Forthcoming VAND. L. REV. (2023); comments welcome!

HUMANS IN THE LOOP

Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II[*]

From lethal drones to cancer diagnostics, complex and artificially intelligent algorithms are increasingly integrated into decisionmaking that affects human lives, raising challenging questions about the proper allocation of decisional authority between humans and machines. Regulators commonly respond to these concerns by putting a "human in the loop": using law to require or encourage including an individual within an algorithmic decisionmaking process.

Drawing on our distinctive areas of expertise with algorithmic systems, we take a bird's eye view to make three generalizable contributions to the discourse. First, contrary to the popular narrative, the law is already profoundly (and problematically) involved in governing algorithmic systems. Law may explicitly require or prohibit human involvement and law may indirectly encourage or discourage human involvement, all without regard to what we know about the strengths and weaknesses of human and algorithmic decisionmakers and the particular quirks of hybrid human-machine systems. Second, we identify "the MABA-MABA trap," wherein regulators are tempted to address a panoply of concerns by "slapping a human in it" based on presumptions about what humans and algorithms are respectively better at doing, often without realizing that the new hybrid system needs its own distinct regulatory interventions. Instead, we suggest that regulators should focus on what they want the human to do—what role the human is meant to play—and design regulations to allow humans to play these roles successfully. Third, borrowing concepts from systems engineering and existing law regulating railroads, nuclear reactors, and medical devices, we highlight lessons for regulating humans in the loop as well as alternative means of regulating human-machine systems going forward.

[*] RC is Assistant Professor of Law, University of Richmond School of Law. MEK is Associate Professor of Law, Colorado Law School and Director, Privacy Initiative, the Silicon Flatirons Center. WNP is Professor of Law, University of Michigan and Core Partner, Center for Advanced Studies in Biomedical Innovation Law, University of Copenhagen Faculty of Law. For detailed feedback on earlier drafts, we thank Nick Bagley, Stefan Bechtold, Douglas Bernstein, Charleton Copeland, Dan Deacon, Rebecca Eisenberg, Michael Froomkin, Jim Gibson, Patrick Gudridge, Kristian Hammond, Corinna Lain, Matt Sag, Andres Sawicki, Alex Stremitzer, Charlotte Tschider, and Christopher Yoo. For helpful comments and conversations, we thank Luca Baltensberger, Rabia Benhalim, Ana Bracic, Fiona Illi, Izzy Longstaff, Aileen Neilson, Nadav Orien-Peer, Gabriel Rauterberg, Blake Reid, Nikkita Rivera, Sloan Speck, Elizabeth Stalfort, and Harry Surden. This paper benefited from generous feedback at the ETH Zurich/University of Zurich/University of St. Gallen Workshop on the Law & Economics of Innovation, the Northwestern/Penn/Stanford Junior Faculty Forum for Law and Stem, the Richmond Law School Faculty Workshop, the University of Miami Law School Legal Theory Workshop, the University of Michigan Law School Governance Workshop, the University of Michigan AI & Law Seminar, and the Techlaw Junior Scholars Workshop. We thank Nathan Fuller and Phoebe Roque for excellent research assistance. WNP's work was supported by the Novo Nordisk Foundation (Grant No. NNF17SA0027784).

Table of Contents

INTRODUCTION

Artificially intelligent algorithms are being integrated into decisionmaking processes at mind-boggling speed and scale.[1] Governments use AI for law enforcement, managing the spread of infectious disease, and distributing benefits.[2] Hospitals are creating AI-powered systems to identify brain hemorrhages,[3] catch life-threatening sepsis,[4] and suggest which patients need more assistance to stay out of the hospital.[5] Militaries are researching, developing, and fielding AI-enabled autonomous weapon systems[6] and increasingly incorporating AI

[1] We use "algorithms" as a catch-all term for everything from automated to artificially intelligent systems. "Algorithms" are sets of instructions that can be executed when triggered; "Artificial intelligence" is composed of groups of algorithms that can be modified in response to learned input. *See* European Commission, Proposal for a Regulation of the European Parliament and the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (Apr. 21, 2021), Title I Art. 3(1), p. 39 (defining "AI" expansively to include machine-learning approaches, logic- and knowledge-based approaches, and statistical approaches) [hereinafter Draft E.U. AI Act].

[2] *See, e.g.*, Cary Coglianese & Lavi M. Ben Dor, *AI in Adjudication and Administration*, 87 BROOK. L. REV. (forthcoming) (manuscript at 2) (on file with authors) ("This article seeks to capture the state of the art of the current uses of digitization, algorithmic tools, and machine learning in domestic governance in the United States."); Aziz Z. Huq, *A Right to a Human Decision,* 160 VA. L. REV. 611, 651 (2020); Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L. J. 797, 800 (2021) ("A little over a decade ago, the problems associated with automating public-benefits determinations came into view. In the public benefits arena, programmers embedded erroneous rules into the systems, more often by mistake or inattention than by malice or intent. Systems cut, denied, or terminated individuals' benefits without explanation in violation of due process guarantees."); Hamsa Bastani et al., *Efficient and Targeted COVID-19 Border Testing via Reinforcement Learning*, 599 NATURE 108, 108 (2021) (describing the Greek government's use of AI to use limited testing resources to most effectively identify asymptomatic travelers infected with COVID-19).

[3] *See* Mohammad R. Arbabshirani et al., *Advanced Machine Learning in Action: Identification of Intracranial Hemorrhage on Computed Tomography Scans of the Head with Clinical Workflow Integration,* npj DIGIT. MED., April. 4, 2018, at 1.

[4] *See, e.g.*, Mark Sendak et al., *Real-World Integration of a Sepsis Deep Learning Technology into Routine Clinical Care: Implementation Study*, 8 JMIR MED. INFORM. 1 (2020) (discussing Duke Health's Sepsis Watch program).

[5] *See, e.g.*, Rebecca Robins & Erin Brodwin, *An Invisible Hand: Patients Aren't Being Told About the AI Systems Advising Their Care*, STAT NEWS (July 15, 2020), https://www.statnews.com/2020/07/15/artificial-intelligence-patient-consent-hospitals/.

[6] *See, e.g.*, PAUL SCHARRE, AUTONOMOUS WEAPONS AND OPERATIONAL RISK (2016) (noting that, as of 2016, over thirty states already have "air, rocket, and missile defense systems with human-supervised autonomous modes"); Kai-Fu Lee, *The Third Revolution in Warfare*, ATLANTIC (Sept. 11, 2021), https://www.theatlantic. com/technology/archive /2021/09/i-weapons-are-third-revolution-warfare /620013/; Gerrit D. Vynck, *The U.S.*

decision assistants to collect information, crunch data, assess threats, and recommend strategic moves or specific targets.[7] In these and a host of other fields—agriculture, commerce, education, employment, energy, housing, law, philanthropy, transportation—there is growing interest in integrating algorithms into decisionmaking processes, either as decision aids or decisionmakers.

This proliferation has prompted questions of where and how humans should be involved in algorithmic decisionmaking processes—or, conversely, if certain weighty or irreversible decisions should be delegated to non-human entities at all.[8] Regulators frequently respond to these concerns by either explicitly requiring or implicitly encouraging retaining a "human in the loop"—which we define as an individual involved in a single, particular algorithmic decision.[9] In the United States, for example, over forty government policies now require human oversight or involvement in various algorithmic decisionmaking processes.[10] However, regulators often deploy humans sloppily, in ways that set up the human (and the greater system) to fail. As algorithms are being integrated into more and more decisionmaking processes, policymakers need better guidance on how to use law to foster productive hybrid decisionmaking.

Says Humans Will Always Be in Control of AI Weapons. But the Age of Autonomous War Is Already Here, WASH. POST. (July 7, 2021), https://www.washingtonpost.com/technology/2021/07/07/ai-weapons-us-military/.

[7] *See, e.g.*, Ashley S. Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 NAT'L SEC. LAW & POL'Y 1 (2019); Ashley Deeks, *Predicting Enemies*, 104 VA. L. REV. 1529 (2018).

[8] Indeed, many have tackled this question. *See, e.g.*, Ben Green, The Flaws of Policies Requiring Human Oversight of Government Algorithms (Sept. 23, 2021) (unpublished manuscript) (on file with authors); Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1, 6, 44–54 (2019); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 804–05 (2014). For the purposes of this Article, we take the loop as a given to evaluate how law affects its structure.

[9] This description reflects how most regulators tend to think about humans in hybrid systems; using it allows us to explore the issues with the "slap a human in it" regulatory strategy. However, it is far from the only possible one. As we discuss later, more expansive definitions better highlight the myriad ways humans affect algorithmic systems. *See infra* Part V.

[10] Green, *supra* note 8, at 9–14. Some of Green's examples involve "human in the loop" systems (where a human is involved in an algorithmic decisionmaking process); we would describe others as "human on the loop" systems (where a human oversees an algorithmic decisionmaking process), both of which are often contrasted with "human off the loop" systems (algorithmic decisionmaking processes without human involvement or oversight).

Informed by our disparate areas of expertise[11] and starkly different baseline assumptions regarding the utility and drawbacks of human and algorithmic decisionmaking,[12] we make three generalizable contributions, applicable to regulating human-in-the-loop systems across contexts. We collectively neither advocate for nor against placing a human in the loop; instead, we argue that if policymakers are going to insert humans to accomplish certain policy aims, they need to do a better job. As we detail, there are pitfalls to be avoided and strategies for crafting more effective regulations.

