Harvard Journal of Law & Technology
Volume 31, Number 2 Spring 2018

THE ARTIFICIAL INTELLIGENCE BLACK BOX AND THE FAILURE OF INTENT AND CAUSATION

Yavar Bathaee[*]

TABLE OF CONTENTS

[*] Associate, Sullivan & Cromwell LLP. J.D., Fordham University School of Law; B.S., Computer Science and Engineering, University of California Davis. For Audrey and Elliot. With immense gratitude to my wife Jacqueline for her support and help with this article. Opinions expressed in this article are my own.

I. INTRODUCTION

There is a heated debate raging about the future of artificial intelligence, particularly its regulation,[1] but little attention is being paid to whether current legal doctrines can properly apply to AI.[2] Commentators, for example, are asking important questions about potential risks, such as whether AI will pose an existential threat to humanity,[3] or whether AI technology will be concentrated in the hands of the few.[4] Many have forcefully called for regulation before these risks manifest, but there is a more pressing problem looming on the horizon: the law

[1] There has been a forceful call to regulate AI. For example, five of the largest developers of AI technology plan to form a consortium to devise objective ethical standards for the development and use of AI. John Markoff, *How Tech Giants Are Devising Real Ethics for Artificial Intelligence*, N.Y. TIMES (Sept. 1, 2016), https://www.nytimes.com/2016/09/02/technology/artificial-intelligence-ethics.html (last visited May 5, 2018). Likewise, the One Hundred Year Study on Artificial Intelligence's Study Panel released a report identifying several regulatory problems concerning, *inter alia*, privacy, innovation policy, civil and criminal liability, and labor. *See* STANFORD UNIV., ARTIFICIAL INTELLIGENCE AND LIFE IN 2030: ONE HUNDRED YEAR STUDY ON ARTIFICIAL INTELLIGENCE 46–47 (Sept. 2016) [hereinafter ONE HUNDRED YEAR STUDY].

[2] The report of the One Hundred Year Study on Artificial Intelligence, for example, acknowledges that AI may cause problems with civil and criminal liability doctrines, such as intent, but notes that a detailed treatment is beyond the scope of the report. One Hundred Year Study, *supra* note 1, at 45–46. Although other commentators have identified problematic interactions between current legal doctrines and AI or machine learning, I am aware of no attempt to address the problems in detail or to propose a broader solution. *See, e.g.*, Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1193 (2017) (discussing the difficulty in establishing discriminatory intent when a federal agency uses AI to guide its decisions).

[3] *See* James Vincent, *Elon Musk Says We Need to Regulate AI Before It Becomes a Danger to Humanity*, VERGE (July 17, 2017, 4:43 AM), https://www.theverge.com/2017/7/17/15980954/elon-musk-ai-regulation-existential-threat [https://perma.cc/EY2Q-2R2P].

[4] The battle for control over AI focuses largely on capturing the top AI talent. At present, large companies such as Amazon, Google, Microsoft and IBM "account for 40% of open AI positions." Stacy Jones, *Automation Jobs Will Put 10,000 Humans to Work, Study Says*, FORTUNE, (May 1, 2017), http://fortune.com/2017/05/01/automation-jobs-will-put-10000-humans-to-work-study-says/ [https://perma.cc/M3YD-WSSE]. AI researchers, who are regarded "among the most prized talent in the modern tech world," are aggressively sought out by large companies, which also aggressively purchase AI startups in their incipiency to ensure primacy over budding technology and talent. *See* Cade Metz, *The Battle for Top AI Talent Only Gets Tougher from Here*, WIRED (Mar. 23, 2017, 11:00 AM), https://www.wired.com/2017/03/intel-just-jumped-fierce-competition-ai-talent/ [https://perma.cc/3LNM-APEV].

is built on legal doctrines that are focused on human conduct,[5] which when applied to AI, may not function. Notably, the doctrines that pose the greatest risk of failing are two of the most ubiquitous in American law — intent and causation.

The reason intent and causation may fail to function is because of the nature of the machine-learning algorithms on which modern AI are commonly built.[6] These algorithms are capable of learning from massive amounts of data, and once that data is internalized, they are capable of making decisions experientially or intuitively like humans.[7] This means that for the first time, computers are no longer merely executing detailed pre-written instructions but are capable of arriving at dynamic solutions to problems based on patterns in data that humans may not even be able to perceive.[8] This new approach comes at a price, however, as many of these algorithms can be black boxes, even to their creators.[9]

It may be impossible to tell how an AI that has internalized massive amounts of data is making its decisions.[10] For example, AI that relies on machine-learning algorithms, such as deep neural networks, can be as difficult to understand as the human brain.[11] There is no straightforward way to map out the decision-making process of these

[5] As Justice Oliver Wendell Holmes, Jr. observed, "[t]he life of the law has not been logic: it has been experience." OLIVER WENDELL HOLMES, JR., THE COMMON LAW 1 (1881). As this Article claims, the law is presently at an inflection point, as never before has the law encountered thinking machines. The experience of the law is limited to the criminal, business, and artistic endeavors of humans, powered only by their own actions and the actions of others they control.

[6] As will be discussed *infra* in Part II of this Article, machine-learning algorithms are computer programs that are capable of learning from data. *See infra* Section II.A.

[7] *See* TOSHINORI MUNAKATA, FUNDAMENTALS OF THE NEW ARTIFICIAL INTELLIGENCE 1–2 (2d ed. 2008) (listing abilities such as "inference based on knowledge, reasoning with uncertain or incomplete information, various forms of perception and learning, and applications to problems such as control, prediction, classification, and optimization").

[8] Since the 1940s, artificial intelligence has evolved from its roots in programs that merely executed instructions specified by the programmer into machine-learning algorithms that "can learn, adapt to changes in a problem's environment, establish patterns in situations where rules are not known, and deal with fuzzy or incomplete information." MICHAEL NEGNEVITSKY, ARTIFICIAL INTELLIGENCE 14 (2d ed. 2005). These modern AI can arrive at solutions or solve problems without the need for a human programmer to specify each instruction needed to reach the given solution. Thus, AI may solve a particular problem or reach a solution that its programmer never anticipated or even considered.

[9] This is the central claim of Part II of this Article, which demonstrates how machine-learning algorithms may be black boxes, even to their creators and users. *See infra* Section II.B. For an excellent description of the problem and how researchers are struggling to ease transparency problems with AI, see Davide Castelvecchi, *Can We Open the Black Box of AI?*, NATURE (Oct. 5, 2016) (characterizing "opening up the black box" as the "equivalent of neuroscience to understand the networks inside" the brain).

[10] *See id.*

[11] *See id.* (quoting a machine-learning researcher stating that "even though we make these networks, we are no closer to understanding them than we are a human brain").

complex networks of artificial neurons.[12] Other machine-learning algorithms are capable of finding geometric patterns in higher-dimensional space,[13] which humans cannot visualize.[14] Put simply, this means that it may not be possible to truly understand how a trained AI program is arriving at its decisions or predictions.

The implications of this inability to understand the decision-making process of AI are profound for intent and causation tests, which rely on evidence of human behavior to satisfy them. These tests rely on the ability to find facts as to what is foreseeable,[15] what is causally related,[16] what is planned or expected,[17] and even what a person is thinking or knows.[18] Humans can be interviewed or cross-examined; they leave behind trails of evidence such as e-mails, letters, and memos that help answer questions of intent and causation;[19] and we can draw on heuristics to help understand and interpret their con-

[12] *Id.* ("But this form of learning is also why information is so diffuse in the network: just as in the brain, memory is encoded in the strength of multiple connections, rather than stored at specific locations, as in a conventional database.").

[13] By space I refer here to a mathematical space, such as the notion of a vector space, where every element of the space is represented by a list of numbers and there are certain operations defined, such as addition, in the space. *See generally* Vector Space, WOLFRAM ALPHA, http://mathworld.wolfram.com/VectorSpace.html [https://perma.cc/DC6F-DHLS].

[14] A two-dimensional space can be visualized as a series of points or lines with two coordinates identifying the location on a graph. To represent a third dimension, one would add a third axis to visualize vectors or coordinates in three-dimensional space. While four dimensions can be visualized by adding a time dimension, five dimensions and higher are impossible to visualize. This is discussed further as part of the discussion of dimensionality. *See infra* Section II.C.2.

[15] *See, e.g.*, Owens v. Republic of Sudan, 864 F.3d 751, 794 (D.C. Cir. 2017) (stating that to establish proximate cause, plaintiff's injury must have been "reasonably foreseeable or anticipated as a natural consequence of the defendant's conduct" (citation omitted)); Palsgraf v. Long Island R.R., 162 N.E. 99 (N.Y. 1928) (marking the beginning of the modern formulations of proximate cause).

[16] For example, as discussed further *infra* in Section IV.B.2, Article III standing requires that the alleged injury be fairly traceable to the allegedly unlawful conduct at issue.

[17] As discussed further *infra* in Section III.B, certain intent tests require that the effects of the conduct (such as market manipulation in the securities and commodities laws) to be intentional. *See, e.g.*, Braman v. The CME Group, Inc., 149 F. Supp. 3d 874, 889–90 (N.D. Ill. 2015) ("A manipulation claim requires a showing of specific intent, that is, a showing that 'the accused acted (or failed to act) with the purpose or conscious object' of influencing prices." (quoting *In re* Soybean Futures Litig. 892 F.Supp. 1025, 1058–59 (N.D. Ill. 1995))).

[18] The reliance test in the securities fraud context is a classic example of such a test. A plaintiff must have believed the alleged misrepresentation in order to prevail. *See, e.g.*, Basic Inc. v. Levinson, 485 U.S. 224, 243 (1988).

[19] Indeed, e-mails, documents and other such evidence often serve as circumstantial evidence of intent. *See, e.g.*, Koch v. SEC, 793 F.3d 147, 155 (D.C. Cir. 2015) (noting that e-mails and recorded phone conversations provided circumstantial evidence of defendant's intent); United States v. Patel, 485 F. App'x 702, 708 (5th Cir. 2012) ("Intent to defraud is typically proven with circumstantial evidence and inferences" (citing United States v. Ismoila, 100 F.3d 380, 387 (5th Cir. 1996))); ACP, Inc. v. Skypatrol, L.L.C., No. 13-cv-01572-PJH, 2017 U.S. Dist. LEXIS 77505, at *33 (N.D. Cal. May 22, 2017) (noting that e-mails could provide sufficient circumstantial evidence of fraudulent intent); United States v. Zodhiates, 235 F. Supp. 3d 439, 447 (W.D.N.Y. 2017) (noting that e-mails could be used by jury to infer knowledge and intent).

duct.[20] If an AI program is a black box, it will make predictions and decisions as humans do, but without being able to communicate its reasons for doing so. The AI's thought process may be based on patterns that we as humans cannot perceive, which means understanding the AI may be akin to understanding another highly intelligent species — one with entirely different senses and powers of perception. This also means that little can be inferred about the intent or conduct of the humans that created or deployed the AI, since even they may not be able to foresee what solutions the AI will reach or what decisions it will make

Two possible (but ultimately poor) solutions to these problems are (1) to regulate the degree of transparency that AI must exhibit, or (2) to impose strict liability for harm inflicted by AI. Both solutions are problematic, incomplete, and likely to be ineffective levers for the regulation of AI. For example, a granular regulation scheme of AI transparency will likely bring new startups in AI technology to a halt, as new entrants would have to bear the high costs of regulatory compliance and wrestle with regulatory constraints on new designs.[21] Moreover, there is no guarantee certain AI programs and machine-learning algorithms can be developed with increased transparency. The future may in fact bring even more complexity and therefore less transparency in AI, turning the transparency regulation into a func-

[20] For example, a court may use heuristics such as consciousness of guilt to assist with the intent inquiry. *See, e.g.*, United States v. Hayden, 85 F.3d 153, 159 (4th Cir. 1996) ("Evidence of witness intimidation is admissible to prove consciousness of guilt and criminal intent under Rule 404(b), if the evidence (1) is related to the offense charged and (2) is reliable." (citations omitted)). Rules of evidence frequently include such heuristics — for example, the peaceful character of a victim is admissible in a murder case to rebut the notion that that the victim was the first aggressor, FED. R. EVID. 404(a)(2)(C), and evidence of a past crime can be used to infer a defendant's motives and intent, *id.* at 404(b)(2). Other heuristics include the notion of a reasonable man — that is, an idealization of the risks and conduct that one would expect writ large. *See* RESTATEMENT (SECOND) OF TORTS § 283 cmt. b (AM. LAW INST. 1965) (defining a reasonable person as "a person exercising those qualities of attention, knowledge, intelligence, and judgment which society requires of its members for the protection of their own interests and the interests of others."). These heuristics contain implicit assumptions about how and why people typically behave or ideally should behave and are often used to control the conclusions that can be inferred from the evidence.

[21] Banking regulations illustrate the effect of a complex regulatory scheme. As the Federal Reserve's website notes, "[s]tarting a bank involves a long organization process that could take a year or more, and permission from at least two regulatory authorities." *How Can I Start a Bank?*, BOARD OF GOVERNORS OF THE FED. RES. SYS. (Aug. 2, 2013), https://www.federalreserve.gov/faqs/banking_12779.htm [https://perma.cc/HNR5-EQL7]. After obtaining approval for deposit insurance from the Federal Deposit Insurance Corporation (FDIC), the new entrant must then meet the "capital adequacy guidelines of their primary federal regulator" and "demonstrate that it will have enough capital to support its risk profile, operations, and future growth even in the event of unexpected losses." *Id.* Technology startups, however, are infamous for their scrappiness, with notable examples beginning in garages. *See* Drew Hendricks*, 6 $25 Billion Companies that Started in a Garage*, INC. (Jul. 24, 2014), https://www.inc.com/drew-hendricks/6-25-billion-companies-that-started-in-a-garage.html [https://perma.cc/CU5B-CEX2].

tional prohibition on certain classes of AI that inherently lack transparency.[22] Strict liability is also a poor solution for the problem because if one cannot foresee the solutions an AI may reach or the effects it may have, one also cannot engage in conduct that strict liability is designed to incentivize, such as taking necessary precautions or calibrating the level of financial risk one is willing to tolerate.[23]

A better solution is to modify intent and causation tests with a sliding scale based on the level of AI transparency and human supervision. Specifically, when AI merely serves as part of a human-driven decision-making process, current notions of intent and causation should, to some extent, continue to function appropriately, but when AI behaves autonomously, liability should turn on the degree of the AI's transparency, the constraints its creators or users placed on it, and the vigilance used to monitor its conduct.

This Article proceeds in five parts. Part II provides a framework for understanding what this Article calls the Black Box Problem. Specifically, Part II describes two machine-learning algorithms that are widely used in AI systems — deep neural networks and support vector machines — and demonstrates why these algorithms may cause AI built on them to be black boxes to humans. Deep networks of artificial neurons distribute information and decision-making across thousands of neurons, creating a complexity that may be as impenetrable as that of the human brain.[24] So-called "shallow" algorithms such as support vector machines operate by finding geometric patterns in higher-dimensional space that humans cannot visualize. This dimensionality renders these models similarly opaque to humans. Part II attempts to provide a clear definition of the constraints imposed by these problems and posits for the purposes of the analysis in this Article a weak and strong form of these constraints.

Part III discusses three categories of intent tests and demonstrates that when an AI is a black box, these intent tests can rarely be satisfied. Effect Intent tests, such as those that appear as part of market manipulation claims in securities and commodities law,[25] assess

[22] As this Article argues *infra* in Section II.C, the black box nature of AI arises from the complexity of distributed elements, such as in deep neural networks, and from the inability of humans to visualize higher-dimensional patterns. As machine-learning algorithms become more sophisticated, networks are likely to become more complex and the number of dimensions will grow with the amount of data that machine-learning models have the capacity to balance and optimize at once. In other words, there is little reason to believe that the problem can be solved simply by regulatory fiat.

