm to be even safer than humans.

Under such circumstances, we want to be cautious about exposing manufacturers to liability. If the algorithm's mistake resulted from an inherent weakness that is not currently possible to efficiently improve, we would not want to hold the algorithm unreasonable and the manufacturer liable, because the algorithm in such a scenario did not cause damage that a human would not have, and has no efficient means of improvement. If, however, the algorithm could efficiently be made safer than humans and avoid the mistake a human would have made, then we would want to hold the manufacturer liable for not undertaking the measures necessary to do so. Under such scenarios, the fact that, on average, the algorithm is just as safe as a human decision-maker would not suffice. Adding the second prong—the reasonableness of the safety means taken by the manufacturer—will assist in distinguishing the former cases from the latter. If the second prong's lenient reasonableness analysis shows that the manufacturers applied reasonable safety means, this indicates that this is likely not a scenario where the algorithm could be easily and efficiently made safer. If, however, the second prong's lenient assessment reveals that no adequate safety measures were undertaken, this indicates that the algorithm could have efficiently been rendered safer, and thus the initial finding of reasonableness on the part of the algorithm itself compared to human tortfeasors would not suffice.

A second and more troubling scenario arises when the positive comparison to human reasonableness results in unreasonableness, meaning that the algorithm caused damage that a person would not have. Generally speaking, in such cases we would apply greater emphasis on promoting safety because the technology is not as safe as a human equivalent would be. The same level of safety measure taken by the manufacturer that sufficed for a reasonableness finding in the former scenario, therefore, would not suffice when the algorithm itself was unreasonable. At the same time, given our assumption that algorithmic decision-makers tend to be safer than human ones, we would still have an interest in promoting said technology and be concerned about a potential chilling effect if manufacturers are held to standards they cannot meet. If a manufacturer satisfies the elevated bar of reasonableness of the safety measures it took, it will not be found liable. On the other hand, if the manufacturer cannot show it took all available safety precautions (or some other elevated standard of "reasonableness" of the safety measures that a court would apply), we would be less concerned about a chilling effect since the algorithm is less safe than a human and there are efficient means of improving it. Liability in this case would be appropriate, as it would not put an excessive burden on a manufacturer and would promote safety.

The proposed two-prong analysis of algorithmic reasonableness warrants further development and discussion.[167] It does, however, set a practical method for applying a "reasonable algorithm" analysis to algorithmic tortfeasors in a manner that allows for a fine-tuned balance between the goals of tort law and technological advancement.

Conclusion

Algorithms possess unique traits. The flexible, adaptive, and neutral standard of the reasonableness analysis, be it the positive or the normative one, is compatible with these traits. The reasonableness analysis may be applied to various types of algorithmic tortfeasors, regardless of their specific characteristics and abilities, and may be applicable even when the reasons for their actions remain unknown.

Unlike humans, algorithms lack the common sense that enables them to understand, balance, and follow vague standards rather than concrete rules. Applying a straight normative reasonableness approach to algorithms would therefore be unhelpful. Moreover, applying a straight positive standard to algorithms would also be problematic, since there is not currently a sufficient corpus of other algorithms with which to compare.

The two-pronged model I propose combines the reasonableness of the algorithm compared to a human with the reasonableness of the safety measures taken by the algorithm manufacturer. By applying both measures, courts and other stakeholders could easily assess the reasonableness of the algorithm itself, while at the same time reaching the appropriate balance between the competing policy desires to promote safety and the need to avoid chilling effects on technological development.

[167] Addressing questions such as: when safety measures would meet both levels of reasonableness required in the second prong depending on the outcome of the first, whether it could be based on the current test for products liability and a design defect, how to apply said analysis when there is also fault by the users or by other parties, etc.