Many bemoan law's seeming inability to "keep up" with this new technology. Accordingly, a draft E.U. law[13] that would regulate AI systems across sectors has been widely hailed as "the first-ever legal framework on artificial intelligence."[14] Our first contribution is to highlight, contrary to this popular narrative, that there already is a "law of humans in the loop"—law that influences the inclusion and capabilities of humans who affect what would otherwise be purely algorithmic decisions.

Somewhat surprisingly, as of yet there has been no comprehensive evaluation of the law of the loop. There is a robust and growing body of legal scholarship discussing different elements of regulating algorithmic decisionmaking,[15] but these forward-looking works rarely acknowledge

[11] Collectively, we have expertise in the law of armed conflict, free expression, health law, intellectual property, international law, national security law, privacy law, property, and torts.

[12] For a "Pollyannaish, techno-utopian" take (per Rebecca Crootof), see W. Nicholson Price II, *Medical AI and Contextual Bias*, 33 HARV. J.L. & TECH. 66, 101–04 (2019) (arguing that medical AI will do tremendous good and physicians will make lots of mistakes and often won't make AI better); for a more "skeptical, progress-hating monster" perspective (per Nicholson Price), see Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 U. PA. L. REV. 1347 (concluding that autonomous weapon systems will inevitably make mistakes, with horrific consequences); Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1538–40 (2019) 1538–40 ("Calls for algorithmic decision-making often start from the premise that it replaces something worse: decision-making by humans . . . . But even complex algorithms are simplifications of reality, and these simplifications involve human choices along a number of axes.").

[13] Draft E.U. AI Act, *supra* note 1.

[14] *E.g.*, Eve Gaumond, *Artificial Intelligence Act: What Is the European Approach for AI?*, LAWFARE (June 4, 2021), https://www.lawfareblog.com/artificial-intelligence-act-what-european-approach-ai.

[15] The substantial work on the law of algorithmic decisionmaking comes in several interrelated strands. It includes works addressing implementation questions when regarding encoding law and policies. *See, e.g.*, Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014) (arguing for procedural regularity and oversight to ensure fairness and accuracy

how law already shapes human involvement in a decisionmaking process.[16] To the extent there is scholarship discussing relevant existing law, it has been largely siloed by field and focused on a particular topic or technology, such as international humanitarian law and autonomous

in artificially intelligent scoring systems); Pasquale, *A Rule of Persons, Not Machines*, *supra* note 8 (discussing ambiguities that arise in the context of translating health privacy law into code); Harry Surden, *The Variable Determinacy Thesis*, 12 COLUM. SCI. & TECH. L. REV. 1, 6–8 (2011) (proposing guiding principles for automating legal reasoning, on the theory that some legal concepts are relatively determinable); *see also* Lisa A. Shay, Woodrow Hartzog, John Nelson & Gregory Conti, *Do Robots Dream of Electric Laws? An Experiment in the Law as Algorithm*, *in* ROBOT LAW 274 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016) (demonstrating that attempts to encode even seemingly clear laws—like traffic speed limits—could be problematically indeterminate in practice). Some work considers current and potential second-order effects. *See, e.g.*, RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE (2019); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018); Calo & Citron, *supra* note 2 (arguing that, as agencies increasingly rely on automated decisionmaking, they lose the expertise and flexibility that justified their existence and authority); Rebecca Crootof, '*Cyborg Justice' and the Risk of Technological-Legal Lock-in*, 119 COLUM. L. REV. F. 233 (2019) (arguing that translating law and judicial decisionmaking processes into code might create an additional barrier to legal evolution and thereby foster legal stagnation and a loss of judicial legitimacy); Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242 (2019) (arguing that incorporating AI in the common law judicial process will encourage a shift in societal values and expectations around judging, from a focus on equity to a focus on quantifiable results—which in turn will affect who aspires to the bench); Rory Van Loo, *The Corporation as Courthouse*, 33 YALE J. ON REG. 547 (2016) (discussing private automation complaint-resolution mechanisms). Other scholarship evaluates related social and governance considerations. Kaminski, *Binary Governance*, *supra* note 12 (discussing how states and industry might work together to govern the use of algorithms); Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633 (2019) (advocating for public governance of data to avoid the future likelihood of private AI governance).

[16] A number of works consider aspects of the relationship between humans and algorithms in decisionmaking systems, and we draw on their insights extensively. *See, e.g.,* Kiel Brennan-Marquez, Karen Levy & Daniel Susser, *Strange Loops: Apparent Versus Actual Human Involvement in Automated Decision-Making*, 24 BERK. TECH. L.J. 745 (2019) (discussing the relevance of appearing to have a human in the loop, even if that individual does not have any actual authority to affect the decisionmaking process); Aziz Z. Huq, *Constitutional Rights in the Machine-Learning Rights*, 105 CORNELL L. REV. 1908–10 (reporting that human oversight may not adequately address due process concerns regarding the quality of AI decisions because a human in the loop does not always reduce the number of false positives and false negatives); Meg Leta Jones, *The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles*, 18 VAND. J. ENT. & TECH. L. 77, 90–91 (2015) (describing how humans, who operate under the theory that they are "unreliable and inefficient," automate the easiest tasks, thereby increasing the amount of time they must use to tackle the harder ones—while overseeing failable automated systems); Andrea Roth, *Trial by Machine*, 104 GEO. L. J. 1245 (2016) (discussing the role of AI in criminal adjudication).

weapon systems,[17] health or privacy law and black-box medicine,[18] or administrative law and automated benefit disbursement.[19] But comparing systems across contexts highlights law's current influence.[20] Sometimes, human decisionmakers are explicitly or implicitly (and often inadvertently) required.[21] Other times, legal systems may indirectly incentivize keeping a human in the decisionmaking process.[22] Conversely, law may discourage or prohibit retaining human influence.[23]

Identifying the existing law of the loop also allows us to better see its problems.[24] Namely, it often operates haphazardly and inadvertently, sometimes placing humans in the loop without regard to their skills and frailties. Take autonomous vehicles: The most dangerous time for a human to take control of an autonomous vehicle is during a split-second emergency, when the handoff itself may cause deadly delay or errors. Nevertheless, the risk of tort liability may incentivize autonomous

[17] *See, e.g.*, Rebecca Crootof, *The Killer Robots Are Here: Legal Policy and Implications*, 36 CARDOZO L. REV. 1861 & n.79 (2015).

[18] *See, e.g.*, W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421 (2017).

[19] *See, e.g.*, Calo & Citron, *supra* note 2, at 800.

[20] Given this fast-developing field and the early stages of many of these policy discussions, we take a relatively capacious view of what constitutes "law." In addition to binding legislation, treaties, and other formal regulation, we include rules that have not yet been adopted but which nonetheless influence relevant actors, see, e.g., Draft EU Artificial Intelligence Act; agency documents that are treated as authoritative or quasi-binding by relevant stakeholders, see, e.g., FDA guidance or Guidelines from the European Data Protection Board (EDPB); and even soft-law recommendations by respected bodies that are frequently adopted and thus might be considered proto-law, see, e.g., International Committee of the Red Cross, ICRC Position on Autonomous Weapon Systems and Background Paper (proposing regulations for autonomous weapon systems). We endeavor throughout to be clear about the nature of the "law" we consider.

[21] For example, the European Parliament's proposed Regulation on Artificial Intelligence (Regulation on AI) sometimes explicitly requires human oversight. *See infra* Part II.A.1. Meanwhile, the 1968 Convention on Road Traffic implicitly assumes a (presumably human) driver. *See infra* Part II.A.2.

[22] For example, makers of software that supports physician decisionmaking can avoid onerous and costly regulatory processes if they design their product to keep a human in the loop. *See infra* Part II.B.1.

[23] For example, in the absence of laws requiring human involvement, high-frequency trading algorithms and defensive cybersecurity systems discourage including humans in the loop. *See infra* Part II.C.1.

[24] Which, in turn, highlights the dangers of presuming that law trails technological change: a reactive perspective obscures how law influences the development and implementation of technology in different contexts. Rebecca Crootof & BJ Ard, *Structuring Techlaw*, 34 HARV. J. L. & TECH. 347 (2021); Meg Leta Jones, *Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw*, 2018 J. L. TECH & POL'Y 101, 108 ("By accepting the pacing problem and chasing new technologies with legal solutions, law and technology scholars, as well as policymakers, unnecessarily accept a degree of irrelevance."); Margot E. Kaminski, *Authorship, Disrupted: AI Authors in Copyright and First Amendment Law*, 51 U.C. DAVIS L. REV. 589, 615 (2017).

vehicle developers to force a handoff to a human driver in precisely such situations, because doing so increases the likelihood that the human driver, rather than the designer, bears the brunt of liability.[25] Tort rules, though not intended to affect the design or operation of the hybrid system, profoundly shape its dynamics.[26] Given this, anyone thinking about how best to regulate these systems cannot assume they have a blank slate; instead, they must consider how new rules might build upon or be undermined by extant ones.

Second, when attempting to craft new regulations, policymakers often fall into "the MABA-MABA trap." Based on what we know about what "Men Are Better At" and what "Machines Are Better At,"[27] rulemakers often insert humans in decisionmaking systems based on assumptions about their respective capabilities. Returning to autonomous vehicles: AI-driven cars may be excellent at repetitive tasks, such as following curves on a highway, but terrible at improvising, as might be required when encountering unidentifiable debris in the road. The MABA-MABA solution? Transfer control to the human when there's unidentifiable debris. There's a seductive simplicity to the "slap a human in it" approach,[28] not least because it's often grounded on truths: at least at present, machines are better at repetitive tasks, while humans are better at contextual analysis.