[23] A classic example of the sort of conduct strict liability incentivizes to allow the appropriate calibration of risk is the purchase of insurance. *See* Todd v. Societe BIC, S.A., 9 F.3d 1216, 1219 (7th Cir. 1993) ("Some products are dangerous even when properly designed, and it is both easier and cheaper for consumers to obtain their own insurance against these risks than to supply compensation case-by-case through the judicial system.").

[24] *See supra* notes 9–12.

[25] *See supra* note 17.

whether a person intended a prohibited outcome, but because the operator of an AI may not know ex ante what decisions or predictions the AI will make, it may be impossible to establish such intent. Basis Intent tests such as those that appear in constitutional,[26] securities,[27] and antitrust[28] law, scrutinize the justifications or reasons for a person's conduct, but if a black-box AI's reasoning is opaque, then such tests will also be impossible to satisfy. Finally, Gatekeeping Intent tests such as the Discriminatory Intent test used in Equal Protection jurisprudence,[29] which limit the scope of a law or cause of action by requiring a showing of intent upfront, may entirely prevent certain claims or legal challenges from being raised in the first place when AI is involved.

Part IV examines two categories of causation tests and argues that these tests also fail when black-box AI is involved. Conduct-Regulating tests attempt to determine the scope of liability for broad claims such as negligence and are designed to encourage or discourage conduct ex ante. Proximate cause is the most prominent example of such a test, which requires an inquiry into what was reasonably foreseeable to the creators or users of AI.[30] When the AI is a black box, foreseeability cannot be proven because the creators or users of the AI will not know ex ante what effects the AI will have. The sec-

[26] Basis Intent tests are at the center of constitutional law. The rational basis test, the intermediate and strict standards of scrutiny, and the undue burden test are some of the most prominent examples, with each test requiring some justification for challenged government conduct. *See, e.g.*, Lawrence v. Texas, 539 U.S. 558, 579 (2003) (O'Connor, J., concurring) ("Under our rational basis standard of review, 'legislation is presumed to be valid and will be sustained if the classification drawn by the statute is rationally related to a legitimate state interest.'" (quoting Cleburne v. Cleburne Living Center, 473 U.S. 432, 440 (1985))).

[27] For example, claims under the Securities Act of 1933 for omissions relating to opinion statements require an examination of the basis of the challenged opinion. *See* Omnicare, Inc. v. Laborers Dist. Council Constr. Indus. Pension Fund, 135 S. Ct. 1318, 1326 (2015). *Omnicare* is discussed in more detail *infra* in Section III.C.

[28] For example, in response to a claim under Section 2 of the Sherman Act for a refusal to deal by a monopolist, the defendant may make a defensive showing that it had valid business or economic justifications for its refusal. *See* Morris Commc'ns. Corp. v. PGA Tour, Inc., 364 F.3d 1288, 1295 (11th Cir. 2004) ("[R]efusal to deal that is designed to protect or further the legitimate business purposes of a defendant does not violate the antitrust laws, even if that refusal injures competition." (citing Aspen Skiing Co. v. Aspen Highlands Skiing Corp., 472 U.S. 585, 604 (1985)). This test from the antitrust laws will be discussed in more detail *infra* in Section III.B.

[29] In addition to showing a disparate impact, a plaintiff challenging a law designed to serve neutral ends under the Equal Protection Clause of the U.S. Constitution must also prove discriminatory intent by the government. *See* Washington v. Davis, 426 U.S. 229 (1976). The discriminatory intent test will be discussed further *infra* in Section III.D.

[30] *See* DAN B. DOBBS, THE LAW OF TORTS 447 (2000) ("The defendant must have been reasonably able to foresee the kind of harm that was actually suffered by the plaintiff (or in some cases to foresee that the harm might come about through intervention of others)."). As Dobbs observes, the term "foreseeability is itself a kind of shorthand" that stands for the proposition that "the harm must be the kind that [the defendant] should have avoided by acting more carefully." *Id.* The proximate cause test therefore also polices the legal scope of reasonable conduct. *See id.*

ond category of causation tests consists of what this Article refers to as Conduct-Nexus tests. These tests, which include, for example, reliance tests and the causation element of Article III standing,[31] attempt to address whether there is some minimum connection between the unlawful conduct and the injury suffered. Because the reasons why AI may have made a particular decision or prediction may be opaque to analysis, it may be impossible to establish the threshold causation required to satisfy these tests. A plaintiff challenging AI used by a federal agency, for example, may not be able to prove that the AI improperly considered or weighed information and was therefore responsible for his injury.[32]

Part V examines and rejects two possible solutions to the problems with intent and causation. First, this part examines the option of imposing minimum transparency standards on AI. This option assumes that transparency can be improved for the powerful machine-learning algorithms currently being used as part of AI, but such an assumption may be flawed. For example, given that the Black Box Problem arises from complexity in artificial neural networks, there is only reason to believe that complexity is likely to become greater as AI advances. Likewise, as AI becomes capable of handling larger amounts of information, the dimensionality problem is also likely to become more acute. Therefore, such transparency standards may stifle AI innovation by prohibiting major categories of AI. Additionally, such standards may increase market concentration as a result of regulatory compliance costs and require regulators to make design decisions they are likely unequipped to make. Second, Part V rejects strict liability as a potential solution because the unpredictability of AI eliminates the positive effects of strict liability. For instance, if the creator or user of AI cannot predict the effects of the AI ex ante, he cannot take precautions for the injury inflicted. Strict liability may only deter smaller firms from developing AI because they would risk outsized liability should the AI cause any injury. This would favor established and well-capitalized participants in the field and erect significant barriers to entry and innovation.

[31] *See* Bank of Am. Corp. v. City of Miami, 137 S. Ct. 1296, 1302 (2017) ("To satisfy the Constitution's restriction of this Court's jurisdiction . . . a plaintiff must demonstrate constitutional standing. To do so, the plaintiff must show an 'injury in fact' that is 'fairly traceable' to the defendant's conduct and 'that is likely to be redressed by a favorable judicial decision.'" (quoting Spokeo, Inc. v. Robins, 136 S. Ct. 1540, 1547 (2016)). Reliance tests also serve a similar function. *See* Basic Inc. v. Levinson, 485 U.S. 224, 243 (1988).

[32] The difficulty in tying federal administrative action to injury in fact is not new. *See, e.g.*, Simon v. E. Ky. Welfare Rights Org., 426 U.S. 26 (1976) (holding that the plaintiff could not show that the IRS's regulations providing tax-exempt status were the cause of the hospitals' refusal to serve indigent patients absent medical emergency). The Simon case and the fairly traceable standard it established is discussed in more detail *infra* Section IV.B.2.

Part VI proposes an approach that takes into account the degree of the AI's transparency as well as the extent to which the AI is supervised by humans. If AI is given complete autonomy to, for example, trade in the securities markets, the threshold for liability should be lower (i.e., it should be easier to meet the burden of proof for a claim) and evidence of poor constraints, design, and limitations on data access should weigh more heavily in favor of liability. Where an algorithm is designed to merely assist a human being in making a decision or performing a task, the human's intent or the foreseeability of the effects of the AI's decision should weigh more heavily and the constraints on the algorithm, the nature of the design, or the data available to the AI algorithm should play a lesser role in the question of liability.

II. AI, MACHINE-LEARNING ALGORITHMS, AND THE CAUSES OF THE BLACK BOX PROBLEM

This section attempts to demonstrate how transparency problems arise directly from the nature of certain machine-learning algorithms that are widely used in AI. Specifically, this section discusses two commonly used machine-learning algorithms to demonstrate why AI that relies on them may be a black box to humans. The first algorithm discussed in this section is the deep neural network, which often involves the use of thousands of artificial neurons to learn from and process data. The complexity of these countless neurons and their interconnections makes it difficult, if not impossible, to determine precisely how decisions or predictions are being made. The second algorithm, the support vector machine, is used to illustrate how shallow (i.e. less complex) algorithms can also create a black-box problem because they process and optimize numerous variables at once by finding geometric patterns in higher-dimensional, mathematically-defined spaces. This high "dimensionality" prevents humans from visualizing how the AI relying on the support vector machine is making its decisions or from predicting how the AI will treat a new data. Finally, this section more precisely defines the constraints imposed by what I refer to throughout this Article as the Black Box Problem. This Article further subdivides the Black Box Problem into a strong and a weak form to aid with the rest of the Article's analysis. As explained further throughout this Article, there may be different implications for intent and causation tests depending on whether AI is a weak or strong black box.

A. What Is Artificial Intelligence?

Artificial intelligence refers to a class of computer programs designed to solve problems requiring inferential reasoning, decision-making based on incomplete or uncertain information, classification, optimization, and perception.[33] AI programs encompass a broad range of computer programs that exhibit varying degrees of autonomy, intelligence, and dynamic ability to solve problems. On the most inflexible end of the spectrum are AI that make decisions based on preprogrammed rules from which they make inferences or evaluate options.[34] For example, a chess program that evaluates every possible move and then selects the best move according to a scoring formula would fall within this category. On the most flexible end are modern AI programs that are based on machine-learning algorithms that can learn from data. Such AI would, in contrast to the rule-based AI, examine countless other chess games and dynamically find patterns that it then uses to make moves — it would come up with its own scoring formula.[35] For this sort of AI, there are no pre-programmed rules about how to solve the problem at hand, but rather only rules about how to learn from data.[36]

To further illustrate the difference between AI that learns from data and AI that simply evaluates rules or possible outcomes, consider the following hypothetical computer program designed to choose whether to admit students to a university. The program is tasked with reviewing each applicant's file and making an admission decision based on a student's SAT score, grade-point average and a numerical score assigned to the difficulty of his or her high school's curriculum. The first computer program applies hard rules — multiply the SAT

[33] *See supra* note 7.

[34] Early AI was focused on solving problems with static rules, which were in most cases mathematically defined. *See* IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 2 (2016) ("Several artificial intelligence projects have sought to hard-code knowledge about the world in formal languages. A computer can reason automatically about statements in these formal languages using logical inference rules. This is known as the knowledge base approach to artificial intelligence." (emphasis omitted)). That approach was largely unsuccessful. *Id.* Today, AI is more focused on solving problems the way humans do, by using intuition — problems such as image recognition, identification of patterns in large amounts of data, or language and voice processing. *Id.* at 1–2. These are tasks that may be easy for humans to perform, but hard to describe. *Id.*

[35] For a comparison of how early AI chess programs evaluated possible moves versus the more intuitive, experience-based method used by modern AI chess programs, see Dave Gershgorn, *Artificial Intelligence Is Taking Computer Chess Beyond Brute Force*, POPULAR SCI. (Sept. 16, 2015), http://www.popsci.com/artificial-intelligence-takes-chess-beyond-brute-force [https://perma.cc/PYR4-7DW2].

[36] To illustrate, consider the algorithm for performing a least-squares regression. That algorithm does not, for example, provide rules or instructions for assessing the data from a drug's clinical trial or the polling data for an election. Rather, the algorithm contains instructions for performing the regression and is capable of applying to different data without regard for the context.

score by 10, the grade-point average by 6, and then adjust based on difficulty by multiplying by the high school difficulty score. Then, rank the scores and the students in the top 10% of the scores are admitted. The second computer program is given the same data about the candidates for admission — SAT score, grade-point average, high school difficulty — but is also given historical admissions decisions as well as the corresponding SAT, grade-point, and difficulty scores. Because the second computer program is not given hard and fast rules, it must devise its own way of determining which students to admit and which to reject, based on its knowledge of past data.

This Article is concerned with this second type of AI that learns from data and solves problems dynamically.[37] This class of AI will often use machine-learning techniques, such as the ones described in this section, to arrive at a dynamic solution to a problem. Fundamentally, the defining characteristic of the AI at issue in this Article is their ability to learn from data.[38]

B. How Do Machine-Learning Algorithms Work?

Many modern machine-learning algorithms share their pedigree with the vast array of statistical inference tools that are employed broadly in the physical and social sciences.[39] They may, for example, use methods that minimize prediction error, adjust weights assigned to various variables, or optimize both in tandem.[40] For instance, a machine-learning algorithm may be given three pieces of data, such as a person's height, weight, and age, and then charged with the task of predicting the time in which each person in a dataset can run a mile. The machine-learning algorithm would look through hundreds or thousands of examples of people with various heights, weights and ages and their mile times to devise a model. One simple way to do so

[37] While I refer in this Article to machine-learning algorithms presently being used to build AI, I implicitly include as part of my analysis more powerful AI that will be developed in the future. Other commentators have made similar assumptions about the progress of AI. *See, e.g.*, NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 124–25 (2014) (positing that AI may reach a point where it is capable of improving itself, resulting in a feedback loop that significantly advances its own intelligence, perhaps beyond that of its human creators).

[38] *See* ETHEM ALPAYDIN, INTRODUCTION TO MACHINE LEARNING, at xxv (2004) ("We need learning in cases where we cannot directly write a computer program to solve a given problem, but need example data or experience. One case where learning is necessary is when human expertise does not exist, or when humans are unable to explain their expertise."); GOODFELLOW ET AL., *supra* note 34, at 96.

[39] *See* GOODFELLOW ET AL., *supra* note 34, at 19–20.

[40] A simple example of such an algorithm is the logistical regression, which optimizes a likelihood function to generate the probability of an event or outcome. *See id.* at 3; PETER FLACH, MACHINE LEARNING: THE ART AND SCIENCE OF ALGORITHMS THAT MAKE SENSE OF DATA 282–86 (2012).

would be to assign some co-efficient or weight to each piece of data to predict the mile time. For example:

```
Predicted Mile Time = A x Height +
B x Weight + C x Age
```

The algorithm may continue to adjust A, B and C as it goes through the examples it has been given to look for the values for A, B and C that result in the smallest error — that is, the difference between each person in the training data's actual mile time and the algorithm's predicted mile time. Most people will recognize this example as the same framework for a least-squares regression,[41] in which the square of the error of the predicting equation is minimized.[42] Many machine-learning algorithms are directed at a similar task but use more mathematically sophisticated methods to determine weights for each variable or to minimize some defined error or "loss function."[43]

Machine-learning algorithms are often given training sets of data to process.[44] Once the algorithm trains on that data, it is then tested with a new set of data used for validation. The goal of tuning a machine-learning algorithm is to ensure that the trained model will generalize,[45] meaning that it has predictive power when given a test dataset (and ultimately live data).[46]

Machine-learning algorithms commonly (though not necessarily)[47] make predictions through categorization.[48] These "classifiers" are able to, for example, look at millions of credit reports and classify individuals into separate credit risk categories or process images and separate the ones containing faces from the ones that do not. If a ma-

[41] For a description of least-squares regression, see generally WILLIAM MENDENHALL, III, ROBERT J. BEAVER & BARBARA M. BEAVER, INTRODUCTION TO PROBABILITY AND STATISTICS 482–529 (14th ed. 2013).

[42] *See* FLACH, *supra* note 40, at 196–207.

[43] Loss functions are functions that machine-learning algorithms seek to minimize or maximize. *See* GOODFELLOW ET AL., *supra* note 34, at 80. They are sometimes referred to as "cost functions" or "error functions." *Id.* Note that not all machine-learning algorithms share the framework described above (i.e., a series of weights for each variable and a loss function to be optimized). For example, many popular algorithms are based on mathematical descriptions of trees of possible decisions or outcomes. *See generally* FLACH, *supra* note 40, at 129–56.

[44] SEBASTIAN RASCHKA, PYTHON MACHINE LEARNING 11 (2015).

[45] *See* GOODFELLOW ET AL., *supra* note 34, at 20.

[46] *See id.*

[47] As discussed above, in addition to classification, machine-learning algorithms may also directly predict particular values instead of merely classifying data.

[48] *See* FLACH, *supra* note 40, at 52. Classifiers that choose between two possible classifications are called binary classifiers. There are also models that can map inputs to multiple classifications. *See id.* at 81–82. Some classifiers can also be designed to provide probability estimates that a particular input should be mapped to a given class. *See generally id.* at 72–76. Classifiers can also be designed to provide rankings of various possible classes to which an input belongs. *See id.* at 61–62.

chine-learning algorithm is properly generalizing, it will correctly predict the appropriate classification for a particular data point.

C. Two Machine-Learning Algorithms Widely Used in AI and the Black Box Problem

One possible reason AI may be a black box to humans is that it relies on machine-learning algorithms that internalize data in ways that are not easily audited or understood by humans. This section provides two illustrative examples. First, a lack of transparency may arise from the complexity of the algorithm's structure, such as with a deep neural network, which consists of thousands of artificial neurons working together in a diffuse way to solve a problem. This reason for AI being a black box is referred to as "complexity." Second, the lack of transparency may arise because the AI is using a machine-learning algorithm that relies on geometric relationships that humans cannot visualize, such as with support vector machines. This reason for AI being a black box is referred to as "dimensionality." This section provides a description of deep neural networks to illustrate how complexity arises, and likewise provides a description of support vector machines to demonstrate how dimensionality can limit transparency.

1. Deep Neural Networks and Complexity

The deep neural network is based on a mathematical model called the artificial neuron. While originally based on a simplistic model of the neurons in human and animal brains, the artificial neuron is not meant to be a computer-based simulation of a biological neuron. Instead, the goal of the artificial neuron is to achieve the same ability to learn from experience as with the biological neuron.[49] Multi-layered networks of these interconnected artificial neurons were not possible until the mid-1980s, when a method of training such networks was rediscovered and further developed.[50] Since then, the ability to connect layers of neural networks has yielded staggering results. What has emerged is the so-called "deep" architecture of artificial neurons,

[49] The notion that deeply interconnected networks can solve computational problems is called connectionism, which gained traction in the 1980s. *Id.* at 16 ("The central idea in connectionism is that a large number of simple computational units can achieve intelligent behavior when networked together. This insight applies equally to neurons in biological nervous systems as it does to hidden units in computational models.").

[50] The backpropagation algorithm for training layers of artificial neurons was first developed by Paul Werbos in 1974 but went largely unnoticed and unused until brought to prominence by others in 1985. *See* Paul Werbos, Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences (Aug. 1974) (unpublished Ph.D. thesis, Harvard University) (on file with the Gordon McKay Library, Harvard University); Bernard Widrow & Michael A. Lehr, *30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation*, 78 PROCS. IEEE 9 (1990).

referred to as Deep Neural Networks, where several layers of interconnected neurons are used to progressively find patterns in data or to make logical or relational connections between data points. Deep networks of artificial neurons have been used to recognize images, even detecting cancer at levels of accuracy exceeding that of experienced doctors.[51]

No single neuron in these networks encodes a distinct part of the decision-making process.[52] The thousands or hundreds of thousands of neurons work together to arrive at a decision.[53] A layer or cluster of neurons may encode some feature extracted from the data (e.g., an eye or an arm in a photograph), but often what is encoded will not be intelligible to human beings.[54] The net result is akin to the way one "knows" how to ride a bike. Although one can explain the process descriptively or even provide detailed steps, that information is unlikely to help someone who has never ridden one before to balance on two wheels. One learns to ride a bike by attempting to do so over and over again and develops an intuitive understanding.[55]

Because a neural network is learning from experience, its decision-making process is likewise intuitive. Its knowledge cannot in

[51] For example, one AI program has been able to detect breast cancer with accuracy rates exceeding that of experienced doctors. *See, e.g.*, Martin Stumpe & Lily Peng, *Assisting Pathologists in Detecting Cancer with Deep Learning*, GOOGLE RES. BLOG (Mar. 3, 2017), https://research.googleblog.com/2017/03/assisting-pathologists-in-detecting.html [https://perma.cc/YWK2-2FAS] ("In fact, the prediction heatmaps produced by the algorithm had improved so much that the localization score (FROC) for the algorithm reached 89%, which significantly exceeded the score of 73% for a pathologist with no time constraint.").

[52] This is sometimes referred to as a "distributed representation," meaning that "each input to a system should be represented by many features, and each feature should be involved in the representation of many possible inputs." GOODFELLOW ET AL., *supra* note 34, at 16; *see also supra* notes 9–12.

[53] Modern neural networks may feature tens of thousands of interconnected artificial neurons and at the high-end, even hundreds of thousands. Interestingly, however, most neural networks do not even exceed the number of neurons in the nervous system of a frog. *See* GOODFELLOW ET AL., *supra* note 34, at 21.

[54] To be sure, in some cases, such as with image recognition, one can examine what some group of neurons has identified — they may encode a portion of a type of image — but even in those cases, neural networks will identify features in the data that will look like visual noise to human beings. *See* Castelvecchi, *supra* note 9; Andrej Karpathy, *Understanding and Visualizing Convolution Neural Networks*, CS231N CONVOLUTIONAL NEURAL NETWORKS FOR VISUAL RECOGNITION, http://cs231n.github.io/understanding-cnn/ [https://perma.cc/P8R4-LS5E] (describing approaches to understanding a class of neural networks used to analyze images).

[55] As Siddhartha Mukherjee notes in his article in the New Yorker, the distinction is between two types of knowledge that British philosopher Gilbert Ryle referred to as "knowing that" and "knowing how." *See* Siddhartha Mukherjee, *A.I. Versus M.D.*, NEW YORKER (Apr. 3, 2017), http://www.newyorker.com/magazine/2017/04/03/ai-versus-md [https://perma.cc/Q2L6-ZLYJ]. Knowing some factual propositions about a task can be characterized as "knowing that", "[b]ut to learn to ride a bicycle involves another realm of learning. A child learns how to ride by falling off, by balancing herself on two wheels, by going over potholes. Ryle termed this kind of knowledge — implicit, experiential, skill-based — knowing how.'" *Id.*

most cases be reduced to a set of instructions, nor can one in most cases point to any neuron or group of neurons to determine what the system found interesting or important.[56] Its power comes from "connectionism," the notion that a large number of simple computational units can together perform computationally sophisticated tasks.[57] The complexity of the large multi-layered networks of neurons is what gives rise to the Black Box Problem.

2. Support Vector Machines and Dimensionality

Some machine-learning algorithms are opaque to human beings because they arrive at decisions by looking at many variables at once and finding geometric patterns among those variables that humans cannot visualize. The support vector machine ("SVM") illustrates this. The SVM was invented in 1963[58] and modified to classify data in 1991.[59] To understand the principle underlying the support vector machine, consider a two-dimensional example.

Assume our SVM is tasked with taking height and weight and determining whether a person is male or female. If we plotted each person's height and weight on a two-dimensional graph as in Figure 1, we can then attempt to draw a dividing line through the data that we can use to make a prediction. If a height / weight combination falls on one side of the line, the person is predicted to be male; if the person falls on the other side, they are predicted to be female. As Figure 2 shows, there are multiple ways one could draw the dividing line, but line *b* is clearly the best for making predictions. Line *b* reflects the key insight upon which the SVM is based: the line that creates the largest distance or margin between one class and the other is probably the most predictive and one that generalizes the best.[60]

[56] *See id.*; *see also supra* notes 9–12 and accompanying text.

[57] GOODFELLOW ET AL., *supra* note 34, at 16.

[58] The support vector machine was invented by Vladimir Vapnick in the 1960s. Mr. Vapnick recently joined Facebook's AI team in 2014. Jordan Novet, *Facebook AI Team Hires Vladimir Vapnick, Father of the Modern Vector Machine Algorithm*, VENTUREBEAT (Nov. 25, 2014), https://venturebeat.com/2014/11/25/facebooks-ai-team-hires-vladimir-vapnik-father-of-the-popular-support-vector-machine-algorithm/ [https://perma.cc/A3TQ-HBJS].

[59] *See* Isabelle Guyon, *Data Mining History: The Invention of Support Vector Machines*, KDNUGGETS (July 2016), http://www.kdnuggets.com/2016/07/guyon-data-mining-history-svm-support-vector-machines.html [https://perma.cc/N459-CRUY] (describing the history of the SVM by one of the scientists that modified the algorithm in the 1990s).

[60] The SVM mathematically arrives at this optimal solution by arriving at a maximized margin between each category being classified and the dividing line it draws. The margin is between "support vectors" near the dividing line. *See* FLACH, *supra* note 40, at 211–16.

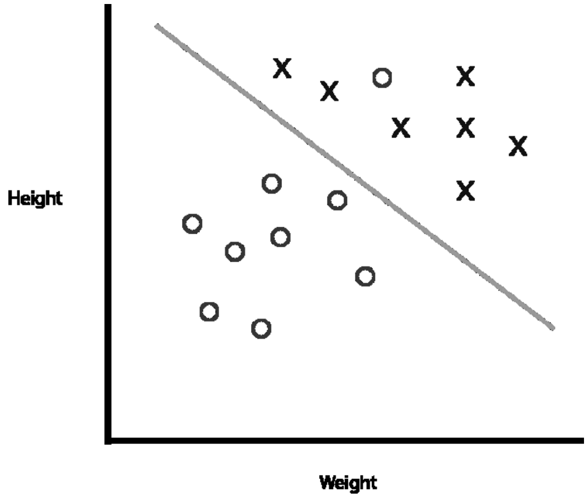

Figure 1: If we graph the men as Xs and the women as Os, we can see that the dividing line depicted above correctly classifies all of the men and most of the women. Only 1 woman is misclassified out of a total of 9, meaning our model has an approximately 11% error rate.

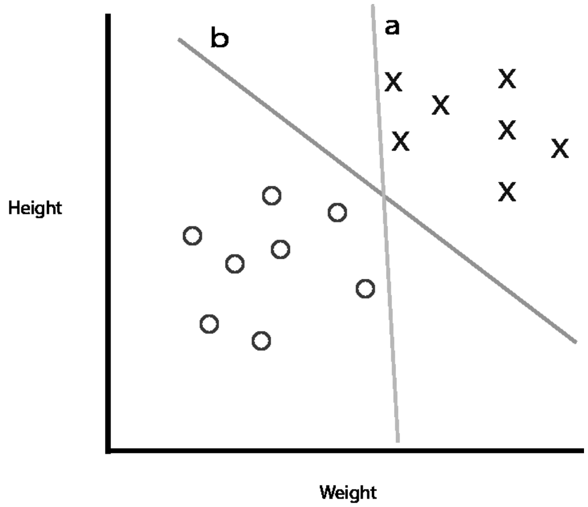

Figure 2: The graph in figure two has two dividing lines, a and b. Both dividing lines have the same accuracy — that is, they classify all of the data correctly. The key insight exploited by an SVM is that line b is likely better suited for generalizing on new data than line a because line b maximizes the distance between the two classes and the dividing line (the margin).

What is important for the purposes of this Article is to note that the dividing line is a line when there are only two features or variables provided to the model. When there are three variables, the dividing line will be a plane. If, however, we provide the model with 17 variables or even 1000 variables, the human mind is unable to visualize what that dividing line looks like. Human brains simply cannot visually process high dimensionality.[61] Moreover, not all SVMs use straight lines to divide the data — that is, a mathematical method used with SVMs allows for non-linear (i.e., curved) divisions.[62] Thus, when the number of variables or features provided to an SVM becomes large, it becomes virtually impossible to visualize how the model is simultaneously drawing distinctions between the data based on those numerous features. [63] An AI that uses an SVM to process dozens or perhaps hundreds of variables would thus be a black box to humans because of the dimensionality of the model, despite being a shallow (i.e. less complex) model relative to deep neural networks.[64]

D. Weak and Strong Black Boxes

Generally, the Black Box Problem can be defined as an inability to fully understand an AI's decision-making process and the inability to predict the AI's decisions or outputs. However, whether an AI's lack of transparency will have implications for intent and causation tests depends on the extent of this lack of transparency. A complete lack of transparency will in most cases result in the complete failure of intent and causation tests to function, but some transparency may allow these tests to continue functioning, albeit to a limited extent. It therefore makes sense to further subdivide the Black Box Problem into two categories.

[61] Humans generally cannot visualize higher-dimensional patterns and shapes without using some method of chunking the information into three dimensions at a time or compressing the image into three dimensions (thus losing information). *See generally* Sean Carroll, *Why Can't We Visualize More than Three Dimensions?*, DISCOVER BLOG (Mar. 30, 2009), http://blogs.discovermagazine.com/cosmicvariance/2009/03/30/why-cant-we-visualize-more-than-three-dimensions/ [https://perma.cc/Q3FV-3NCX]; *see also* CARL SAGAN, COSMOS 279 (Ballantine Books 2013) (1980) (considering what a three-dimensional object would look like to a two-dimensional being).

[62] This mathematical method is referred to as the Kernel Trick and is often used with SVMs to create non-linear models. *See* FLACH, *supra* note 40, at 224–27.

[63] The only possible description of such a model's decision-making is a mathematical one, but for lawyers, judges, juries, and regulators, an expert may be required to describe the model mathematically, and in many cases, even an expert is unlikely to be able to describe (mathematically or otherwise) how the model is making decisions or predictions, let alone translate that description for a regulator or fact finder.

[64] *See* Li Deng & Dong Yu, *Deep Learning: Methods and Applications*, 7 FOUND. & TRENDS SIGNAL PROCESSING 197, 205 (2013) (noting SVMs are examples of "shallow" models).

Strong Black Boxes: Strong black boxes are AI with decision-making processes that are entirely opaque to humans. There is no way to determine (a) how the AI arrived at a decision or prediction, (b) what information is outcome determinative to the AI, or (c) to obtain a ranking of the variables processed by the AI in the order of their importance. Importantly, this form of black box cannot even be analyzed ex post by reverse engineering the AI's outputs.

Weak Black Boxes: The decision-making process of a weak black box are also opaque to humans. However, unlike the strong black box, weak black boxes can be reverse engineered or probed to determine a loose ranking of the importance of the variables the AI takes into account. This in turn may allow a limited and imprecise ability to predict how the model will make its decisions. As explained further *infra* in Parts III and IV, weak black boxes may not entirely cause intent and causation tests to cease to function, though they still pose serious challenges for both legal doctrines.[65]

III. THE BLACK BOX PROBLEM AND THE FAILURE OF INTENT

Intent tests appear throughout the law and have developed over centuries to help courts and juries understand and regulate human conduct. Intent, for example, is a means of finding out whether a person intended to cause a particular outcome to occur.[66] Intent may also determine whether the severity of a penalty is appropriate for the particular conduct.[67]

Machines and computer programs have no intent. The most we can glean from how they work and how they are designed is what goals their users or creators sought to achieve and the means they permitted their machine or program to use to achieve them.[68] It there-

[65] To be sure, the strong and weak black boxes are not perfect descriptions of how the Black Box Problem manifests itself in real-world applications. They are idealizations used in this Article to demonstrate the effect on intent and causation tests. Real AI may exhibit some subset of the constraints imposed by these concepts.