But even if humans do their jobs to the best of their abilities, MABA-MABA can be a trap. Human-in-the-loop systems are more than the sum of their parts: rather than marrying the best of both, hybrid systems can exacerbate the worst of each, while adding new sources of error as information is lost in translation. When designing autonomous vehicles, for example, systems that hand off control to the driver need to deal with the reality that drivers in autonomous vehicles lose focus and may drift off; systems might need to include alerts and sufficient time for the operator to take control. Nor is placing an ineffective human in the loop

[25] Ryan Calo, Robots in American Law 36 (unpublished manuscript) (on file with authors) (observing that judges have a tendency to attribute liability to the person "in the loop" over a robotic system).

[26] *See* K.C. Webb, *Products Liability and Autonomous Vehicles: Who's Driving Whom?,* 23 RICH. J.L. & TECH. 9, 38 (2016) ("Certain design features of AVs are responsive to legal requirements. For example, California law requires that all AVs be equipped with a steering wheel and a driver at the ready. However, it is not just statutory and regulatory reform driving the incorporation of certain design elements. Products liability concerns exert a similar influence.").

[27] Jones, *Ironies of Automation*, *supra* note 16, at 1.

[28] *See, e.g.*, Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood,* 47 SOC. STUD. SCI. 216, 224 (2017) (describing the E.U. use "of the human in the loop as a regulatory tool.").

harmless. It can serve as "ethics washing," distract from other, more effective forms of regulation, and all too often, set up the system to fail and leave the human to shoulder the blame.

While we are not the first to note the dangers of a MABA-MABA approach,[29] many policymakers seem to have not yet gotten the message.[30] This may be due in part to the absence of an alternative framework. The human in the loop is a concrete and identifiable entity, and thus an easy regulatory target. While policymakers should be wary of weighing in on *where* humans should be located in hybrid systems based on a limited understanding of human or machine capabilities, policymakers are well-situated to articulate *why* a human should be incorporated in a decisionmaking system. In furtherance of this shift, we detail the panoply of roles that humans in the loop might be expected to play, ranging from the benign-if-challenging (correcting errors or respecting dignity) to the understandable-but-problematic (serving as a hook upon which to hang liability for accidents) to the mildly orthogonal (making sure humans get to keep jobs in an automated economy).

Third, policymakers tend to have a blinkered understanding of scope of "the loop." They focus on regulating individuals involved in particular decisions, rather than on whether the system as a whole enables humans to play their expected roles. But effectively regulating a human-in-the-loop system requires regulating the system, rather than the human. While acknowledging the terrifying complexity of effectively regulating these systems, we synthesize findings and lessons learned from human factors engineering to suggest alternative legal interventions.

Here's what's coming. In Part I, we acknowledge that there is no one correct definition of a "human in the loop" and explain that we adopt a version that reflects lawmakers' apparent narrow focus on humans involved in a particular decision. Part II is both descriptive and normative: we describe how legal systems already shape the structure of human/machine decisionmaking processes, often inadvertently and problematically. We then shift to issues regulators wishing to craft new rules must face, beginning with the MABA-MABA trap in Part III. After discussing what humans are better at, what machines are better at, and the special challenges of hybrid systems, we synthesize findings from the regulation of railroads, nuclear reactors, and medical devices to demonstrate what grappling with regulating human-in-the-loop systems

[29] Roth, *supra* note 16, at 1297 (criticizing the MABA-MABA approach and arguing that we should look to systems engineering to learn how to design human-machine systems so the system as a whole works better).

[30] *See, e.g.,* Jones, *The Right to a Human in the Loop, supra* note 28.

really looks like. In Part IV, we suggest that, rather than focusing solely on regulating humans in the loop directly, regulators consider what role the human is expected to play—which might include corrective, justificatory, dignitary, accountability, "warm body," and interface roles.

We had hoped to be able to glean simple cross-cutting insights that would allow us to provide a neat, tidy, and comprehensive solution to these governance problems. Instead, we make recommendations that are inherently messy, because regulating human in the loop systems is inherently messy. Still, Part V offers actionable suggestions for how policymakers might better craft regulations going forward: (1) be specific about what humans in the loop are there to do (recognizing that some roles may conflict); (2) take context into account; and (3) learn from human factors engineering about how to situate humans to succeed. Using the draft AI Act as a case study, we showcase how to better regulate human-in-the-loop systems—which will sometimes require considering alternative regulatory interventions.

We close on a perhaps unsatisfying but honest note: Governing human-algorithm hybrid systems is *hard*. There are myriad, cross-cutting, and influential background laws. Inserting a human into the loop isn't the convenient regulatory intervention it first appears to be. There is no straightforward, one-size-fits-all solution. Instead, as demonstrated by railroads, nuclear reactors, and medical devices, regulating human-in-the-loop systems is messy, complicated, and contextual. It's hard. But it's doable, and regulators can do it better.

I. DEFINING A "HUMAN IN THE LOOP"

For the purposes of this paper, we define a "human in the loop" as an individual who is involved in a single, particular decision made in conjunction with an algorithm.[31] Human-in-the-loop systems include ones where (1) an individual decides to use an algorithmic system to

[31] It is important not to conflate the question of whether a system has the capacity to be autonomous with the question of whether there is a human in the loop. For example, the Society of Automotive Engineers' five levels of automation range from a human performing all tasks (level zero) to an automated vehicle performing all tasks in all conditions, even where a human may have an option to control the vehicle (level five). *See The Society of Automated Engineering's Levels of Automation*, *Automated Vehicles for Safety,* NAT'L HIGHWAY TRAFFIC SAFETY ADMIN, https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety (last visited Dec. 13, 2021). Level five automation focuses on the capabilities of the system—not the potential involvement of a human in practice. By our definition, some SAE level 5 automated systems will not have humans in the loop. Others may, such as in circumstances where humans face legal incentives to exercise their option to control the car.

reach a decision, such as the doctor who chooses whether to use an AI diagnostic tool when treating a patient; (2) an individual and algorithm which pass off tasks or performs tasks in concert, such as the pilot who performs some tasks manually while relying on an autopilot for others; (3) an individual alters an algorithm mid-determination, such as the lawyer who reconfigures the parameters of an e-discovery tool mid-process; and (4) an individual determines whether to implement an algorithmically-informed conclusion, such as the commander who decides against engaging a recommended target. Our definition thus includes more actors than those which focus only on humans engaged in oversight or review.[32]

It might be helpful to contrast such systems with one without humans in the loop.[33] Let's say a company decides to use an AI or other algorithm to screen all job applicants.[34] That process could be entirely automated: a candidate submits her resume and uploads a video answering written questions, and the algorithm screens the resume for particular terms and the interview for particular personality characteristics such as "the willingness to work hard and persevere" (really) and rejects any candidates who don't meet certain criteria.[35] If the process is supervised by a human who does not intervene in particular decisions, there is a human "on" (rather than "in") the loop. If no human even reviews the process or decisions, humans are "off" the loop.

Other definitions are far broader than ours, as they zoom out to highlight the myriad indirect ways humans affect the design, training data and inputs, and ex post evaluation of decisionmaking processes, in ways which are often obscured when we focus overmuch on a discrete application of a system or its particular results.[36] In her writing about

[32] *See, e.g.*, Brennan-Marquez, Levy & Susser, *supra* note 16, at 749; Green, *supra* 8, at 3, n. 1.

[33] The "human in the loop"—the human involved in an algorithmic decisionmaking process—is often contrasted with the "human on the loop"—the human overseeing an algorithmic decisionmaking process—and a "human off the loop"—algorithmic decisionmaking processes without human involvement or oversight.

[34] *See, e.g.*, Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, REUTERS (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G; Nathan Mondragon, *Creating AI-Driven Pre-Hire Assessments*, HIREVUE (June 6, 2021), https://www.hirevue.com/blog/hiring/creating-ai-driven-pre-employment-assessments (which now heavily emphasizes its use as a tool rather than in lieu of people).

[35] Mondragon, *supra* note 34.

[36] *See, e.g.*, KATE CRAWFORD, ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE 9 (2021) (discussing how AI is conceptually often

automation and "man-machine systems,"[37] for example, Meg Jones observes that there is no such thing as a purely algorithmic decision; instead, humans have a "permanent role in the loop."[38] Similarly, Andrew Selbst, danah boyd, Sorelle Friedler, Suresh Venkatasubramanian and Janet Vertesi observe that all "technical systems are subsystems" within social and human contexts.[39] Like all technical systems, even human-off-the-loop systems are part of sociotechnical systems, which always include humans.[40]

But our definition is purposive: we focus on the individual involved in a particular decision because lawmakers tend to focus on such individuals.[41] The human in the loop connected to a particular outcome is a concrete and identifiable entity, and thus a frequent regulatory target. In contrast to ex ante systems design and ex post oversight systems, which require expertise usually outside of lawmakers' ambit, regulating individuals is familiar ground. Accordingly, when defining or crafting regulations for algorithmic systems, regulators often ignore other, less obvious individuals who might also be considered "humans in the loop" under a broader definition of the term.[42]

disembodied from the systems of extraction and power relations between humans that undergird its use); Hannah Bloch-Wehba, Democratizing Technology (unpublished manuscript) (on file with authors) (discussing how oversight in the design stage includes transparency requirements and external stakeholder involvement); David Lehr & Paul Ohm, P*laying with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 655 (2017).

[37] Jones, *Ironies of Automation*, *supra* note 16, at 84.

[38] *Id.* at 104.