[66] Civil forms of intent such as intent tests used in tort law, for example, define intent as the state of mind of a person that "either (1) has a purpose to accomplish that result or (2) lacks such a purpose but knows to a substantial certainty that the defendant's conduct will bring about the result." DOBBS, *supra* note 30, at 48.

[67] The most obvious example is the various degrees of homicide that are defined in criminal statutes, with more severe forms of unlawful killings warranting more severe sentences. *See, e.g.*, CAL. PENAL. CODE. §§ 187–192 (Deering, LEXIS through Ch. 6 of 2018 Reg. Sess.) (defining sentences for various forms of homicide, with diminishing penalties for less culpable mens rea).

[68] Indeed, some laws directly reference the purpose of software in defining unlawful conduct. The Digital Millennium Copyright Act ("DMCA"), for example, provides that "[n]o person shall manufacture, import, offer to the public, provide, or otherwise traffic in any technology, product, service, device, component, or part thereof, that — (a) is *primarily designed or produced for the purpose of* circumventing a technological measure that effec-

fore makes sense to speak about the intent of the designer or user. For example, we may infer from a computer program designed to break into a computer system that its creator intended to use it for that purpose.[69] In some cases, we can look at the computer program's instructions to determine what the designer of the program was trying to accomplish and what means could be used by the program to accomplish that goal.

Black-box AI, however, may function in a manner well outside of what the program's creators could foresee. To be sure, we may be able to tell what the AI's overarching goal was, but black-box AI may do things in ways the creators of the AI may not understand or be able to predict.[70] An AI securities-trading program, for example, may be given the overarching goal of maximizing profit, but how it makes its trading decisions or whether it meets its objective through market manipulation may be entirely unclear ex ante to its creators and even ex post to courts and regulators. Because we cannot look to the program's instructions or design to determine the intent of its creator or user, intent tests become impossible to satisfy.

If intent tests cannot be satisfied, laws relying on them will cease to function. Most critically, because intent tests appear in the law where penalties are most severe (such as in criminal statutes), the most dangerous or noxious conduct may go unregulated if it is AI, rather than a human, that engages in it. AI would be exempt from the most stringent aspects of our criminal, securities,[71] and antitrust

tively controls access to a work protected under this title." 17 U.S.C. § 1201(a)(2)(A) (2012) (emphasis added).

[69] Inferring intent from the instructions of a computer program will often be far from straightforward. Sometimes an algorithm's purpose is not universally malicious or unlawful but may be used for unlawful purposes. For example, just because a criminal defendant uses encryption, does not necessarily imply that he was doing so in furtherance of a crime. *See* Neal Kumar Katyal, *Criminal Law in Cyberspace*, 149 U. PA. L. REV. 1003, 1060 (2001); *see also* United States v. Boyajian, No. CR 09-933(A) CAS, 2013 U.S. Dist. LEXIS 116492, at *5–6 (C.D. Cal. Aug. 14, 2013) (holding that the use of encryption does not warrant the inference of consciousness of guilt).

[70] It is precisely this property of some machine-learning algorithms that allow them to be used to dynamically devise forms of encryption that AI can use to securely communicate with each other. *See, e.g.*, Martin Abadi & David G. Andersen, *Learning to Protect Communications with Adversarial Neural Cryptography*, ARXIV (Oct. 24, 2016), https://arxiv.org/pdf/1610.06918v1.pdf [https://perma.cc/SWB9-5W55]. Similar machine-learning algorithms can even be designed to dynamically generate their own language, which even the creator of the computer program may not be able to interpret or understand. *See* Metz, *supra* note 4.

[71] Both civil claims for securities fraud as well as criminal charges (which carry a potential prison sentence of up to 25 years, 18 U.S.C. § 1348 (2012)), require proof of scienter. *See* Dura Pharms., Inc. v. Broudo, 544 U.S. 336, 346 (2005); United States v. Litvak, 808 F.3d 160, 178 (2d Cir. 2015) (requiring scienter in criminal securities fraud case).

laws,[72] which often require a showing of intent to give rise to the most serious forms of civil and criminal liability.[73]

This section is about the several ways intent tests break down when AI is involved. First, this section discusses the recent wave of non-intelligent algorithms used by securities and commodities traders and their interactions with intent tests. Next, this section discusses three categories of intent tests — Effect Intent, Basis Intent, and Gatekeeping Intent — and how each type of test interacts with strong and weak black-box AI. This section argues that these intent tests fail largely for three reasons:

(1) An AI's conduct or decisions may tell us nothing about its designer's or user's intent, which means that intent tests based on a person's intent to achieve an unlawful outcome become unsatisfiable;

(2) Because it may be impossible to determine the bases of an AI's decision or prediction, intent tests that scrutinize the bases or justifications for conduct become unsatisfiable; and

(3) Because intent tests often serve as a gatekeeper, limiting the scope of claims, they may entirely prevent certain claims or legal challenges from being raised when AI is involved.

All of these problems threaten to leave AI unregulated either because defendants that use AI may never be held liable (e.g., the government's use of AI may prevent a showing of discriminatory intent) or claimants that rely on AI may be left without legal redress (e.g., because a plaintiff that uses AI to make investment decisions is unable to show reliance).

A. Non-AI Algorithms and Early Cracks in Intent

High frequency trading ("HFT") algorithms, which were used by financial firms to trade securities and commodities in fractions of sec-

[72] Civil antitrust claims can result in trebled damages, *see* 15 U.S.C. § 15(b) (2012), and often require an inquiry into the justifications for allegedly anticompetitive conduct. *See, e.g.*, Retractable Techs., Inc. v. Becton Dickinson & Co., 842 F.3d 883, 892 (5th Cir. 2016) ("To determine whether conduct is exclusionary, the court looks to the 'proffered business justification for the act'. 'If the conduct has no rational business purpose other than its adverse effects on competitors, an inference that it is exclusionary is supported.'" (quoting Taylor Publ'g Co. v. Jostens Inc., 216 F.3d 465 (5th Cir. 2000); Stearns Airport Equip. Co. v. FMC Corp., 170 F.3d 518 (5th Cir. 1999)).

[73] *See, e.g.*, Stuart P. Green, *Moral Ambiguity in White Collar Criminal Law*, 18 NOTRE DAME J.L. ETHICS & PUB. POL'Y 501, 502–07 (2004).

onds,[74] were some of the first algorithms to expose the potential problems with the intent tests. HFTs are for the most part based on hard-coded rules that allow computer systems to react faster than any human being.[75] The central strategy for many of these HFTs is to identify an inefficiency in the market and to trade them away before anyone else.[76] The same speed that allows these algorithms to exploit market inefficiencies also allows them to engage in conduct that may be unethical or border on being unlawful.[77] For example, an HFT algorithm can be used to beat other orders to market by fractions of a second, allowing the algorithm to (a) determine that someone was seeking to buy a security at a certain price, (b) buy the security before the other person does, and (c) sell it to them at a higher price.[78] They may also be used to engage in conduct such as "spoofing," where the algorithm places phantom orders on markets, only to withdraw them once the market has moved in a desired direction.[79]

Early lawsuits against firms that used or facilitated HFTs were largely unsuccessful.[80] Some, for example, never made it past motions to dismiss largely because allegations of intent are required at the onset of a lawsuit. As one court held in the commodities context, because the Commodities and Exchange Act ("CEA") requires that a person act with the "conscious object of influencing prices," it is not enough to allege "knowledge that certain actions might have an impact on the futures market" to bring a private claim under the CEA.[81] For HFTs in the securities markets, courts have taken a similar approach, defining manipulation of markets as "intentional or willful conduct designed to deceive or defraud investors by controlling or

[74] In 2011, it was estimated that "high-frequency trading made up about 60% of U.S. equity trading and 35 to 40% of European equity trading." Tom C.W. Lin, *The New Financial Industry*, 65 ALA. L. REV. 567, 575 (2013).

[75] *See* Yesha Yadav, *The Failure of Liability in Modern Markets*, 102 VA. L. REV. 1031, 1077 (2016) ("For HF algorithms to maintain execution speeds measured in microseconds and milliseconds, they must be preset and predictively model market behavior.").

[76] *See, e.g.*, United States v. Coscia, 866 F.3d 782, 786 (7th Cir. 2017) ("[D]iscrepancies [in price] often last a very short period of time (i.e., fractions of a second); speed in execution is therefore an essential attribute for firms engaged in this business.").

[77] *See id.* at 786 ("Although high-frequency trading has legal applications, it also has increased market susceptibility to certain forms of criminal conduct. Most notably, it has opened the door to spoofing.").

[78] *See* Yadav, *supra* note 75, at 1065–66 ("The HF trader can earn steady profits by being a constant counterparty for investors, particularly if it can trade tens of thousands of times over a day and incrementally earn small spreads on each deal.").

[79] *See id.* at 1069 (describing the spoofing scheme at Trillium Brokerage Services, in which "Trillium submitted waves of false buy and sell orders with a view to inducing other market participants to transact").

[80] The Seventh Circuit, for example, only recently affirmed the very first criminal conviction for HFT-based spoofing. *Coscia*, 866 F.3d at 803.

[81] Braman v. The CME Group, Inc., 149 F. Supp. 3d 874, 889–90 (N.D. Ill. 2015).

artificially affecting the price of securities."[82] Because "manipulation" is defined as the result of *willful* conduct, courts have dismissed claims that do not clearly explain how algorithms to buy and sell securities intentionally affected the prices of those securities."[83]

Both types of intent tests require some ability by the users of the HFTs to foresee the effects of the HFT's conduct on the market or on prices. Where an HFT is not clearly designed for an unlawful purpose, it will therefore be difficult to prove that the HFT had an illegitimate price or market impact, let alone that the firm using the algorithm intended such an impact. What is more, the speed at which these algorithms execute transactions creates a degree of unpredictability, as contagion and feedback effects from an error made by a single HFT can interact with other market dynamics (or other HFTs) to cause rapid price movements.[84]

It is thus unsurprising that in *Coscia*, the only criminal conviction for spoofing to date, the conviction was based on the testimony by the programmer of the HFT program as to what it was programed to do.[85] As the court explained:

> The designer of the programs, Jeremiah Park, testified that Mr. Coscia asked that the programs act "[l]ike a decoy," which would be "[u]sed to pump [the] market." Park interpreted this direction as a desire to "get a reaction from the other algorithms." In particular, he noted that the large-volume orders were designed specifically to avoid being filled and accordingly would be canceled in three particular circumstances: (1) based on the passage of time (usually measured in milliseconds); (2) the partial filling of the large orders; or (3) complete filling of the small orders.[86]

In most cases, there will be no such direct testimony that an algorithm was designed for unlawful purposes. And, where transactions are on their face legitimate, proving intent becomes even more difficult. The test applied by some courts to open-market transactions is illustrative. For example, some courts have held that where the conduct underlying a market manipulation is an open-market transaction,

[82] *In re* Barclays Liquidity Cross and High Frequency Litig., 126 F. Supp. 3d 342, 361 (S.D.N.Y. 2015).

[83] *Id.* at 364.

[84] *See* Yadav, *supra*, note 75, at 1079 ("HF algorithms are preprogrammed to respond instantly to new information and to the errors and mischiefs of other traders — with human beings unable to intervene in real time to correct mishaps.").

[85] *See Coscia*, 866 F.3d at 802–03.

[86] *Id.* at 789 (citation omitted).

such as a purchase or short sale, there must be a showing that the conduct lacked any legitimate economic reason before there can be liability under Section 10(b).[87] The court explained:

> [I]f an investor conducts an open-market transaction with the intent of artificially affecting the price of the security, and not for any legitimate economic reason, it can constitute market manipulation. Indeed, "the only definition [of market manipulation] that makes any sense is subjective — it focuses entirely on the intent of the trader."[88]

Where an algorithm is designed to use simple buy and sell transactions at rapid speeds to manipulate prices, such intent tests would find no liability. For example, an algorithm may be designed to enter into legitimate transactions 90% of the time and "spoof" the other 10% of the time — that is, place market orders only to rapidly withdraw them. In such a case, it will be difficult to demonstrate that the algorithm was designed to engage in spoofing, particularly when the designer of the algorithm can point to hundreds of thousands of legitimate transactions on a motion for summary judgment.[89]

The courts' early experience with HFTs brings several intent problems to light. First, intent tests, such as those resulting in the dismissal of the HFT cases, which require a showing that some unlawful effect is intended, will fail if the designers or users of the computer program cannot predict the effects of the algorithm ex ante. Second, where the law uses an intent test to prevent legitimate transactions from giving rise to liability, the user of a computer program may be insulated from liability by the program's unpredictability or speed, particularly where there is no direct evidence of the unlawful purpose of the program. Finally, intent tests can be a bar at the threshold of litigation to bringing a claim, resulting in early dismissal. How each of these problems is compounded by black-box AI will be discussed further in this section.

B. AI and Effect Intent

Consider the following hypothetical: a system of deep neural networks is designed to devise a profitable trading strategy in the eq-

[87] *See* SEC v. Masri, 523 F. Supp. 2d 361, 372 (S.D.N.Y. 2007).

[88] *See id.* (citation omitted).

[89] Indeed, in *Coscia*, there was evidence that the defendant entered orders that were never ultimately filled, with an order-to-fill ratio of 1600%, whereas other market participants had ratios of 91%. 866 F.3d at 789. A more innocuous fill rate may therefore obfuscate spoofing.

uities markets. It is given access to a broad range of data, including a Twitter account, real-time stock prices of thousands of securities, granular historical price data, and access to popular business news feeds. Within months of training on data, the algorithm is able to consistently turn a profit. It is unclear what strategy the AI has stumbled upon, but it is rapidly placing trading orders, consummating some of them and rapidly withdrawing or changing others. Interestingly, the system has learned to "retweet" news articles on Twitter and often does so before and after trades.[90] The designer of the system is not able to tell what role the retweets have in the overall trading strategy, nor is he able to tell why certain trade orders are consummated and others withdrawn. All he can tell is that his AI is working and is profitable.

Within days, the price of one of the securities that the AI frequently trades crashes steeply within seconds. The AI, which can either take a long or short position in the security, has, however, managed to make a profit. When private investors learn about the AI system's participation in the market for that security, they bring suit under the anti-fraud provisions of the Securities Exchange Act — namely, Section 10(b) and SEC Rule 10b-5.[91] The lawsuit alleges (a) market manipulation through phantom orders, and (b) that the proponent of the AI made false or misleading statements about the stock prior to buying and selling the securities by retweeting false factual statements.

The creator of the AI would demure that he has no idea the extent to which either form of alleged conduct is even integral to the AI algorithm's trading strategy. Although he will also make a causation argument, which we will discuss *infra* in detail,[92] he will likely argue that the element of scienter, which must be proven in anti-fraud cases,[93] is lacking. Specifically, he will argue that although he gave the computer program full access to all of the functions of a Twitter account, he never designed the algorithm to retweet information or to place phantom orders. He merely gave it the broad objective of maximizing profits. He had no intent to manipulate prices or to make any statements in connection with the purchase or sale of securities. In fact, he may argue that it surprised him entirely that the AI system learned to retweet, as he never designed it to use a Twitter account in that way.

[90] Retweets are simply the repeating by one user of another user's twitter message. *See Retweet FAQs*, TWITTER, https://support.twitter.com/articles/77606 [https://perma.cc/QHD3-UVYD].

[91] 15 U.S.C. § 78j (2012); 17 C.F.R. § 240.10b-5 (2017).

[92] *See infra* Part IV.

[93] *See* Dura Pharms., Inc. v. Broudo, 544 U.S. 336, 346 (2005).

This hypothetical highlights a problem with a specific class of intent tests that this Article refers to in this section as Effect Intent tests. Effect Intent is a test that requires that a person intended the unlawful consequences of his action or that he engaged in the conduct with the intent to accomplish something unlawful. As shown in the last section, this may be the sort of test applied by courts in market-manipulation cases, where there must be evidence that the designer of the algorithm intended to have a distortive effect on market prices.[94]

In the above hypothetical, the designer of the program never provided the AI with anything other than the lawful objective of making a profit. The AI then devised a strategy that in part employs a potentially prohibited strategy — it retweets information of the sort that in the past has been able to move markets. It may be that a certain kind of tweet containing misinformation about a company has the most market impact. The AI has no way of vetting the accuracy of the data before tweeting, just its impact. The only intent we can examine is the designer's, and there is no evidence that the designer intended that the AI engage in the particular strategy it chose.

The Black Box Problem renders the AI's decision-making process impenetrable. It may even be impossible to prove what the AI's motivation was for a particular retweet, and with both strong and weak forms of the Black Box Problem, it is impossible to tell what strategy the AI had adopted. We can only look at the effect of the AI's actions. If the test is whether the creator of the AI intended the effect, there will almost never be any liability.[95]

In the case of a weak black box, where one can loosely rank the value of the data processed by the AI, we may be able to tell what sort of tweet the AI found interesting, but we cannot determine what overall strategy the AI executed. Moreover, we cannot determine the effect that any individual piece of information had on the overall decision-making of the AI.[96] In the case of both strong and weak black boxes, the AI's conduct or decision-making is dynamically generated based on past data, so the AI's creator or user will not have known ex ante what decisions or strategies the AI would ultimately use. The AI's creator would therefore not be able to foresee the effects

[94] *See, e.g.*, *In re* Barclays Liquidity Cross and High Frequency Litig., 126 F. Supp. 3d 342, 361 (S.D.N.Y. 2015).

[95] Because Effect Intent tests require intent to cause a particular outcome, it will not be enough to point to the creator's negligence or failure to place constraints on the AI to satisfy such a test. For example, one Effect Intent test applied to market-manipulation claims under the Commodities Exchange Act requires that the defendant have a "purpose or conscious object of causing or effecting a price or price trend in the market." CFTC v. Johnson, 408 F. Supp. 2d 259, 267 (S.D. Tex. 2005). A failure to constrain the conduct of the AI would not likely meet that standard if the AI's particular strategy was not foreseeable to its creator.

[96] *See supra* Section II.D.

of the AI's decisions, nor would he even be able to fully understand the AI's decision-making process or conduct ex post.

C. AI-Assisted Opinions and Basis Intent

The hypothetical *supra* in Section III.B, also highlights a different problem. Some forms of intent require an examination of a decision-maker or actor's basis for conduct. The AI in that section placed trading orders, consummating some of them and rapidly withdrawing or changing others. It is possible that the algorithm was engaging in spoofing. The evidence, however, will likely be equivocal. The designer of the algorithm will have tens of thousands of legitimate transactions to point to for every dozen or so withdrawn orders. The Black Box Problem ensures that there is no way to determine what the AI's particular strategy is. Unlike the first generation of algorithms discussed earlier, there will not be instructions somewhere in the AI's programming that are designed to engage in spoofing,[97] so there will not be testimony from the AI's designer to that effect, as there was in *Coscia*.[98] In a neural network, for example, a series of interconnected artificial neurons may be mimicking data from the past, which may have simply reflected that a rise in price was correlated with placing and withdrawing trades.[99] If the AI is a black box, there is no way of knowing. While the author of the AI's programming could have expressly prohibited this sort of conduct, the failure to do so is likely the result of negligence, not an intentional design decision. This would fall short of the sort of intent required for most criminal laws or civil fraud causes of action.

This problem becomes even more intractable where the particular basis for a decision is the central question of a litigation or regulatory investigation. For example, consider another hypothetical from the securities laws. A large financial institution uses AI to appraise homes that will serve as collateral for mortgage-backed loans that it will make and ultimately package into mortgage-backed securities. The financial institution then provides the appraised values in its offering documents for the mortgage-backed security.[100] An investor later

[97] *See supra* Section III.A.

[98] United States v. Coscia, 866 F.3d 782, 790 (7th Cir. 2017).

[99] *See supra* Section II.C.1.

[100] This fact pattern has appeared before, but with non-intelligent computer programs valuing the houses. During the height of the 2005–2008 mortgage-backed-security bubble, automated valuation models ("AVMs"), which are "computer programs that use statistical models to reach objective estimates of the market value of real property," were widely used as part of the due diligence process for mortgage loans that would be packaged into mortgage-backed securities. *See* Mass. Mut. Life Ins. Co. v. DB Structured Prods., 110 F. Supp. 3d 288, 293 (D. Mass. 2015). Government Sponsored Entities, such as Fannie Mae, also used AVMs to review loans they guaranteed as well as loans that were packaged into securities they purchased. *See* Fed. Hous. Fin. Agency v. Nomura Holding Am., Inc., 60 F. Supp.

sues, alleging that the home values stated in the offering documents were incorrect and seeking rescission and damages under Sections 11 and 12 of the Securities Act of 1933.[101]

For years, courts have held that valuations are statements of opinion; thus, to be actionable under Section 11 or 12 of the Securities Act, the opinion must not only be wrong, but the speaker of the opinion must not have believed the opinion to be true.[102] That is, statements of opinion "affirm[] one fact: that the speaker actually holds the stated belief."[103] In other words, as the Supreme Court held in *Omnicare*, even if an opinion turns out to be wrong, it is not actionable — the law "does not allow investors to second-guess inherently subjective and uncertain assessments."[104] The securities laws are simply not "an invitation to Monday morning quarterback an issuer's opinions."[105]

Under this rubric, our hypothetical would almost never result in liability, even if the AI renders demonstrably problematic valuation opinions. The financial institution can always say that it designed the AI with the utmost care and verified its accuracy on past data. It will almost always be able to argue that it subjectively believed in the valuation opinions it was publishing — it designed a highly sophisticated machine to carefully look at every house underlying every mortgage. A plaintiff may be able to argue that the AI was not accurate enough for such reliance, that the AI was inadequately tested on out-of-sample data, and even that the issuer had some duty to sanity check the results, but none of these arguments will likely suffice to allege a subjective disbelief of the opinions. There will have to be something more, particularly at the motion to dismiss stage, that would make plain that algorithm's design or opinions were problematic and that the issuer knew it. Again, as with the Effect Intent cases, the opinion-statement test will be actionable only at the margins — where there is particularly obvious and egregious conduct.

Black-box AI ensures that the problems will be impossible to detect without access to the AI and the ability to probe it to determine why it makes particular decisions. Even then, the AI may be a com-

3d 479, 491–92 (S.D.N.Y. 2014). The next generation of valuation model will no doubt be based less on statistical models and more on machine-learning algorithms and artificial intelligence, which may make them comparable in accuracy to human appraisers.

[101] *See* 15 U.S.C. § 77k (2012) (Section 11 of the Securities Act); 15 U.S.C. § 77l (Section 12). Section 11 provides for damages arising from a false statement in a registration statement, 15 U.S.C. § 77k(e), and Section 12 provides for rescission or rescissionary damages, 15 U.S.C. § 77l(a).

[102] *See, e.g.*, Fait v. Regions Fin. Corp., 655 F.3d 105, 110 (2d Cir. 2011); Rubke v. Capitol Bancorp, Ltd., 551 F.3d 1156, 1162 (9th Cir. 2009).

[103] Omnicare, Inc. v. Laborers Dist. Council Constr. Indus. Pension Fund, 135 S. Ct. 1318, 1326 (2015).

[104] *Id.* at 1327.

[105] *Id.*

plete black box. It is almost impossible for a plaintiff, such as a purchaser of mortgage-backed securities, to be able to make allegations that the AI was designed or tested poorly or made decisions that put the user of the AI on alert that something was wrong.

Omission claims complicate things further. That is, a deeper problem occurs when courts require a detailed probe of the basis for an opinion. The Supreme Court in *Omnicare*, for example, not only set forth the standard for opinion statements that turn out to be false, but also for opinion statements that allegedly omit a material fact about the opinion or its basis.[106] In that context, the Supreme Court has rejected the notion that "[a]s long as an opinion is sincerely held . . . , it cannot mislead as to any matter."[107] The Court explained: "a reasonable investor may, depending on the circumstances, understand an opinion statement to convey facts about how the speaker has formed the opinion — or, otherwise put, about the speaker's basis for holding that view."[108] Thus, "if the real facts are otherwise, but not provided, the opinion statement will mislead its audience."[109]

Prior to looking at how our hypothetical AI would fare under this standard, consider the hypothetical set of facts the Court set forth in its opinion:

> Consider an unadorned statement of opinion about legal compliance: 'We believe our conduct is lawful.' If the issuer makes that statement without having consulted a lawyer, it could be misleadingly incomplete. In the context of the securities market, an investor, though recognizing that legal opinions can prove wrong in the end, still likely expects such an assertion to rest on some meaningful legal inquiry — rather than, say, on mere intuition, however sincere. Similarly, if the issuer made the statement in the face of its lawyers' contrary advice, or with knowledge that the Federal Government was taking the opposite view, the investor again has cause to complain: He expects not just that the issuer believes the opinion (however irrationally), but that it fairly aligns with the information in the issuer's possession at the time. Thus, if a registration statement omits material facts about the issuer's inquiry into or knowledge concerning a statement of opinion, and if those facts conflict with what a reasonable investor

[106] *Id.* at 1328–29.

[107] *Id.* at 1328.

[108] *Id.*

[109] *Id.*

> would take from the statement itself, then § 11's omissions clause creates liability.[110]

The Court's statement presents a serious problem for AI-based opinion statements. If the AI gave too little weight to a particular factor or piece of information, there would be no way of knowing it. Even if the AI is a weak black box, the user of the AI may not be able to tell if a particular piece of information was outcome determinative. How would a plaintiff allege that the speaker of the opinion omitted important information if he cannot explain how the speaker's AI weighed particular information or came to the valuation opinion?

More specifically, in the hypothetical above, if an investor sues under Section 11 of the Securities Act claiming that the AI ignored, for example, the square footage of the homes entirely, the proponent of the AI may respond that the AI was given information about square footage. If the AI is a weak black box, he may even be able to explain that the square footage of the homes was less important to the AI than some other variable, such as the number of bathrooms in the house. There would, however, be no way to prove that the AI inappropriately weighed the square footage, that it would have reached a more accurate decision if it gave the square footage more weight, or that the AI was not accurate enough to rely upon. The net effect would be that the user of the AI would be functionally immune from omissions-based securities lawsuits.[111]

If the AI is a black box to the financial institution, it will often be impossible to probe a statement of opinion that states, for example: "Our valuations are statements of opinion and are based on state-of-the-art artificial intelligence, tested to industry standards, and provided all information that could possibly be relevant to a human appraiser." The implicit statement in that disclosure is that the maker of the statement used powerful technology to arrive at its opinion and that it subjectively believes the opinion is correct.[112]

Although that example from the Securities Act is illustrative, it is not difficult to imagine this problem occurring outside of securities law. The problem will arise anywhere the law requires justifications for conduct or the basis of a belief. Antitrust law, for example, often focuses on whether particular conduct has a legitimate economic or

[110] *Id.* at 1328–29 (footnotes omitted).

[111] This, of course, assumes that the AI was designed and tested appropriately, such that a belief that the AI was reaching appropriate valuations can be justified.

[112] As the Court in *Omnicare* pointed out, a statement that "we believe we are obeying the law," for example, would not be actionable simply because it turned out not to be the case — "a sincere statement of pure opinion is not an 'untrue statement of material fact,' regardless whether an investor can ultimately prove the belief wrong." 135 S. Ct. at 1327.

business reason motivating it.[113] A telltale sign of anticompetitive conduct is often that a monopolist's conduct makes no sense other than to harm a competitor.[114]

For instance, there is a justification test in antitrust law's refusal-to-deal jurisprudence.[115] Generally, the rule is that a firm has the unfettered discretion to choose whom it will deal with.[116] Indeed, there is no duty to deal with one's competitors.[117] The only exception appears in the Supreme Court's decision in *Aspen Skiing Co*. v. *Aspen Highlands Skiing Corp.*[118] There, the Supreme Court affirmed a jury verdict that the owner of three ski resort mountains with monopoly power unlawfully refused to deal with the owner of the fourth mountain.[119] The fourth mountain sought to purchase lift tickets for the defendant's three mountains to bundle it with its own lift tickets, but the three-mountain resort refused to sell its lift tickets even at full retail price.[120] It made no sense for the defendant, a monopolist, to refuse to sell its tickets to its competitor at its retail price. The only reason it would engage in such conduct would be to exclude its competitor from the market.[121] In other words, there were no legitimate business justifications for the conduct.

Courts since *Aspen Skiing* have held that refusals to deal are not actionable unless it can be proven that there is no legitimate business justification for the conduct or that the business justifications offered by a monopolist are a mere pretext.[122] Given this rule, if a monopolist's conduct was determined by an AI program with a decision-making process that is a black box to human beings, there would be no way to determine whether the monopolist's conduct was legitimate or anticompetitive.

[113] The inquiry into whether a party had legitimate justification for allegedly anticompetitive conduct is an integral part of the "rule of reason", which is applied to a broad swath of antitrust claims. FTC v. Actavis, Inc., 570 U.S. 136, 156 (2013).

[114] The most widespread example of such an inquiry in antitrust law is the "no economic sense test." For a detailed treatment of such tests in antitrust law, see generally Gregory J. Werden, *The "No Economic Sense" Test for Exclusionary Conduct*, 31 J. CORP. L. 293 (2006).

[115] *See, e.g.*, LePage's Inc. v. 3M, 324 F.3d 141, 163 (3d Cir. 2003) (considering whether actions had "valid business reasons").

[116] *See* Pac. Bell Tel. Co. v. Linkline Commc'ns., Inc., 555 U.S. 438, 448 (2009) (citing United States v. Colgate & Co., 250 U.S. 300, 307 (1919) ("As a general rule, businesses are free to choose the parties with whom they will deal, as well as the prices, terms, and conditions of that dealing.").

[117] *See id.*

[118] 472 U.S. 585 (1985).

[119] *See id.* at 611.

[120] *See id.* at 593–94.

[121] *See id.* at 608.

[122] *See, e.g.*, MM Steel, L.P. v. JSW Steel (USA) Inc., 806 F.3d 835, 845 (5th Cir. 2015) (affirming jury finding that the reasons for a refusal to deal were pretextual, thus giving rise to a finding that a horizontal conspiracy existed between distributers (citing Ross v. Standard Roofing, Inc. 156 F.3d 452, 478 (3d Cir. 1998)).

Consider the following hypothetical. A software manufacturer uses AI to remove poorly performing features from its upcoming product, which other developers rely on for their own products. If there is evidence that the software manufacturer intended to remove a feature solely to harm its competitor and the AI ultimately removes that feature as part of its performance sweep, how will one determine whether the removal of the feature was because of poor performance or a desire to impose costs on the monopolist's competitor?

The obvious question will be whether the AI was under instructions to remove that particular feature or whether the designers of the AI knew that the particular feature was a likely target. If the AI is a black box, then there is no way to tell. It may well have been that the AI determined that removing the particular feature would improve overall performance of the software. It may also be that the software was deployed because it would likely remove the feature that a competitor would rely on. Perhaps the AI was provided with the profitability resulting from removing various features and determined that removing features relied upon by a competitor happened to be the most profitable thing to do. None of these potential reasons can be verified by examining the AI ex post.