[39] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian & Janet Vertesi, Fairness and Abstraction in Sociotechnical Systems 59 (Jan. 29, 2021) (unpublished manuscript) (on file with authors), http://sorelle.friedler.net/papers/sts_fat2019.pdf.

[40] *Id.* at 60.

[41] *See, e.g.*, Regulation 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), art. 22(3), 2016 O.J. (L 119) (EU) [hereinafter GDPR] (characterizing a "human in the loop" as being a person involved in the decision when the algorithm is being used (as opposed to when it is designed or trained)); Artificial Intelligence Video Interview Act, 820 Ill. Comp. Stat. 42/1 et seq. (regulating employers that "use[] an artificial intelligence analysis of the applicant-submitted videos" to determine whether an applicant qualifies for an in-person interview); Colo. Rev. Stat. § 10-3-1104.9 (prohibiting insurers from using discriminatory algorithims as a part of their insurance rating process). In recent years, some regulatory proposals have shifted to more systemic approaches. *See, e.g.,* Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019); Draft E.U. AI Act, *supra* note 1 (discussed further below).

[42] One aim of this Article is to highlight the limitations of this narrow approach. *See infra* Part V.

Note that nothing in our definition implies that the human in the loop must be effective. Other definitions, including those advanced in some of our prior individual writings,[43] suggest that human oversight or intervention must be "meaningful" to count—and that a human who merely rubberstamps an algorithmic decision does not constitute a human in the loop.[44] But our definition here, which mirrors the law of the loop, encompasses humans who are unable to effectively achieve a regulator's aims. The law rarely ensures that these individuals are able to successfully play their roles.

Finally, before we turn to exploring the law of the loop, it is worth taking a moment to acknowledge the importance of framing, insofar as whether a human appears to be in the loop will depend on how we define the scope of a system's tasks. Consider the elevator: If the task is described as physically moving the elevator up and down, there is no human in the loop—we no longer have elevator operators.[45] But if the scope of the task includes choosing what floor to go to and deciding when to call and command the elevator, the human user now still does most information processing and decision-making, relegating the elevator to the limited role of implementation.[46] Or take a more pressing modern example of the import of scope: when debating potential regulations for autonomous weapon systems, the International Committee of the Red Cross noted that any regulatory definition should focus on "autonomy in the critical functions of selecting and engaging targets."[47] The Committee explained that "[a]utonomy in other functions (such as movement or navigation)" would not be relevant in a discussion as to what distinguishes autonomous weapon systems from those controlled more

[43] Crootof, *Killer Robots Are Here*, *supra* note 17, at 1861 & n.79 (2015) (arguing that autonomous weapon systems with nominal human involvement should be considered "effectively autonomous weapon systems" rather than semi-autonomous weapon systems); *see also* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189 (2019) (arguing that, contrary to others' assessments, the GDPR's Article 22 requires "meaningful" human involvement for a system to be considered a human-in-the-loop system and thus, when regulating "solely" automated decisionmaking, the GDPR actually regulates systems that would be considered "human-in-the-loop" systems under this Article's definition).

[44] Brennan-Marquez, Levy, and Susser also require the human to have "some degree of meaningful influence" to be considered in the loop, *supra* note 16, at 749, while Green appears to require human discretion to use or reject an AI decision, but not necessarily meaningful capacity to do so. *See, e.g.*, Green's differentiation of "oversight" policies into those "emphasizing human discretion" versus those "requiring 'meaningful' human input." Green, *supra* note 8, at 10, 12.

[45] Jones, *Ironies of Automation*, *supra* note 16, at 108.

[46] *Id.*

[47] Toward Limits on Autonomy in Weapon Systems, INT'L COMM. RED CROSS (April 9, 2018), https://www.icrc.org/en/document/towards-limits-autonomous-weapons.

directly by humans.[48] In doing so, it dramatically circumscribed which humans would constitute humans in the (lethal) loop.

II. THE LAW OF THE HUMAN IN THE LOOP

As surveyed here, there is already a "law of the human in the loop": a complex web of regulation that has a surprisingly profound influence on the presence and contours of human roles in algorithmic systems.[49] While the majority of this paper focuses on regulatory interventions that intentionally affect the human in the loop, we begin with an overview of the many explicit and indirect ways in which law already impacts humans in the loop. Obviously, law can drive whether humans are in the loop at all by requiring or forbidding their presence. Less obviously, law creates incentives that encourage or discourage their presence.

Collecting these different examples together highlights common problems. Namely, law's influence is frequently incidental and path-dependent,[50] rather than grounded on thoughtful evaluations of the intended role of a human in the loop and how law might facilitate success in that role.[51] Some laws place a human in the loop badly. Some were drafted before it was technologically relevant to ask whether a human should be involved in a decisionmaking process; others are written without awareness of how they already shape answers to that question. All too often, the law of humans in the loop sets humans up to fail.

A. Requiring

As algorithmic systems proliferate, a growing number of proposed and enacted laws explicitly require a human in the loop. Many existing regimes also effectively mandate humans in the loop, even if their creators did not have that intention or even imagine the possibility that the rules might apply to algorithmic systems.

1. New Explicit Mandates

Various rules explicitly mandate human involvement in algorithmic

[48] *Id.*

[49] Of course, law is far from the only relevant regulatory modality. For example, where a human presence may minimize reputational harms or social censure, market forces and social norms may dovetail to encourage the inclusion of humans in the loop.

[50] *See infra* Part IV.

[51] Policymakers might have entirely unrelated goals, such as avoiding injury or compensating those injured. But many laws have the side effect of creating incentives to keep humans in the loop.

decisionmaking. Ben Green has compiled over forty policies that prescribe human oversight of algorithms employed in governmental decisionmaking. Some policies prohibit decisions made "solely" by algorithms, some emphasize that human oversight and discretion is necessary to address algorithmic error, and some require "meaningful" human oversight.[52]

The newly proposed E.U. AI law ("the draft AI Act") would explicitly require "high risk" AI to be built for oversight by a human in the loop.[53] It obliges providers of "high risk" AI systems to design and develop AI systems so that "they can be effectively overseen by natural persons during the period in which the AI system is in use."[54] "High risk" AI systems include certain biometric systems, like facial or gait recognition systems; systems that operate critical infrastructure; systems that assess students; and some kinds of employment-related AI, among others.[55] AI are also required to build systems with "appropriate human-machine interface tools."[56] They must enable humans in the loop to understand the capacities of the system "and be able to duly monitor its operation"; detect and address anomalies; correctly interpret the AI system's output; reject the output; and stop the system's operation.[57] Providers must also somehow design the system so that humans in the loop remain aware of automation bias.[58] Aspirationally, the draft Act's preamble suggests the human in the loop should also "have the necessary competence, training and authority to carry out that role;" however, nothing in the Act itself requires this.[59]

The draft AI Act would also create a humans-in-the-loop requirement for certain biometric systems, insofar as it mandates sign-off by two

[52] Green, *supra* note 8, at 9–14.

[53] Draft E.U. AI Act, *supra* note 1, art. 14 ("High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use."). The draft AI Act, broadly speaking, creates obligations that apply to two sets of actors: AI "providers," who build the systems, and AI users, who use them. The draft AI Act divides and regulates AI systems according to levels of risk, ranging from "unacceptable" systems that are banned to low-risk systems whose builders and users voluntarily self-regulate. *See* Michael Veale & Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, 22 COMPUT. L. REV. INT'L. 97, 98. Title II of the draft E.U. AI Act regulates "unacceptable risks" and "high risks"; Title IV regulates "limited risks"; and Title IX regulates "minimal risks." Draft E.U. AI Act, *supra* note 1.

[54] *Id.* art. 14(1).

[55] *Id.* annex III. The European Commission can later add to this list.

[56] Draft E.U. AI Act, *supra* note 1, art. 14(1).

[57] *Id.* art. 14(4).

[58] *Id.* art. 14(4)(b).

[59] *Id.* at recital 48.

humans before biometric identification could be used in certain contexts.[60] With respect to biometric systems in particular, the draft Act requires providers to build AI and instruct its users such that "no action or decision is taken by the user on the basis of the identification resulting from the system unless this has been verified and confirmed by at least two natural persons."[61] This requirement positions the human(s) at the end of the loop, as gatekeepers to prevent inaccurate action.

In addition to existing law, some call for new law requiring a human in the loop. For instance, a civil society-led movement is campaigning for regulations which would require human involvement in decisions to employ lethal force in war.[62]

2. Existing De Facto and Implicit Mandates

A variety of existing laws result in de facto requirements for humans in algorithmic systems, even if they were not written with algorithmic systems in mind. For instance, two international conventions on road traffic require that every road vehicle must have a driver able to control the system while it is in motion. If applied to autonomous vehicles, this provision could be interpreted to require a human driver in the loop.[63] Similarly, medical prescriptions may only be written by a specified set of professionals. Accordingly, any algorithmic system which recommends the use of prescription drugs would need to have a human in the loop to actually do the prescribing.[64] A number of laws might similarly require

[60] *Id.* art. 14(5); *see also* Veale & Zuiderveen Borgesius, *supra* note 53, at 103 ("A 'four-eyes' principle requires biometric identification systems to be designed so that two natural persons can sign off on any identification and have their identities logged, and for instructions to specify that they must.").

[61] *Id.* art. 14(5).