Nor would there be evidence that the monopolist knew ex ante what the AI would do. All the monopolist would have to do is point to the design of the computer program to argue that it was built to ferret out performance bottlenecks, and it had no idea what features the program would target. There were certainly no instructions to target the particular feature that is alleged to have been anti-competitively removed. Again, as with the Securities Act hypothetical, the law will insulate the creator of the AI from liability so long as he can demonstrate that it was designed for a particular purpose and was reasonably accurate and effective in accomplishing that purpose.

Throughout the law, we see the same legal construct — an intent test designed to determine what justifications an actor had for their conduct. These tests are easily bypassed when black-box AI, rather than a human being, is the actor or decision-maker.

D. AI and Gatekeeping Intent

Intent tests exist to limit the universe of claims that can be brought — they serve a gatekeeping function. For example, in *Washington v. Davis*, the Supreme Court held that a statute designed to serve neutral ends that has a discriminatory impact on a particular race is not unconstitutional unless there is evidence of discriminatory intent.[123] The Court reasoned that otherwise, a disparate-impact-only

[123] Washington v. Davis, 426 U.S. 229, 248 (1976).

test would invalidate various statutes "that may be more burdensome to the poor and to the average black than to the more affluent white."[124]

It is easy to see that such a rule breaks down entirely when state action is based on black-box AI.[125] Consider this hypothetical: a judge uses a sophisticated AI system as a tool to help him sentence criminals. The AI is designed to examine years of past sentencing decisions, the nature of the past crimes, and the attributes of past defendants who were convicted to provide a recommended sentence that reflects the likelihood of recidivism.[126] If there is any kind of bias in the past data that the AI uses to train, that bias may translate directly into a bias in its decisions.[127] For example, if a criminal defendant's zip code correlates highly with race and there is a history of racial bias in the sentencing data used to train the AI, then a zip code may become an outsized and outcome-determinative parameter for the AI.[128] The AI would then propagate the racial discrimination implicit in the data it learned from.

Someone seeking to challenge the decisions of a judge assisted by such an AI would only be able to point to the repeated decisions by that AI over time and show a racially disparate impact. Tests such as the one articulated in *Davis*, however, would require additional evidence of discriminatory intent that would be impossible to obtain. The judge would have no discriminatory intent if he followed the AI most

[124] *Id.*

[125] As Coglianese and Lehr acknowledge in their article defending the use of machine-learning algorithms for administrative regulation and adjudication, the discriminatory intent requirement would mean that constitutional challenges against AI-driven government or agency action would be difficult, if not impossible. *See* Coglianese et al., *supra* note 2 ("[A]lgorithms that include variables indicating protected class membership will seldom if ever trigger heightened scrutiny, at least in the absence of any explicit showing of discriminatory intent or animus.").

[126] This hypothetical is not far from what courts are currently considering. The Wisconsin Supreme Court, for example, recently held that the use of actuarial data to predict recidivism did not offend a defendant's due process rights, even though the data and methodology was not disclosed to the court or the defendant. *See* State v. Loomis, 88 N.W.2d 749 (Wis. 2016). There was nothing more than a written disclaimer to the judge about the dangers of the methodology, which as some commentators have noted, does little to inform a judge as to how much to discount the assessment of recidivism risk. Case Comment, *Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing*: State v. Loomis, 130 HARV. L. REV. 1530, 1534 (2017). The Court nonetheless accepted the warning as a reasonable safeguard.

[127] *See* ONE HUNDRED YEAR STUDY, *supra* note 1, at 10. ("[I]t remains a deep technical challenge to ensure that the data that inform AI-based decisions can be kept free from biases that could lead to discrimination based on race, sexual orientation, or other factors.").

[128] In other contexts, race-neutral variables have been shown to correlate with race, which has been a common refrain in response to disparate impact data. *See, e.g.*, Sean Hecker, *Race and Pretextual Traffic Stops: An Expanded Role for Civilian Review Board*, 28 COLUM. HUM. RTS. L. REV. 551, 568 (1997) ("'[R]ace' tends to correlate with race-neutral factors that police do find probative, such as 'nervousness,' out-of-state license plates, driving at less than the posted speed limit, and driving certain automobile models.").

of the time.[129] Moreover, the AI itself has no intent, and if it is a strong black box, there would be no way to determine what combination of parameters were dispositive in its sentencing decisions.[130]

Even if the AI is a weak black box and certain factors can therefore be ranked, the fact that a zip code or employment status is ranked higher than other parameters would not prove discriminatory intent. Indeed, they could be interpreted as merely proving the sort of economic aspects of the laws that would burden on-average poorer individuals more than the affluent.[131] More problematically, we may know that a particular parameter, such as zip code was ranked third or fourth most important overall, but where AI is making decisions in a non-linear way, a parameter such as employment status may be dispositive in one case, but not in other cases, depending on what other factors are present or not present.[132] Put simply, it may be impossible to know how important any particular parameter truly is to the AI's decision.[133]

The danger here is that a gatekeeping test put in place to draw a line may have worked well when dealing with humans, but when AI is involved, the test functionally immunizes the user of the AI from liability. It may also allow biases in data to propagate through the AI's decisions, potentially worsening the bias through a feedback loop. Again, intent tests leave AI conduct largely unregulated. With gatekeeping tests, there is the added problem that most cases may never reach discovery, meaning that an expert will not have occasion to analyze the AI's decisions (at least in the cases where some analysis of the AI's decision-making is possible).

[129] The cases that survive the *Davis* test will often be the ones where there is very clear evidence that a state actor intended to discriminate or where there is an extensive history of intentional discrimination. *See, e.g.*, Winfield v. City of New York, No. 15CV5236-LTS-DCF, 2016 U.S. Dist. LEXIS 146919, at *24–26 (S.D.N.Y. Oct. 24, 2016) (sustaining a complaint because a city had a history of discriminatory practices and officials had made statements "characteriz[ing] the policy in racial terms"); *see also* KG Urban Enters., LLC v. Patrick, No. 11-12070-NMG, 2014 U.S. Dist. LEXIS 2437, at *25–26 (D. Mass. Jan. 9, 2014) (holding that statements by government official were sufficient to establish prima facie case and shift the burden to the government). In the case of our sentencing judge, it is the AI that embodies past discrimination, so there will likely never be a statement by the sentencing judge targeting individuals of a particular race, just a disparate impact reflecting the one in the data.

[130] Indeed, the *Davis* test as applied to human actors already frequently results in the dismissal of Equal Protection claims or summary judgment in favor of the state actor because no discriminatory intent can be proven. *See, e.g.*, Lu v. Hulme, 133 F. Supp. 3d 312, 332 (D. Mass. 2015) (explaining that plaintiff's disparate impact claim fails as a matter of law because of a lack of evidence of discriminatory intent).

[131] *See* Washington v. Davis, 426 U.S. 229, 248 (1976).

[132] *See supra* Section II.C.2.

[133] *See id.*

IV. The Failure of Causation

Causation tests also fail when black-box AI is involved. Most causation tests are used to limit the scope of far-reaching causes of action, such as ordinary negligence. Doctrines such as proximate cause ensure that only reasonably foreseeable effects give rise to liability.[134] Such a doctrine encourages individuals to act reasonably and penalizes those who do not.[135] The proximate cause standard is thus a means of tying the scope of liability to the nature of the conduct at issue. Other related doctrines, such as reliance, require the injured to prove that the harm they suffered was related to the allegedly unlawful conduct.[136] Thus, a fraud claim will often fail unless the plaintiff can prove that the misrepresentation was something that the plaintiff took as true and that informed or caused the plaintiff to act to his detriment.[137] Loss causation, a doctrine that also appears in fraud-based claims, will also place similar limits on claims: only losses that stem from the alleged misrepresentation will be redressed.[138]

This section discusses these two types of causation. The first form of causation, which includes doctrines such as proximate cause, the Article will refer to as Conduct-Regulating Causation. This section explains that when AI is a black box, causation doctrines, such as proximate cause, fail because the causation inquiry will focus on what is foreseeable to the creator or user of the AI.[139]

The section also discusses what this Article refers to as Conduct-Nexus Causation, which ensures that the unlawful conduct and the resulting harm are sufficiently connected. This form of causation, which includes doctrines such as reliance and the causation element of Article III standing, is often predicated on the assumption that one can determine whether a defendant's conduct is connected to the alleged harm. When AI is a strong black box, this sort of inquiry is nearly impossible to undertake.[140] When AI is a weak black box, examina-

[134] *See* Owens v. Republic of Sudan, 864 F.3d 751, 794 (D.C. Cir. 2017); *see also* Palsgraf v. Long Island R.R., 162 N.E. 99, 104–05 (N.Y. 1928).

[135] *See* Mark F. Grady, *Proximate Cause Decoded*, 50 UCLA. L. Rev. 293, 322 (2002) ("The basic purpose of the reasonable foresight doctrine is to reduce the liability of people who may have been efficiently (reasonably, in a larger scheme of things) negligent.").

[136] *See* Basic Inc. v. Levinson, 485 U.S. 224, 243 (1988) ("Reliance provides the requisite causal connection between a defendant's misrepresentation and a plaintiff's injury.").

[137] *See, e.g.*, APA Excelsior III L.P. v. Premiere Techs., 476 F.3d 1261, 1271 (11th Cir. 2007) (finding no reliance where plaintiff could not have possibly relied on allegedly false statements in the registration statement).

[138] *See In re* Vivendi, S.A. Sec. Litig., 838 F.3d 223, 260 (2d Cir. 2016) ("Loss causation is the causal link between the alleged misconduct and the economic harm ultimately suffered by the plaintiff. In some respects, loss causation resembles the tort-law concept of proximate cause, which generally requires that a plaintiff's injury be the 'foreseeable consequence' of the defendant's conduct." (citations omitted)).

[139] *See supra* Section II.C.2.

[140] *See supra* Section II.D.

tion of a version of the AI as it existed at a particular point in time — a snapshot — may be the only way to find facts that can satisfy the required evidentiary burden.[141] This is because it is possible to obtain a ranking of the importance of data processed by a weak black box.[142] It may, however, still be impossible to determine if a particular type of data is outcome determinative for the AI.

A. Conduct-Regulating Causation

Causation tests exist in part for the purpose of setting the scope of liability. The most common of such tests is proximate cause, which allows courts to balance the breadth of a law's application against the administrative burdens of enforcing it.[143] Other fields of law have derived similar tests but with different names, which means that the same sort of causal analysis echoes throughout the law.[144] Fundamentally, proximate cause asks whether the result of the conduct was one that could have been foreseen by a reasonable person.[145] At its core is the assumption that a person should not be liable for results having nothing to do with what he could have done to limit the risk of harm, nor should there be liability for the flukes of chance.[146]

This sort of causation essentially asks the same question an Effect-Intent test asks — could the effect of the conduct have been foreseen? The reason this question is so critical is that the law seeks to deter behavior that causes harm to others or society, and holding individuals liable for effects they should have foreseen will encourage

[141] *See id.*

[142] *See id.*

[143] *See* Holmes v. Sec. Inv'r Prot. Corp., 503 U.S. 258, 268 (1992) ("At bottom, the notion of proximate cause reflects 'ideas of what justice demands, or of what is administratively possible and convenient.'" (quoting W. KEETON, D. DOBBS, R. KEETON, & D. OWEN, PROSSER AND KEETON ON LAW OF TORTS § 41, 264 (5th ed. 1984))); *see also* Palsgraf v. Long Island R.R., 162 N.E. 99, 103 (N.Y. 1928) (Andrews, J., dissenting) ("What we do mean by the word 'proximate' is, that because of convenience, of public policy, of a rough sense of justice, the law arbitrarily declines to trace a series of events beyond a certain point.").

[144] For example, as the Supreme Court has recognized, the antitrust standing analysis is similar to proximate cause analysis. *See* Associated General Contractors v. Cal. State Council of Carpenters, 459 U.S. 519, 535–36 (1983). Indeed, since *Associated General Contractors*, courts have held that proximate cause is an element of antitrust injury. *See, e.g.*, *In re* Aluminum Warehousing Antitrust Litig., 833 F.3d 151, 162 (2d Cir. 2016). The Supreme Court has likewise held that proximate cause is required for certain RICO claims. *See* Holmes v. Sec. Inv'r Prot. Corp., 503 U.S. 258, 270 (1992). A proximate cause requirement is also an element of a securities fraud claim. *See* Dura Pharms., Inc. v. Broudo, 544 U.S. 336, 346 (2005).

[145] See Owens v. Republic of Sudan, 864 F.3d 751, 794 (D.C. Cir. 2017); *Palsgraf*, 162 N.E. at 104–05.

[146] *See* Grady, *supra* note 135, at 294 ("Probably the most obvious limitation is that a person should not be liable when the only connection between his lapse and the plaintiff's injury was the purest chance, a total coincidence.").

them to take precautions (or perhaps discourage them from risky behavior that would cause injury).[147]

In the case of black-box AI, the result of the AI's decision or conduct may not have been in any way foreseeable by the AI's creator or user. For example, the AI may reach a counter-intuitive solution, find an obscure pattern hidden deep in petabytes of data, engage in conduct in which a human being could not have engaged (e.g., at faster speeds), or make decisions based on higher-dimensional relationships between variables that no human can visualize.[148] Put simply, if even the creator of the AI cannot foresee its effects, a reasonable person cannot either. Indeed, if the creator of AI cannot necessarily foresee how the AI will make decisions, what conduct it will engage in, or the nature of the patterns it will find in data, what can be said about the reasonable person in such a situation?

Already, with the first generation of algorithms, used principally in the financial markets over the last few years, the effects of the algorithms' conduct have been highly unpredictable. Flash crashes, for example, were in most cases difficult to predict due to algorithms trading rapidly.[149] More importantly, the speed to which contagion may spread throughout the markets as a result of algorithms interacting with other traders (and with other algorithms) have caused impacts on prices with magnitudes beyond what anyone likely could have predicted.[150] As some commentators have noted, this unpredictability makes it very unlikely that the law can appropriately encourage or deter certain effects, and more problematically, the failure of our legal structures will allow people using the algorithms to externalize costs to others without having the ability to pay for the injuries they inflict.[151]

The inability to foresee harm is even greater with black-box AI because there is little or no ability to foresee how the AI will make decisions, let alone the effects of those decisions. As with the hypothetical of the AI-assisted sentencing judge in Part II, the AI may serve to perpetuate biases that exist in the past.[152] If the creator or user of the AI cannot ex ante predict the nature or the extent of the effect of the AI's conduct or decisions, the tuning function of a Conduct-Regulating Causation test fails entirely because the scope of liability no longer reflects the sort of precautionary measures or risk calculus

[147] *See id.*

[148] *See supra* Section II.C.2.

[149] *See supra* note 84 and accompanying text.

[150] *See* Felix Salmon & Jon Stokes, *Algorithms Take Control of Wall Street*, WIRED (Dec. 27, 2010 12:00 PM), https://www.wired.com/2010/12/ff_ai_flashtrading/ [https://perma.cc/B4DF-JPJM].

[151] *See* Yadav, *supra* note 75, at 1039, 1083.

[152] *See supra* Section III.C.

the law expects of a reasonable person. Put simply, the causation test becomes an arbitrary cutoff for liability.

B. Conduct-Nexus Causation

Another form of causation that breaks down is the class of causation test that examines the ties between the allegedly unlawful conduct and the harm being redressed. There are examples of such tests throughout the law. The doctrine of reliance, for example, requires that the harm suffered relate to the alleged conduct or misstatement by the defendant.[153] Likewise, the test for Article III standing contains a requirement that the alleged injury suffered be "fairly traceable" to the allegedly unlawful conduct.[154] These tests all examine the nexus between the allegedly unlawful conduct and the harm. Unlike Conduct-Regulating causation, which sets the scope of liability, nexus tests serve a gatekeeping function. As this section argues (using reliance and Article III standing as illustrative examples), these causation tests also break down when applied to black-box AI.

1. Reliance

Consider a large institutional investor that uses black-box AI to make decisions about which privately traded securities to purchase. Assume that one of the companies it invests in made misstatements about the progress of a key research and development effort it has undertaken. In fact, the research and development project was a sham, with virtually no product in the pipeline. The first argument the institutional investor will encounter is that there was no reliance on any of the statements about the research and development project. The investor suing the issuer may be able to point to the sort of data it fed to its AI, and it may be able to argue that some of that data was somehow dependent on the research and development project's existence and progress (e.g., the AI relied on research and development expenses in the company's financials). But, if the burden of proving reliance is on the institutional investor, how will it satisfy that burden?

If the institutional investor does not retain a snapshot of the AI it used to make its decisions, it cannot hand it over to an expert to run experiments on it — for example, to change inputs to the AI and determine the effect on its decisions, in order to probe the rules the AI

[153] *See* Basic Inc. v. Levinson, 485 U.S. 224, 243 (1988).

[154] *See* Bank of Am. Corp. v. City of Miami, 137 S. Ct. 1296, 1302 (2017) ("To satisfy the Constitution's restriction of this Court's jurisdiction to 'Cases' and 'Controversies,' Art. III, § 2, a plaintiff must demonstrate constitutional standing. To do so, the plaintiff must show an 'injury in fact' that is 'fairly traceable' to the defendant's conduct.") (quoting Spokeo v. Robins, 136 S. Ct. 1540, 1547 (2016)).

had established for itself.[155] If the AI is a strong black box, then it will in any event be impossible to tell whether any information relating to the R&D project was given any weight at all.

One may argue that the fact that the AI took into account information that somehow depended on the existence of the R&D project should be enough to establish reliance, but that overlooks the possibility that the AI may have attached absolutely no weight to the information in this particular case.[156] It is also possible that in almost every possible case, that information would never be outcome determinative to the AI.[157] Under such circumstances, it is difficult to justify a finding of reliance on the alleged misrepresentation.

If a human being had made the investment decision, it would be a task familiar to courts and regulators to take evidence on reliance. Witnesses would be interviewed or deposed and ultimately testify at trial, and e-mails and documents would be produced and analyzed. It is clear that a strong black box, however, cannot be interrogated. Its decision-making process cannot be audited. This means reliance would be nearly impossible to prove unless the reliance standard is significantly relaxed.[158]

If the AI is a weak black box, it may be possible to prove reliance because a loose ranking of the parameters fed to the model is available for analysis.[159] However, the investor would still face arguments that there is no way of proving that a particular piece of information would generally be outcome-determinative. This sort of argument can only be overcome if there is a version of the AI as it existed at the time the investment decision was made that can be analyzed, likely by an expert witness.

[155] *See supra* Section II.D.

[156] A quick examination of the simplistic equation *supra* in Section II.B makes clear how this is possible. One of the coefficients for a particular feature (variable) may be set to a very low number — or perhaps even 0. In such a case, the model's decision would virtually ignore the contributions from that variable. With multi-layered neural networks, however, it is often not as simple as examining a single weight — data may be weighed differently depending on the existence, absence, or degree of other factors.

[157] Consider a non-linear version of the SVM described *supra* in Section II.C.2. The combination of one feature at a particular value coupled with dozens of other features at particular values may have one outcome, and the slight modification of any one of those features may change the model's decision entirely. Because of dimensionality, it is virtually impossible in most cases for humans to visualize how such a non-linear model has reached a dividing boundary in higher-dimensional space. *See supra* note 62.

[158] One traditional way to relax the reliance requirement is to shift the burden, but that will only shift the problem posed by the black-box AI to the other party. In such a case, the defendant bears the burden of rebutting reliance and would face the same intractable evidentiary burden when black-box AI is involved.

[159] *See supra* Section II.D.

2. Article III Standing

Another form of nexus test that breaks down appears in federal standing jurisprudence. To have standing, a plaintiff must show, *inter alia*, injury in fact that is "fairly traceable" to the conduct alleged to be unlawful.[160] This standing requirement ensures that the alleged injury flows from the conduct of the defendant before the court.[161] The causation doctrine in the Article III standing inquiry first emerged in a Supreme Court decision denying standing to indigent plaintiffs suing the IRS over a regulation governing the tax-exempt status of hospitals that refused to provide anything beyond emergency services to indigent patients; no hospital was party to the lawsuit.[162] The Court stated that the "'case or controversy' limitation of Art. III still requires that a federal court act only to redress injury that fairly can be traced to the challenged action of the defendant, and not injury that results from the independent action of some third party not before the court."[163]

In denying standing, the Court reasoned that it was not clear that the injury plaintiffs suffered could be traceable to the IRS's revenue ruling.[164] By the time the Supreme Court decided the seminal *Lujan v. Defenders of Wildlife*,[165] the causation requirement had solidified — there would have to be "a causal connection between the injury and the conduct complained of."[166]

This Conduct-Nexus test is part of a constitutional inquiry and sits at the threshold of every federal claim. It is thus unsurprising that black-box AI has the potential to cripple the ability of courts to assess standing in a large swath of federal claims. To begin with, a strong black box will be nearly impossible to probe for causation. A deep neural network may be impossible to audit to determine what information it found outcome-determinative or how it is making decisions. Even if the AI is a weak black box, the intensive expert-driven audit necessary to establish causation will not likely occur at the onset of

[160] Bank of Am. Corp. v. City of Miami, 137 S. Ct. 1296, 1302 (2017). In addition to injury and fact and causation, a plaintiff must also show that a judicial decision is capable of redressing the alleged injury. *Id.* Standing may also depend on prudential standing requirements, which are not the subject of this section. *See, e.g.*, Lexmark Int'l, Inc. v. Static Control Components, Inc., 134 S. Ct. 1377, 1387 (2014) (stating prudential standing may require a court to determine "whether a legislatively conferred cause of action encompasses a particular plaintiff's claim" based on whether, for example, the plaintiff's claim falls within the "zone of interests" of the statute).

[161] Town of Chester v. Laroe Estates, Inc., 137 S. Ct. 1645, 1650 (2017) (noting that Article III standing requires plaintiff to have "suffered an injury in fact" which "is fairly traceable" to the defendant's conduct).

[162] *See* Simon v. E. Ky. Welfare Rights Org., 426 U.S. 26 (1976).

[163] *Id.* at 41–42.

[164] *See id.*

[165] 504 U.S. 555 (1992).

[166] *Id.* at 560.

litigation when Article III standing is first rigorously assessed. This is because that assessment is often made under a pleading-based standard (such as that of a motion to dismiss) when no discovery or evidence is available or appropriately considered.

To see how this problem may arise, consider the following hypothetical. A federal agency uses an AI program that is a strong black box to allocate oil drilling rights on federally held land.[167] Assuming that the federal agency promulgates regulations based on the optimization done by the AI, a plaintiff challenging such a regulation may face a significant constitutional hurdle. For example, if a plaintiff owning a plot of land adjacent to a federal plot of land licensed to an oil company and the plaintiff sustains damage to his property because of the agency's regulations, the plaintiff will have to show that the damage sustained was fairly traceable to the agency's regulation. If the regulation is based on the AI's opaque optimization of a host of variables, the plaintiff will bear the burden of proving causation at the onset of litigation when Article III standing is first rigorously assessed.

This sort of hypothetical makes it clear how difficult it is for a plaintiff to challenge regulations based on an AI program's decisions. The broader point is that because causation is a constitutional requirement that must be met for every federal claim,[168] claims involving opaque AI may fail at the very onset of litigation, and because Article III standing is a question of subject matter jurisdiction, an inability to prove causation at later points in litigation — indeed, at any point in the litigation — may deprive a court of jurisdiction entirely.[169]

V. The Problems with Transparency Standards and Strict Liability

To some, the obvious solution to the intent and causation problems posed in this Article will be to either increase the transparency of AI through standards-based regulation or to impose strict liability. This section argues that both approaches would risk stifling innovation in AI and erecting steep barriers to entry.

[167] Federal agencies are beginning to use machine-learning algorithms in connection with their regulatory functions. For a detailed analysis of the use of machine-learning algorithms by administrative agencies, see generally Coglianese et al., *supra* note 2.

[168] Warth v. Seldin, 422 U.S. 490, 498 (1975) (noting that Article III standing is a "threshold question in every federal case, determining the power of the court to entertain the suit").

[169] Challenges to subject matter jurisdiction can be raised at any time in the litigation, and, if successful, require dismissal of the action. *See* FED. R. CIV. P. 12(h)(3).

A. Transparency Regulation

Although one possible way to alleviate the Black Box Problem is to regulate the minimum transparency required for AI, such regulation would be problematic for several reasons.

1. Transparency Is a Technological Problem

It is tempting to think of AI transparency as a problem akin to those addressed by environmental and securities regulations. The securities laws seek to create fairer, more transparent markets by requiring disclosures and registrations with a federal agency.[170] The environmental laws are rife with granular regulations that impose standards, such as minimum and maximum levels of particular chemicals that can be emitted.[171] AI, however, may be too qualitatively different to regulate in this way.

To begin with, it is not clear that certain forms of AI that are based on complex machine-learning algorithms, such as deep neural networks, will become more auditable and transparent in the future. In fact, it may be that as these networks become more complex, they become correspondingly less transparent and difficult to audit and analyze.[172] Indeed, commentators have speculated that AI may eventually become significantly more intelligent than human beings, such that they will surpass the analytical abilities of humans altogether.[173] If this is the trajectory of AI, then it makes little or no sense to impose regulations requiring minimum levels of transparency. It may be that certain technology may never meet the ideal levels of transparency desired by regulators and governments. If the improvement of AI requires, for example, more complexity that will cause a further lack of transparency, imposing transparency requirements will be tantamount to a prohibition on improvement or an invitation for companies to circumvent the rules.

[170] *See* Geoffrey A. Manne, *The Hydraulic Theory of Disclosure Regulation and Other Costs of Disclosure*, 58 ALA. L. REV. 473, 475 (2007). The Securities and Exchange Commission frequently articulates the value of disclosure as the basis for its rules. *See, e.g.*, Disclosure of Accounting Policies for Derivative Financial Instruments and Derivative Commodity Instruments and Disclosure of Quantitative and Qualitative Information About Market Risk, Exchange Act Release No. 33-7386, 62 FED. REG. 6044, 6048 (1997) ("To address this comparability issue, registrants are required to disclose the key model characteristics and assumptions used in preparing the quantitative market risk disclosures. These disclosures are designed to allow investors to evaluate the potential impact of variations in those model characteristics and assumptions on the reported information.").

[171] *See, e.g.*, Cmtys. for a Better Env't v. Envtl. Prot. Agency, 748 F.3d 333, 335 (D.C. Cir. 2014) (noting specific parts-per-million guidelines).

[172] *See* GOODFELLOW ET AL., *supra* note 34, at 1–2.

[173] *See, e.g.*, RAY KURZWEIL, THE SINGULARITY IS NEAR, WHEN HUMANS TRANSCEND BIOLOGY 8 (2005).

2. Regulatory Influence Over Design

Regulating the minimum levels of transparency for AI would at least implicitly be a regulation of design trade-offs. AI designers deciding whether to increase the size and depth of a neural network (thereby losing transparency) may be forced to use a shallower or less complex architecture to comply with regulations, even if such a design decision would result in poorer performance. This essentially makes regulators and legislators arbiters of design — a function that regulators are not only less likely to be proficient at than AI developers, but are also reluctant to perform.[174]

3. Barriers to Entry

Finally, a complex system of regulation would impose significant costs on new entrants into AI markets.[175] Already, AI talent is concentrated in the hands of a few large firms.[176] Imposing the cost of compliance with a byzantine system of regulations would ensure that only large firms could afford to comply with them.[177]

None of this is to say that the regulation of AI would necessarily be a mistake. Indeed, there may be a need to regulate the extent to which firms can externalize risks through the use of AI, the extent of care in design that must be taken, or the constraints on the AI's conduct that must be imposed.[178] All of this may require the same sort of regulatory apparatus that appears in other contexts (i.e., the establishment of an administrative agency that promulgates granular rules or imposes disclosure requirements). Nevertheless, the degree of transparency that AI must exhibit should not be codified into a set of regulatory standards. It may well be that the most powerful AI is much

[174] In the antitrust context, for example, courts confronting allegedly anticompetitive product redesigns will examine the anticompetitive effect of the redesign, rather than the merits of the redesign itself. *Cf.* New York v. Actavis PLC, 787 F.3d 638, 652 (2d Cir. 2015).

[175] *See generally* Chester S. Spatt, *Complexity of Regulation*, HARV. BUS. L. REV. ONLINE (June 16, 2012), http://www.hblr.org/2012/06/complexity-of-regulation/ [https://perma.cc/SPQ6-SE6X]. In addition to direct costs, indirect costs may impede entry and innovation. *See, e.g.*, John C. Coates, *Cost-Benefit Analysis of Financial Regulation*, 124 YALE L.J. 882, 930 (2015) (noting that the Sarbanes-Oxley Act can be criticized for imposing indirect costs, such as "potential reductions in risk-taking, dilution in strategic focus, and the opportunity costs of devoting excessive management time to compliance").

[176] *See* Metz, *supra* note 4.

[177] Courts have recognized that regulatory compliance costs can serve as potential barriers to entry for the purposes of the antitrust laws. *See, e.g.*, Novell, Inc. v. Microsoft Co., 731 F.3d 1064, 1071 (2013).

[178] This task may nonetheless be exceedingly difficult for a regulator because measuring externalities may prove difficult, and in the financial setting, interconnected markets further complicate the ability to predict or measure externalized harm. *See* Coates, *supra* note 175, at 894.

like the human brain — it has an exceptional ability to learn, but its knowledge and experience may be intractably hard to communicate or transfer.[179]

B. Strict Liability

Another seemingly viable option is to use a strict liability standard to alleviate the intent and causation problems that arise from black-box AI. Strict liability is already pervasive in tort law and is used to deter and punish the most dangerous classes of behavior.[180]

Commentators that have considered strict liability regulation for non-intelligent algorithms have dismissed such an approach; this rejection of strict liability is even more justified in the case of AI.[181] This is because strict liability only makes sense if the creator of a computer program can anticipate the program's harmful effects ahead of time and adjust the program accordingly. As computer programs become more intelligent and less transparent, not only are the harmful effects less predictable, but their decision-making process may also be unpredictable.[182] Strict liability, however, assumes some control or predictability, which would allow the AI developer to, for example, predict the potential injury for which it will be liable so that it can obtain adequate insurance.[183]

Moreover, while in the products liability context, there is a fair assumption that the designer of the product is in the best position to control the aspects of the product that may cause injury, no such assumption may be warranted in the case of AI. Indeed, the designer of a product has a lot of data about how often the product will cause injury and the severity of those injuries, so the product designer is in the best position to avoid any potential injury.[184] If sued, what the design-

[179] *See* Castelvecchi, *supra* note 9.

[180] *See* RESTATEMENT (SECOND) OF TORTS § 519(a) (AM. LAW INST. 1965) ("One who carries on an abnormally dangerous activity is subject to liability for harm to the person, land or chattels of another resulting from the activity, although he has exercised the utmost care to prevent the harm.").

[181] *See* Yadav, *supra* note 75, at 1039 (arguing that strict liability is a poor fit for the regulation of algorithms because, *inter alia*, "[p]redictive programming implies an endemic propensity for ad hoc, unpredictable error, meaning that strict liability can give rise to widespread breaches").