[62] *E.g.,* ICRC, *supra* note 20, at 9 (proposing a ban on anti-personnel autonomous weapon systems); *see also* Rebecca Crootof, *Changing the Conversation: The ICRC's New Stance on Autonomous Weapon Systems*, LAWFIRE (May 24, 2021), https://sites.duke.edu/lawfire/2021/05/24/changing-the-conversation-the-icrcs-new-stance-on-autonomous-weapon-systems/ (noting that the ICRC sidesteps calling for a ban on all autonomous weapon systems and raising the concern that a ban on anti-personnel autonomous weapon systems "risks inadvertently limiting the use of future technology that is capable of or even better than humans in context-specific distinction").

[63] Convention on Road Traffic art. 8, Sept. 19, 1949, 125 U.N.T.S. 22; Convention on Road Traffic art. 8, Nov. 8, 1968, 1042 U.N.T.S.15705; *see also* Bryant Walker Smith, *New Technologies and Old Treaties*, 114 AJIL UNBOUND 152 (2020) (discussing how these conventions are challenged by the advent of autonomous vehicles).

[64] *See, e.g.*, *Who Can Prescribe and Administer Prescriptions in Washington State*, WASH. DEP'T HEALTH, https://www.doh.wa.gov/LicensesPermitsandCertificates/ProfessionsNewReneworUpdate/PharmacyCommission/WhoCanPrescribeandAdministerPrescriptions#Prescribe (last visited September 16, 2021).

the involvement of "a person," "an individual," or "a citizen"—and, at least at present, such phrases are probably best understood to mean "a human being."

Some also intentionally read an implicit human-in-the-loop requirement into existing rules. In contrast to those who argue for creating new rules for autonomous weapon systems, for example, some ban advocates claim that this technology is already prohibited under (particular interpretations of) existing law.[65]

B. Encouraging

Law often operates indirectly, by incentivizing rather than requiring certain actions. We identify three mechanisms by which law may encourage the presence of humans in the loop. First, regulatory arbitrage fosters human involvement when it's possible to avoid costly regulations by keeping a human in the loop. Second, liability rules may encourage human inclusion to "absorb" the legal consequences should something go wrong. Third, the presence of a human in the loop may shield certain procedural decisions from challenge or appeal. All of these regulatory categories may exist either as tech-specific rules that apply specifically to algorithmic systems or as more tech-neutral obligations that apply to algorithmic systems among other things.

1. Regulatory Arbitrage

Some laws allow a system's developers or users to avoid more onerous regulation, such as regulatory review or the need to satisfy additional safety standards, by retaining a human in the loop. This often manifests

[65] *E.g.*, Peter Asaro, *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, 94 INT'L REV. RED CROSS 687, 687 (2012) (arguing that "an implicit requirement for human judgement can be found in international humanitarian law governing armed conflict"); Bonnie Docherty, *Shaking the Foundations: The Human Rights Implications of Killer Robots*, HUM. RTS. WATCH (May 12, 2014), https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots ("Fully autonomous weapons threaten to contravene foundational elements of human rights law. They could violate the right to life, a prerequisite for all other rights. Deficiencies in judgment, compassion, and capacity to identify with human beings could lead to arbitrary killing of civilians during law enforcement or armed conflict operations. Fully autonomous weapons could also cause harm for which individuals could not be held accountable, thus undermining the right to a remedy."); Jeffrey Kahn, "*Protection and Empire": The Martens Clause, State Sovereignty, and Individual Rights*, 56 VA. J. INT'L L. 1 (2016) (describing the Martens Clause to the 1899 Hague Convention Respecting the Laws and Customs of War on Land which held that principles of international law applied in all armed conflicts, regardless of their inclusion in any treaty).

in situations where a human putatively serves some role that regulators would otherwise need to perform.

For instance, in the 21st Century Cures Act ("Cures Act"), Congress clarified what sorts of software are subject to FDA's jurisdiction.[66] Software in certain categories is defined as a "medical device," such that marketing it requires going through FDA's premarket approval or clearance pathways, at substantial cost of time and money. All else being equal, it is cheaper for a developer to avoid FDA's processes,[67] and the Cures Act lets developers do exactly that by keeping a human in the loop. If software "is intended to provide decision support for the diagnosis, treatment, prevention, cure, or mitigation of diseases or other conditions" [68] by a human "health care professional" who is able to review its output before use, then the software (called "clinical decision support software" or CDS) is likely not a medical device subject to FDA approval.[69] This

[66] Kind of. *See* Barbara Evans & Frank A. Pasquale, *Product Liability Suits for FDA-Regulated AI/ML Software*, *in* THE FUTURE OF MEDICAL DEVICE REGULATION: INNOVATION AND PROTECTION *3 (I. Glenn Cohen, Timo Minssen, W. Nicholson Price II, Christopher Robertson & Carmel Shachar eds., forthcoming 2022)("[The Act] includes some (but not all) [Clinical Decision Support] software in the definition of a device that the FDA can regulate and provides a jurisdictional rule distinguishing which software is—and which is not—a medical device. In two subsequent draft guidance documents, the FDA has attempted to clarify this distinction, but key uncertainties remain unresolved for CDS software that incorporates artificial intelligence/machine learning (AI/ML) techniques.") (noting the confusion surrounding the Cures Act's definitions of software as medical devices).

[67] Rachel Sachs, *Innovation Law and Policy: Preserving the Future of Personalized Medicine,*, 49 U.C. DAVIS L. REV. 1881, 1890, 1895 (2016).

[68] *Clinical Decision Support Software*, FDA (May 6, 2020), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software.

[69] Within limits, clinical decision support software (CDS) is *not* a medical device if it is intended for the purpose of

> (i) displaying, analyzing, or printing medical information about a patient . . . (ii) supporting or providing recommendations to a health care professional about prevention, diagnosis, or treatment of a disease or condition; and (iii) enabling such health care professional *to independently review the basis for such recommendation*s that such software presents so that it is not the intent that such health care professional rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient.

21st Century Cures Act, Pub. L. 114–255 §306, 130 Stat. 1033, 1130–33 (2016), codified at 21 U.S.C. 360j(o) (emphasis added). *See* W. Nicholson Price II, Rachel Sachs & Rebecca S. Eisenberg, *New Innovation Models in Medical AI*, 96 WASH. U. L. REV. at *23 (forthcoming 2022) ("This exclusion covers some software functions that analyze data and that provide recommendations to a health care professional about prevention, diagnosis, or treatment of a disease or condition . . . ."). Note that the law requires only distinction has not escaped developers! For example, when Duke University developed its Sepsis Watch system for automated alerts to avoid sepsis, "[t]he team worked closely with regulatory officials to ensure that Sepsis Watch qualified as CDS and was not a diagnostic medical device."[70]

Similarly, the E.U. data privacy law, the General Data Protection Regulation (GDPR), imposes additional obligations on the use of "solely automated" decision-making with significant effects.[71] However, these potentially costly requirements[72] do not apply to automated decisions made with significant human involvement,[73] and nor possibly decisions made with even nominal human involvement.[74] Thus companies are

that the designer intend for there to be a human in the loop, not that a human is actually required for the decisionmaking process to function.

[70] Sendak et al., *Real-World Integration, supra* note 4, at 6.

[71] GDPR, *supra* note 41, art. 22.

[72] *Id.*; Article 29 Data Protection Working Party, Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679, 17/EN. WP 251rev.01 (Feb. 6, 2018), at 20, 32 (suggesting impact assessments are required for some systems and audits are best practices) [hereinafter GDPR Guidelines]. *See also* Kaminski, *Binary Governance, supra* note 12, at 1595. The GDPR's Article 2 also contains a right to "human intervention" in an algorithmic decision with significant effects. *See also* Huq, *A Right to a Human Decision, supra* note 2, at 622–24 ("Article 22(1) of the GDPR vests natural persons with a 'right not to be subject to a decision based solely on automated processing . . . . According to the European Commission Data Protection Working Party created by the EU, Article 22(1) applies only if 'there is no human involvement in the decision process.'") (internal citation omitted); Kaminski, *The Right to Explanation, Explained, supra* note 43, at 208 ("Article 22 requires safeguards—even when an exception applies—that, at a minimum, include a right to human intervention, a right to object, and a right to express one's view."). Such intervention, however, is not necessarily a requirement that humans be in the loop in all AI decisionmaking. Rather, it is better understood in conjunction with the GDPR's "right to contest" as an ex post right invocable by an affected individual. Kaminski & Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1989 (2021) ("Contestation rights do not always provide justice. Contestation may occur ex post, when some harms cannot be undone or ameliorated."); Emre Bayamlıoğlu, *Contesting Automated Decisions: A View of Transparency Implications,* 4 EUR. DATA PROTECT. L. REV. 433 (2018) (discussing requirements for effective automatic-decision contestation and incorporating the possibility of ex-post review). What that means in practice has yet to be determined. Huq, *A Right to a Human Decision, supra* note 2, at 623 ("The precise range of automated machine-learning tools captured by the prohibition thus remains up for grabs.").

[73] Decisions where the human is a rubber stamp, however, are likely covered. GDPR Guidelines, *supra* note 72, at 21 ("[I]f someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing.").

[74] Some scholars argue that by using the word "solely" the GDPR means only decisionmaking that doesn't involve a human at all, such that a company could escape regulation by adding even nominal human involvement. Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-*

incentivized to avoid additional regulatory obligations by putting a human in the loop.

2. Liability Rules

Liability rules sometimes encourage the presence of humans in the loop, especially when they allow system designers or operators to avoid liability by including a human decisionmaker.

As detailed in scholarship across fields, algorithmic systems can cause myriad harm when things go wrong (or even when things go right).[75] Automated weapons systems can kill civilians, autonomous vehicles can injure or kill pedestrians, algorithmic diagnostic systems can miss life-threatening illnesses, medical imaging systems can mischaracterize nascent tumors, and misfiring content-moderation or -promotion systems can stifle speech or facilitate the incitement of genocide.[76] Tort, criminal, or administrative liability may follow.

When liability rules tend to allocate liability to the human involved in the moment, system designers will have incentives to ensure that a human is in the loop—even if that human has no effective means of affecting the actual decision.[77] Bad things happen; when they do, it's useful to have a fall guy. Madeleine Elish has referred to these humans as the "moral crumple zone";[78] they could as easily be referred to as the legal crumple zone. Take a diagnostic algorithm. Today, medical practitioners have final say on diagnoses; if an algorithm suggests an incorrect diagnosis and a physician goes with it, the physician is more likely to be held liable than the algorithm's developer (though the

Making Does Not Exist in the General Data Protection Regulation 17 INT'L DATA PRIV. L. 76, 88 (2017). Others (including one of us) point out that guidance envisions only "meaningful" human involvement counting as a "human in the loop." Kaminski, *The Right to Explanation, Explained*, *supra* note 43; *see also* GDPR Guidelines, *supra* note 72, at 21 ("To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the relevant data."). Thus when the GDPR regulates "solely" automated decisionmaking, it actually regulates a significant amount of decisionmaking that we would consider (possibly ineffective) humans in the loop.

[75] *See, e.g.,* Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L REV. 1315, 1318 (2020) ("Medical Al will recommend improper treatment, robo-advisers will wipe out someone's bank account, and autonomous robots will kill or maim.").

[76] *See, e.g.*, Rebecca Hamilton, *Platform-Enabled Crimes*, 63 B.C. L. REV. (forthcoming 2022).

[77] Calo, *supra* note 25.

[78] Madeleine Claire Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAGING SCI., TECH., & SOC'Y 40 (2019).

landscape is still uncertain). The possibility of liability encourages developers to avoid creating algorithms that provide their own autonomous diagnosis.[79]

The problem with this incentive structure is that it doesn't come with a requirement that the human be effective. These humans in the loop may have little to no meaningful ability to affect the outcome, so long as they have enough nominal control to justify holding them legally liable and morally blameworthy.

3. Shielding Decisions from Challenge

At least in the United States, purely algorithmic decisions may be susceptible to challenge on procedural grounds—that is, that they failed to follow a legally adequate decisional process.

To the extent challenges of U.S. government use of complex algorithms have been successful, they have largely been procedural in nature.[80] There is a growing body of caselaw where algorithmic decisions were invalidated on procedural due process grounds.[81] For example, states—prompted by federal government monetary incentives—have adopted various Value Added Model systems for evaluating public-school teachers.[82] These systems' recommendations affected merit pay, the award or revocation of tenure, and even teacher terminations.[83] In a Texas lawsuit, teachers successfully claimed that the lack of transparency regarding how the algorithm reached its conclusions constituted a due process violation.[84]

[79] In fact, one of the few examples of a system that does provide an autonomous diagnosis without human intervention, IDx-DR, carries medical malpractice insurance for precisely that reason.

[80] Huq, *Constitutional Rights in the Machine Learning Stat*e, *supra* note 16, at 5 ("In Houston, a teachers' union brought an action against an algorithmic tool used to evaluate job performance and determine discharges on due process grounds. In Arkansas, state disability recipients filed suit against the Arkansas Department of Human Services alleging that an "unlawful switch to the computer algorithm" had violated the state's administrative procedure act.").

[81] *See*, *e.g.*, Barry v. Lyon, 834 F.3d 706 (6th Cir. 2016) (upholding the district court's determination that the automatic disqualification of food assistance violated, inter alia, constitutional and statutory due process requirements).

[82] RASHIDA RICHARDSON, JASON M. SCHULTZ & VINCENT M. SOUTHERLAND, LITIGATING ALGORITHMS 2019 US REPORT: NEW CHALLENGES TO GOVERNMENT USE OF ALGORITHMIC DECISION SYSTEMS 10 (AI Now Institute ed., 2019) https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf.

[83] *Id.*

[84] Houston Fed'n of Tchrs., Local 2415 v. Houston Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1179–80 (S.D. Tex. 2017) ("[W]ithout access to SAS's proprietary information—the

Incorporating a human in the loop can make such procedural challenges more difficult. Accordingly, there is a strong incentive to have a human play a role—whether justificatory, corrective, or accountability-based—in any U.S. decision that may implicate due process rights.

C. Discouraging

Law can also discourage the presence of humans in the loop. Here, we identify three mechanisms where law implicitly discourages including humans: by remaining silent in the face of countervailing pressures; by creating background legal obligations to maximize performance, which often take the form of a fiduciary duty; and via liability rules which require meeting standards of care that can only be accomplished by algorithms.

1. Silence

In the face of background realities, legal silence can effectively discourage the presence of humans in many algorithmic systems. Efficiency gains provide their own inherent incentive in most contexts; performing better, or with fewer resources, is typically a good result from the system designer perspective.[85] While not all algorithms are designed with efficiency as the main goal—think customer service phone trees, designed to keep irate individuals engaged while they await a human contact—those which foster efficiency in performance-focused environments may incentivize keeping humans out of the loop.

If there is a benefit to making decisions with superhuman speed, as with high-frequency stock trading or defensive cyberoperations, efficiency pressures discourage involving humans. Take stock trading: The vast majority of trades are made entirely algorithmically, with decisions made in millionths of a second—speeds no human could even imaginably reach.[86] These decisions take relentless advantage of

value-added equations, computer source codes, decision rules, and assumptions—EVAAS scores will remain a mysterious 'black box,' impervious to challenge.").

[85] Efficiency is not always the goal; prioritizing leisure motivates some workers to be less efficient. We take efficiency as a common goal nevertheless.

[86] Merritt B. Fox, Lawrence R. Glosten & Gabriel V. Rauterberg, *High-Frequency Trading and the New Stock Market: Sense And Nonsense*, J. APPLIED CORP. FIN. 29, Feb. 20, 2018, at 30, 31. Indeed, speed is so important that tremendous sums are spent buying slightly faster access to markets, including putting the computers running algorithmic trades in buildings physically close to stock exchanges to reduce the infinitesimal lag of fiber optic communications. *See Wall Street's Secret Advantage: High-Speed Trading*,

arbitrage opportunities.[87] For those with the capacity to use high-frequency trading algorithms, putting humans in the loop of quotidien trading decisions would result in an enormous performance hit.

Similarly, the sheer scale of certain decisions—either in terms of the number of factors to be considered or the number of decisions that need to be made—may discourage human involvement. As of February 2020, for example, over 500 hours of video were uploaded to YouTube every minute.[88] With some significant exceptions (notably including intellectual property law and sex trafficking), the platform is not liable for the content its users post.[89] Nor, however, is it liable for taking down legal content.[90] This liability shield, known as Section 230 of the Communications Decency Act, gives platforms enormous leeway to filter or not filter content as they deem best. Between January 2020 and March 2020, Google removed 6,111,008 videos in violation of its Community Guidelines,[91] which include categories such as spam, child safety, nudity or sexual, and violent or graphic.[92] Of these, 5,901,241 videos were removed through automated flagging.[93] To do so with humans would destroy Google's business model.

By remaining silent, law permits other forces to push humans out of decisionmaking systems.[94] But that is a choice; law needn't take that

WEEK (Jan 11. 2015), https://theweek.com/articles/493238/wall-streets-secret-advantage-highspeed-trading.

[87] *See* Fox, Glosten & Rauterberg, *supra* note 86, at 32, 38.

[88] *Hours of Video Uploaded to YouTube Every Minute as of February 2020*, STATISTA, (Sept. 14, 2021), https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

[89] 47 U.S.C. § 230.

[90] *Id.*

[91] *YouTube Community Guidelines Enforcement: Removed Videos by the Numbers*, GOOGLE, https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2020Q1;exclude_automated:all&lu=total_removed_videos (last visited Jan. 28, 2022).

[92] *YouTube Community Guidelines Enforcement: Removed Videos by Removal Reason*, GOOGLE, https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2020Q1;exclude_automated:all&lu=videos_by_reason&videos_by_reason=period:2020Q1 (last visited Jan. 28, 2022).

[93] *YouTube Community Guidelines Enforcement: Removed Videos by the Numbers*, GOOGLE, https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2020Q1;exclude_automated:all&lu=total_removed_videos (last visited Jan. 28, 2022).

[94] *Cf.* Joel Reidenberg, *Lex Informatica*: *The Formulation of Information Policy Rules through Technology*, 76 TEX. L. REV. 553 (1997) (discussing how different regulatory modalities influence each other), LAWRENCE LESSIG, CODE 2.0 (2006) (same). The absence of human involvement may be less controversial when it's possible to correct problems that occur at speed or at scale. After the 2010 flash crash, for example, regulators were

tack. In the face of the huge amount of effort expended in enabling high-frequency trading,[95] one could plausibly imagine a world where the SEC required that a human sign off on every trade. If a society prioritized quality content over quantity, social media companies could be required to engage in more particularized review of approval decisions. In any context where algorithms are thought to do things more cheaply, more efficiently, or more precisely than a human, there will be background pressure to keep humans out of the loop. The social decision to maintain legal silence implicitly fosters human exclusion.