[182] *See id.* at 1083 ("[E]rrors can arise even if traders take every care in trying to assure the safe operation of their algorithms. They can happen without warning or foreseeability. This poses a conceptual difficulty for traders seeking to avoid liability by designing their algorithms to minimize problems.").

[183] *See, e.g.*, Todd v. Societe BIC, S.A., 9 F.3d 1216, 1219 (7th Cir. 1993) ("Some products are dangerous even when properly designed, and it is both easier and cheaper for consumers to obtain their own insurance against these risks than to supply compensation case-by-case through the judicial system.").

[184] *See* Lovell v. P.F. Chang's China Bistro, Inc., No. C14-1152RSL, 2015 U.S. Dist. LEXIS 112101, at *16 (W.D. Wash. Mar. 27, 2015) ("The theory underlying this type of strict liability is that the manufacturer, seller, and/or distributor is in the best position to

er of a defective product knew or should have known can be discovered and presented to a fact finder. None of this is necessarily true with black-box AI.[185]

Finally, strict liability may impose significant barriers to entry. It may simply be too costly, unpredictable, or difficult to produce and deploy AI without risking potentially ruinous liability. The possibility of unpredictable liability would therefore, like a byzantine regulatory structure, provide significant barriers to entry in most markets where there are already large players.[186] Companies that have large balance sheets may be willing to develop and deploy AI and take the risk of strict liability,[187] but new entrants may not dare to do so.

At bottom, it may be that strict liability works well in some settings, particularly where the risk of loss is great (as is the case in tort law),[188] but a blanket strict liability standard would risk a significant chilling of innovation and an increase in long-term market concentration when applied to AI.

VI. A SUPERVISION-TRANSPARENCY APPROACH

This Section sets forth a sliding-scale approach to adapting intent and causation tests that depends on the degree to which the AI is (a) permitted to operate autonomously, and (b) transparent. As this section will argue, AI supervised by humans will pose the least problems for intent and causation tests, whereas autonomous AI will require liability schemes based on negligence, such as those used in agency law for the negligent hiring, training, or supervision of an agent. When the AI operates under human supervision,[189] the degree of transparency may shed light on the creator or user of the AI's intent. When the AI is permitted to operate autonomously, the creator or user of the AI should be held liable for his negligence in deploying or testing the AI. In the most dangerous settings, strict liability may be ap-

know of the dangerous aspects of the product and to translate that knowledge into a cost of production against which liability insurance can be obtained." (internal quotation marks and citation omitted)); *see also* RESTATEMENT (SECOND) OF TORTS § 402A, cmt. c..

[185] *See supra* Section II.D.

[186] *See, e.g.*, Michael D. Stovsky, Comment, *Product Liability Barriers to the Commercialization of Biotechnology*, 6 HIGH TECH. L.J. 363, 379–80 (1991) (arguing that strict liability regimes impose barriers to entry in biotechnology markets).

[187] In consumer-facing settings, the size and structural market power of a firm may signal to a consumer that a firm can pay for, or distribute the cost of, any injury caused by product failure or that it possesses insurance to cover those injuries. For a discussion of the assumptions underlying strict liability rules, see generally Alan Schwartz, *The Case Against Strict Liability*, 60 FORDHAM L. REV. 819 (1992). *See also* Fleming James, Jr., *Some Reflections on the Bases of Strict Liability*, 18 LA. L. REV. 293, 296 (1958).

[188] This question will be explored in more detail in Part VI, *infra*.

[189] References to "supervised" or "unsupervised" in this section refer to the degree of human involvement and oversight, not the technical distinction between AI that is trained on labeled training sets and AI that is given unlabeled data. *See* FLACH, *supra* note 40, at 14.

propriate. The overall picture is a sliding scale of intent and foreseeability required for liability.

A. The Supervised Case

AI that is supervised by a human is unlikely to pose significant problems for traditional intent and causation tests. For example, a human that consults AI to make decisions, such as a judge that consults AI to assist with sentencing decisions,[190] ultimately makes decisions himself. The AI may assist with the decision-making process, but the responsibility for the decision lies largely with the human decision-maker. In such a case, the transparency of the AI will inform intent and causation inquiries. For example, if a human relies on AI that is fully transparent, then he can determine how the AI is making its decisions and will be able to foresee the effect of the AI's decisions. Intent and causation tests will therefore properly apply because the foreseeable consequences of the AI's decisions will be ascertainable.

When AI is a black box, the degree of transparency bears directly on the intent of a human who makes decisions based on the AI. For example, blind reliance on AI that engages in a decision-making process that the human cannot understand and that may have effects that the human cannot foresee may be evidence of unlawful intent such as scienter or willful blindness.[191] The extent to which the AI is a black box thus bears on the human's intent. In such cases, courts and regulators will not need to look to the design of the AI or the foreseeable effects of the AI to determine liability. The central question in such cases will be the degree of the AI's transparency and the culpability or reasonableness of the human's reliance on the black-box AI.

For example, consider the earlier hypothetical of a large financial institution's use of AI to appraise homes for use as collateral for mortgage-backed loans that will ultimately be packaged into mortgage-backed securities. The financial institution would be liable to an investor who relied to his detriment on an inaccurate appraisal value if the financial institution acted unreasonably in relying on the output of

[190] *See supra* Section III.C.

[191] The user of a trading AI that, for example, frequently places and then cancels orders, may lead the trader using the AI to suspect that it may be spoofing and in such a case he may be considered willfully blind to the spoofing if he does not then monitor or limit the AI's conduct. *See* Global-Tech Appliances, Inc. v. SEB S.A., 563 U.S. 754, 769 (2011) (noting that despite Circuit differences, "all appear to agree on two basic requirements: (1) the defendant must subjectively believe that there is a high probability that a fact exists and (2) the defendant must take deliberate actions to avoid learning of that fact."). Likewise, the same conduct may amount to a reckless disregard of the truth, which can give rise to the inference of scienter. *See, e.g.*, Universal Health Servs. v. United States ex rel. Escobar, 136 S. Ct. 1989, 1996 (2016) (concluding that scienter can be proven under the False Claims Act when a person "acts in reckless disregard of the truth or falsity of the information").

the black-box AI. Perhaps the institution ignored an unreasonable result or apparent bias in the AI's input data, or failed to put in place reasonable safeguards or testing regimes. To continue to rely on AI that may be making flawed decisions or that is relying on problematic data may be evidence of willful blindness or may arise to the level of recklessness required for scienter.

B. The Autonomous Case

In the autonomous case, agency law is instructive. The industrial revolution brought with it difficult problems for agency law, many of which stemmed from the independence and lack of direct supervision of agents.[192] One of the doctrines created to deal with the problem of an agent employed with the general task of accomplishing a principal's goals is *respondeat superior*,[193] a form of vicarious liability that holds a principal liable for the conduct of an agent he employs. The impetus for the doctrine is, as Blackstone noted in his commentaries, that a principal's use of an agent is tantamount to a "general command" to accomplish the principal's goals.[194] As Blackstone observed, under that rule, if a "drawer at a tavern sells a man bad wine, whereby his health is injured, he may bring an action against the master" because "although the master did not expressly order the servant to sell it to that person in particular" there was "impliedly a general command" to sell the wine.[195]

Early *respondeat superior* and vicarious liability cases struggled with the intentional torts of an agent. In those cases, it was the conduct of the agent that was performed with the requisite intent for the tort, and there was no obvious reason to impute that intent to the principal absent some evidence of his assent to the agent's conduct.[196] Ultimately, the cases led to the modern rule, which embodies broad vicarious liability that covers even the intentional torts of an agent if the agent's conduct was within the scope of his employment.[197] In

[192] *See* Richard R. Carlson, *Why the Law Still Can't Tell an Employee When It Sees One*, 22 BERKELEY J. EMP. & LAB. L. 295, 304 (2001).

[193] *See* RESTATEMENT (THIRD) OF AGENCY § 2.04 (AM. LAW INST. 2006) ("An employer is subject to liability for torts committed by employees while acting within the scope of their employment.").

[194] 1 WILLIAM BLACKSTONE, COMMENTARIES *73, *77. Blackstone referred to the master-servant relationship, but the principal-agency construct is functionally the same for the purposes of this Article's discussion. I accordingly use principal, master, and employer interchangeably. Likewise, I also refer interchangeably to agents, servants, and employees.

[195] *Id.*

[196] *See, e.g.*, Wright v. Wilcox, 19 Wend. 343 (N.Y. Sup. Ct. 1838).

[197] *See* Burlington Indus. v. Ellerth, 524 U.S. 742, 756 (1998) ("While early decisions absolved employers of liability for the intentional torts of their employees, the law now imposes liability where the employee's 'purpose, however misguided, is wholly or in part to further the master's business.'" (quoting W. KEETON, et al., *supra* note 143)).

addition, agency law also developed causes of action against principals for the negligent hiring, training, and supervision of an agent.[198] Thus, even in the cases when the agent's intentional conduct cannot be imputed to the principal, the principal's negligence may nevertheless give rise to liability.

AI is a new and unprecedented form of agent. When it operates autonomously, it is indistinguishable in some cases from a human being tasked with meeting some objective. Just as a human may behave in an unpredictable manner, AI may also arrive at solutions or engage in conduct that its user or creator never foresaw, particularly when the AI is a black box. Notwithstanding the similarities between an AI and a human agent, a vicarious liability rule, such as *respondeat superior*, would make sense only in certain circumstances. When the AI operates autonomously in a mission-critical setting or one that has a high possibility of externalizing the risk of failure on others, such as when it is used in a highly interconnected market or to perform a medical procedure, the AI's user or creator should be more broadly liable for injury the AI inflicts, and a vicarious liability rule is appropriate. In such cases, a lack of transparency should not insulate the user or creator of the AI from liability. Instead, the risks of deploying a black-box AI autonomously in such settings should fall on the AI's user or creator because the AI was used notwithstanding its unpredictable and impenetrable decision-making. In such a case, the imposition of vicarious liability would be functionally equivalent to a strict liability regime.

When the AI is deployed autonomously in less dangerous or mission-critical settings, a vicarious liability rule may be less appropriate. There may be little risk of harm from the AI's error in these circumstances and holding the user or creator of the AI liable regardless of intent or negligence would chill a large swath of desirable AI applications. Instead, in such cases, the negligent principal rule would be more appropriate. When the AI is transparent, knowledge about how the AI's decision-making process works may be used to establish the existence of or lack of reasonable care. When, however, the AI is a black box, the deployment of the AI in the face of a lack of transparency may be sufficient to establish a lack of reasonable care. The question is similar to the one asked in the agency setting — whether the AI's creator or user was negligent in deploying, testing, or operating the AI. The use of the AI in the face of a lack of transparency bears heavily on that question. In the case where there is a lack of transparency, proximate cause tests should focus on the possible effects of deploying AI autonomously without understanding how it functions, rather than on the specific ability of the user or creator of

[198] RESTATEMENT (THIRD) OF AGENCY § 7.05.

the AI to have predicted the injurious effects of the AI's conduct. Consider the previous example of the AI that re-tweeted false or misleading information, was it reasonable for the creator of the AI to give the program the ability to create its own tweets? The focus in such a case would be on whether the risk of the potentially unlawful conduct was apparent or should have reasonably been addressed with precautions.

C. A Sliding-Scale Approach

Putting the supervised and autonomous cases together, one can imagine four quadrants of liability. First, when there is both supervision of the AI and the AI is transparent, then the intent of the creator or user of the AI can be assessed through conventional means (i.e., fact-finding mechanisms such as depositions and subpoenas) as well as by examining the AI's function and effect. Second, when the AI is supervised but to some degree a black box, intent must be assessed based on whether the creator or user of the AI was justified in using the AI as he did — with limited insight into the AI's decision-making or effect. Third, if the AI is autonomous but supervised, the rule that should apply is the principal-supervision rule from agency law. The question will be whether the creator or user of the AI exercised reasonable care in monitoring, constraining, designing, testing, or deploying the AI. Fourth, when the AI is both autonomous and unsupervised, the sole question will be whether it was reasonable to have deployed such AI at all. The answer may simply be no, which means that the creator or user of the AI would be liable for the AI's effects, even if he could not foresee them and did not intend them.

Table 1 represents a general sketch of this sliding scale approach. Its implementation will require a full-scale revision of a wide range of laws. Indeed, the categories of intent and causation tests discussed *supra* may themselves require entirely different modifications.

Table 1: These four quadrants of liability provide the contours of a sliding-scale approach.

	Transparent	**Black Box**
More Supervision	Traditional intent and causation tests can be applied	Use without transparency bears on the intent of the creator or user of the AI and the foreseeability of the harm caused by the AI
Less Supervision	Relaxed intent and causation; negligent principal standard	Broad scope of liability; creator or user of the AI bears the risks stemming from the AI's lack of transparency

1. Effect Intent and Gatekeeping Intent Tests

Liability rules that employ Effect Intent or Gatekeeping Intent tests will require a threshold inquiry into the autonomy of the AI. In cases where the AI operates autonomously, the intent tests should be relaxed to allow for evidence of negligence. In other words, specific intent should not be required for liability, nor should such tests be used to narrow the scope of potential claims when autonomous AI is involved. Thus, as with the discriminatory intent test used in *Washington v. Davis* (discussed *supra*, Section III.D), the question should be whether the government was sufficiently negligent in its training, testing, or deployment of the AI that it would warrant that a plaintiff be given further discovery and ultimately whether constitutional scrutiny (e.g., rational basis or strict scrutiny) is appropriate.

2. Basis Intent Tests

In cases where a basis for conduct must be articulated, such as in antitrust, securities, or constitutional cases, the use of black-box AI should be prima facie evidence that, apart from the past accuracy of the AI, the user or creator of the AI lacked a sufficient justification for a given course of conduct. This can be accomplished through burden shifting. In other words, in cases involving autonomous AI that lack transparency, the burden should be on the proponent of the AI to prove that the AI's conduct or decisions were justified. Such a test would encourage a human check on the AI's decisions and conduct.

3. Conduct-Regulating and Conduct-Nexus Causation Tests

Proximate cause tests should assess the level of human supervision at the outset. It is only in the autonomous case that foreseeability of harm will be exceptionally difficult to assess. In such cases, the question should be focused on the foreseeability of harm from deploying AI autonomously given its degree of transparency. Thus, the causation analysis for AI that falls into the Black Box / Less Supervision quadrant above should turn not on whether the particular harm caused by the AI was reasonably foreseeable, but whether the harm was a foreseeable consequence of deploying black-box AI autonomously. The question is one of conceivability, not foreseeability in such settings.

Ultimately, how this sliding scale scheme should be implemented is highly fact-dependent. What is clear, however, is that both transparency and autonomy will be central questions in most cases. The most important task will be to build these questions into the intent and causation structures that already exist throughout the law.

VII. CONCLUSION

Modern AI systems are built on machine-learning algorithms that are in many cases functionally black boxes to humans. At present, it poses an immediate threat to intent and causation tests that appear in virtually every field of law. These tests, which assess what is foreseeable or the basis for decisions, will be ineffective when applied to black-box AI.

The solution to this problem should not be strict liability or a regulatory framework of granularly defined transparency standards for AI design and use. Both solutions risk stifling innovation and erecting significant barriers to entry for smaller firms. A sliding scale system is a better approach. It adapts the current regime of causation and intent tests, relaxing their requirements for liability when AI is permitted to operate autonomously or when AI lacks transparency, while preserving traditional intent and causation tests when humans supervise AI or when the AI is transparent.