2. Fiduciary Duties

Whenever there is some benefit to be gained from algorithmic efficiency or speed, law's silence will foster eliminating human influence. Further, fiduciary duties to maximize returns—such as the duty owed to corporate shareholders to manage corporate assets or stock broker duties of best execution owed to customers—strengthen the efficiency effect.

Fiduciary duties may add legal heft to efficiency incentives by at least nominally requiring fiduciaries to pursue efficiency and performance on behalf of their principals. The hard law threat of such duties may be relatively light—courts often defer to the judgment of fiduciaries in contestable cases[96]—but the soft law implications may matter. Fiduciary duties strengthen commercial norms that promote seeking performance and efficiency over values such as dignity or fairness,[97] and they can serve as a buffer to public critique for seemingly cold-hearted decisions. While Martin Shkreli was lambasted for trotting out fiduciary duties to

able to turn back time and reset the market; today, financial markets include various "tripwires" that stop trading when algorithms start acting unusually. Still, this capability is far from a determinative factor: the inability of content moderation algorithms to detect and prevent the spread of misinformation has certainly not prevented their use.

[95] *See* Eric Budish, Peter Cramton & John Shim, *The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response*, 130 Q.J. ECON. 1547, 1548 (2015) (characterizing high-frequency trading as "a never-ending socially wasteful arms race for speed").

[96] *See e.g.*, City of Warren Gen. Employees' Ret. Sys. v. Roche, No. CV 2019-0740-PAF, 2020 WL 7023896 (Del. Ch. Nov. 30, 2020), at *20 ("'Because fiduciaries . . . must take risks and make difficult decisions about what is material to disclose, they are exposed to liability for breach of fiduciary duty only if their breach of the duty of care is extreme.'").

[97] *See* Quadrant Structured Prod. Co. v. Vertin, 102 A.3d 155, 171–72 (Del. Ch. 2014) ("When determining whether directors have breached their fiduciary duties, . . . '[t]he standard of conduct for directors requires that they strive in good faith and on an informed basis to maximize the value of the corporation for the benefit of its residual claimants, the ultimate beneficiaries of the firm's value.'").

defend his price-gouging on life-saving medications,[98] implementers of AI systems may receive more credit for arguments that they had to cut humans out of some algorithmic decision loops to do right by their shareholders.[99]

3. Regulatory Arbitrage

While some forms of regulatory arbitrage might foster including a human in the loop, in other contexts it discourages their inclusion. To the extent having human involvement might implicate additional regulatory burdens, for example, regulated entities will automate decisionmaking processes. The National Security Agency, for example, reportedly minimized human oversight of surveilled material, reasoning that if no human was involved, reviewing and classifying gathered data wouldn't constitute a "search" possibly implicating the Fourth Amendment.[100] Scholars have argued that automated computer analysis of personal data online doesn't violate privacy laws in the way human involvement would.[101] Similarly, YouTube allegedly automates the flagging and takedown of copyrighted material, allegedly because involving a human would trigger a higher expectation of a fair use analysis.[102]

4. Liability Rules

In addition to encouraging human inclusion,[103] liability rules can also function to discourage including a human in the loop, especially when

[98] Dan Diamond, *Martin Shkreli Admits He Messed Up: He Should've Raised Prices Even Higher*, FORBES (Dec. 3, 2015), https://www.forbes.com/sites/dandiamond/2015/12/03/what-martin-shkreli-says-now-i-shouldve-raised-prices-higher/?sh=55536d471362.

[99] *See, e.g., Kevin Roose, The Robots Are Coming for Phil in Accounting,* N.Y. TIMES (Mar. 6, 2021), https://www.nytimes.com/2021/03/06/business/the-robots-are-coming-for-phil-in-accounting.html (connecting AI replacement of workers with shareholder incentives).

[100] Richard Posner, *Our Domestic Intelligence Crisis*, WASH. POST. (Dec. 21, 2005) ("[M]achine collection and processing of data cannot, as such, invade privacy.").

[101] Bruce Boyden, *Can a Computer Intercept Your Email*, 34 CARDOZO L. REV. 669, 726 ([A]utomated processing does not pose any risk to privacy. . . . The [Wiretap] Act has always required at least the prospect of human review.") (2012); Matthew Tokson, *Automation and the Fourth Amendment*, 96 IOWA L. REV. 581 (2011) (arguing that purely automated processing does not violate the Fourth Amendment); *but see* Ryan Calo, *The Boundaries of Privacy Harm*, 86 INDIANA L.J. 1131, 1151 (2011) ("[A]utomated decisions can . . . constitute privacy harms . . . .").

[102] Katharine Trendacosta, *Unfiltered: How YouTube's Content ID Discourages Fair Use and Dictates What We See Online*, ELECTRONIC FRONTIER FOUND., https://www.eff.org/wp/unfiltered-how-youtubes-content-id-discourages-fair-use-and-dictates-what-we-see-online (Dec. 10, 2020).

[103] *See supra* Part III.B.

liability is tied to performance standards and the human hampers it.

For example, Michael Froomkin, Ian Kerr, and Joelle Pineau have suggested that medical malpractice liability will discourage the meaningful presence of humans in the loop for algorithmic systems in medicine.[104] They are concerned that once medical algorithmic systems reach a high enough level of performance, the use of such systems—and deference to their recommendations—will become the standard of care, such that deviations would place liability for consequent injury on the shoulders of any human physicians involved.[105] To be clear, this argument doesn't posit that humans will exit the loop entirely—care providers will still be involved—but their roles will be circumscribed by algorithmic recommendations and will grow less meaningful over time as deference increases and skills atrophy. While this potential outcome appears to be some way off,[106] it highlights how background liability rules could gradually operate to push humans out of the loop.

D. Prohibiting

When a rulemaker determines that machine decisionmaking will always be preferable to human decisionmaking in a particular context, law may explicitly prohibit human involvement.

The idea of constraining human discretion through algorithms is not new. Mandatory sentencing laws, enacted initially to prevent judicial bias and standardize sentencing across judges, arguably provided a rudimentary algorithmic process meant to eliminate human capriciousness.[107] (Over time, however, caselaw reintroduced significant judicial discretion, including for deviations from the guidelines.)

[104] A. Michael Froomkin, Ian Kerr & Joelle Pineau, *When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning*, 61 ARIZ. L. REV. 33 (2019).

[105] *Id.*

[106] The use of AI is far from the current standard of care, and current liability rules are less friendly to algorithmic deference. W. Nicholson Price II, Sara Gerke & I. Glenn Cohen, *Potential Liability for Physicians Using Artificial Intelligence*, 322 J. AM. MED. ASS'N 173 (2019); *see also* W. Nicholson Price II, Sara Gerke & I. Glenn Cohen, *How Much Can Potential Jurors Tell Us About Liability for Medical Artificial Intelligence*, 62 J. NUCLEAR. MED. 15 (2020) (finding that physicians who deviate from non-standard of care AI recommendations are not yet viewed as liable for resulting injuries).

[107] *See* Rachel E. Barkow, *Recharging the Jury: The Criminal Jury's Constitutional Role in an Era of Mandatory Sentencing*, 152 U. PA. L. REV. 33, 85–86 (2003) ("The Guidelines and other mandatory sentencing laws dictate that specified facts will be deemed blameworthy as a general matter and establish punishment that will apply in all cases. . . . [T]here is little room for the trial judge to bend the law as a matter of justice or equity.").

Similarly, workers' compensation tables assign predetermined amounts for injuries that occur in the course of employment, without allowing for individualized tweaks.[108]

Likely due to the relative newness of algorithmic decisionmakers and familiarity with human ones, there are few explicit prohibitions on including a human in the loop, but given interest in curtailing the errors and discrepancies associated with human discretion, it is possible to imagine them being proposed in the future.

* * *

Law, whether old or new, implicit or explicit, profoundly shapes the roles of humans in the loop of algorithmic decisionmaking. Recognizing this reality is important both for scholars studying the regulation of algorithmic systems and for policymakers considering how best to govern them. Though older regimes matter for understanding the constraints and incentives for algorithmic systems, we turn our focus now principally to contemporary and forward-looking regulation—that is, policymakers actively considering the role of humans in the loop.[109]

III. THE MABA-MABA TRAP

There are two foundational—and related—problems with the law of the loop. First, even policymakers who intend to place a human in the loop rarely consider what role they want that human to play or what goals they want that human to accomplish. It's impossible to assess whether a human in the loop is improving a system's performance (or really whether regulation has done a good job of accomplishing its goals by adding a human) if we don't know what it is that the human is intended to contribute to the system. We address this problem in Part IV below.

Second, policymakers appear to think of human-machine systems as the sum of their parts, rather than understanding that hybrid systems are different and extraordinarily hard to get right. These two problems are interrelated: hybrid human-machine systems can't be designed well unless we know what a policymaker wants to accomplish by adding the human (problem one), *and* even if a policymaker can articulate why the

[108] *See, e.g.*, Tenn. Code § 50-6-207 (providing the compensation schedule under Tennessee's Worker Compensation Law).

[109] Some of the insights we generate are also applicable to older regimes; being clear about roles, for instance, matters in evaluating the impact of implicit mandates or incentives from older regimes. But they are not our focus.

human is there, humans in the loop will fail to accomplish policymakers' goals if the system isn't designed to avoid known points of failure (problem two).

Placing a human in the loop well is hard. Most policymakers don't seem to realize this.

To help explain why, in this Part we introduce legally minded audiences to the MABA-MABA trap. For over seventy years, a straightforward, easy, but problematic default view of human-machine systems has been to allocate tasks based on what "Men Are Better At" versus "Machines Are Better At"—MABA-MABA.[110] The original 1951 "Fitts list" identified, for example, that machines are better at performing repetitive, routine tasks, while humans are better at improvising.[111] MABA-MABA is attractive in part because there is an element of truth to it: there are things that humans and algorithms are respectively better at doing. But that's not the whole story. Adding or maintaining a human in the loop of an automated system creates well-known problems, many of which set up the system to fail and the human to shoulder the blame. MABA-MABA's attractive simplicity is thus a trap for regulators.

Nonetheless, MABA-MABA concepts pervade discussions of algorithmic regulation, in part because they do capture important truths. The first two Sections of this Part provide a summary of commonly observed strengths and weaknesses of both human and machine decisionmaking.[112] (Readers versed in these concepts should feel free to skip ahead.) Thanks to MABA-MABA, humans may be—understandably!—placed in the loop, whether by designers or policymakers, based on assumptions about their respective capabilities.

[110] Jones, *Ironies of Automation*, *supra* note 16, at 1. The more gender-inclusive "HEI-MEI" rapidly succumbed to the more fun-to-say "MABA-MABA."

[111] Paul Fitts created the seminal list in 1951; in the context of air traffic control systems, he created two columns that listed what "humans excel in" and what "machines excel in." Paul M. Fitts, Human Engineering For An Effective Air Navigation And Traffic Control System (1951). The more inclusive "HEI-MEI" ("Humans Excel In, Machines Excel In") rapidly succumbed to the more fun-to-say "MABA-MABA." The U.S. Department of Defense's 1987 adaptation of the Fitts list echoes of our observations above. U.S. Dep't Of Def., Human Engineering Procedures Guide, MIL-HDBK-763, 93 (1987) [hereinafter DoD Guide]; Jones, *Ironies of Automation*, *supra* note 16, at 105 (Machines are better at "[d]oing many different things at the same time"; humans are better at "[r]eacting to unexpected low-probability events"). The list remains debated today. *See, e.g.*, Joost C.F. de Winter & Dimitra Dodou, W*hy the Fitts List Has Persisted throughout the History of Function Allocation*, 16 Cognition, Tech. & Work 1 (2014).

But merely inserting a human in the loop does not necessarily result in the best of both human and machine. In fact, it can create or exacerbate well known problems for hybrid human-machine systems—problems policymakers currently largely ignore.[113] Getting humans to work well with algorithmic systems is far more difficult than it may first appear, as evidenced by the fact that an entire subfield of engineering focuses on these problems.[114]

In Part III.C, we introduce the special challenges of hybrid systems—complexities familiar in engineering,[115] but underdeveloped in the legal literature and apparently unknown to many policymakers. It is not enough to ask whether a human should be in a loop; regulators must also attempt to enable the human to interact effectively with the loop so that the system functions well as a whole. We provide three examples of law that takes these engineering concepts to heart: regulation of railroads, nuclear reactors, and medical devices.[116]

A. Human Decisionmaking

A vast and expanding literature explores precisely how humans make decisions;[117] here, we simply highlight a few relevant traits. Perhaps most obviously, humans are, well, human—and under some rubrics, the humanity of the decisionmaker is itself considered a positive, especially insofar as humans internalize social norms that inform their decisions.

Humans are flexible decisionmakers who can choose to deviate from strict rules and exercise discretion—though the exact degree of discretion depends on the organizational infrastructure and relative power. Humans can make contextual decisions, deciding when a rule must be bent, when to incorporate factors a machine might not have or

[113] Roth, *supra* note 16, at 1297 (criticizing MABA-MABA and calling instead for looking to systems engineering for ideas on how to better design human-machine systems).

[114] *See, e.g.,* DAVID D. WOODS, SIDNEY DEKKER, RICHARD COOK, NADINE SARTER & LEILA JOHANNESEN, BEHIND HUMAN ERROR (2nd. 2010); ERIK HOLLNAGEL & DAVID D. WOODS, JOINT COGNITIVE SYSTEMS: FOUNDATIONS OF COGNITIVE SYSTEMS ENGINEERING (2005); RESILIENCE ENGINEERING: CONCEPTS AND PRECEPTS (Erik Hollnagel, David D. Woods & Nancy Leveson, eds., 2006).

[115] *See infra* Part III.C.

[116] We later apply these lessons to the draft E.U. AI Act. *See infra* Part V.C.

[117] *See, e.g.*, CHOICES, VALUES, AND FRAMES (Daniel Kahneman & Amos Tversky, eds. 2000); DANIEL KAHNEMAN, THINKING FAST AND SLOW (2011); RICHARD H. THALER, MISBEHAVING: THE MAKING OF BEHAVIORAL ECONOMICS (2015); *Decision Theory*, STAN. ENCY. PHIL. (Oct. 9, 2020) https://plato.stanford.edu/entries/decision-theory/; Leigh Buchan & Andrew O'Connell, *A Brief History of Decision Making,* HARV. BUS. REV., Jan. 2006, at 32.

might categorize as out of scope, or when to make an analysis at a different level of generality to achieve a preferable result.[118]

Humans are also flexible decisionmakers in that we can generalize and jump across contexts, evaluating questions and applying principles in substantially different areas, such as when a judge jumps from a criminal to an antitrust case. This ability to reason across tasks and contexts is a considerable strength when compared to algorithms; humans can adapt to edge cases (for example, a human driver wouldn't fail to steer around a kangaroo at night, just because she had never seen one before). Human experts have "tacit knowledge"—knowledge that can't always readily be translated into code.[119] Finally, while our internal decisionmaking processes may be opaque, humans can be interrogated on and give reasons for their decisions—though the extent to which such reasons may be post-hoc rationalizations is hotly debated.[120]

Human weaknesses are also well-catalogued (and all too personally familiar). Humans are inconsistent, both individually and as groups: We often reach different conclusions, either because we weigh different factors differently or because we are affected by factors that should be irrelevant.[121] Some of these inconsistencies are due to the fact that we are biased—we are subject both to human decisionmaking biases, like saliency or recency,[122] and to personal prejudices. Relative to algorithms, humans are expensive decisionmakers: We are inherently limited resources who are often costly to train, slow to learn, and sluggish to act.

[118] Kaminski, *Binary Governance*, *supra* note 12, at 1546–47 (describing humans' ability to expand or contract the decisional context (including or excluding information that would be unfair to ignore or consider, respectively) based on cultural knowledge about appropriate decision processes).

[119] PASQUALE, NEW LAWS OF ROBOTICS, *supra* note 138, at 24 ("[W]e know more than we can explain.") (citing philosopher Hubert L. Dreyfus's theories of tacit knowledge).

[120] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEG. ANALYSIS 113, 116 (2018) ("A large body of research from behavioral science . . . tells us that people themselves may not know why and how they are choosing—even (or perhaps especially) when they think that they do.") (citing sources); *cf.* Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851, 1864 (2019) ("Reason giving is a core requirement in conventional decision systems precisely because human decisionmakers are inscrutable and prone to bias and error . . . .").

[121] *See, e.g.,* Pasquale, *A Rule of Persons, Not Machines*, *supra* note 8, at 49 ("[T]here are almost always conflicts among the approaches of multiple courts to similar sets of facts, irreconcilable by logic or reason."); Ozkan Eren & Naci Mocan, *Emotional Judges and Unlucky Juveniles*, 10 AM. ECON. J.: APPLIED ECON. 171, 173 (2018) (finding that juvenile court judges gave higher sentences, on average, the week after the local university unexpectedly lost a football game).

[122] DANIEL KAHNEMAN & AMOS TVERSKY, JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES 11 (1982).

Humans get tired. We get bored. We get hungry. We get injured and we fall ill. We (desperately) need vacations and mental health breaks. Human workers trigger various employment laws. All of these are reasons why institutions both private and public increasingly turn to automation and AI.

B. Algorithmic Decisionmaking

After dwelling for any amount of time on human frailties, it's easy for algorithms' relative strengths to dazzle. Upon closer inspection, however, their comparative weaknesses also become apparent.

First, algorithms are able to make decisions based on far more information and factors than a human would be able to take into account (although whether they analyze and how they weigh a particular piece of information will depend both on how they are designed and what data they can access).[123] Algorithms can store and process vast amounts of information—which, among other things, can be edited or deleted in a way human memory cannot.[124]

Algorithms are fast. They can reach conclusions based on multiple factors blazingly quickly: while the process of training an algorithm can take substantial amounts of time, the actual application to an individual decision can be effectively instantaneous. Algorithms are notably consistent: given the same inputs, they should reliably produce the same outputs without significant variation.[125] This has led some to argue that algorithms discriminate less than human decisionmakers.[126] Algorithms scale. They do not get bored making the same decision over and over and

[123] *E.g.,* Shannon E. French & Lisa N. Lindsay, *Artificial Intelligence in Military Decision-Making: Avoidng Ethical and Strategic Perils with an Option-Generator Model*, at 2 (manuscript on file with authors).

[124] DoD GUIDE, *supra* note 111, at 93; *see also* Jones, *Ironies of Automation*, *supra* note 16, at 105.

[125] There are some gaps in algorithmic consistency. For instance, many machine learning models, especially deep learning models, involve some randomness in the model training process; training the model different times on the exact same data, using the exact same parameters, may result in different final models unless the random element is held constant. Andrew L. Beam, Arjun K. Manrai & Marzyeh Ghassemi, *Challenges to the Reproducibility of Machine Learning Models in Health Care*, 323 J. AM. MED. ASS'N 305, 305 (2020). Nevertheless, once a developed or trained model has been implemented, the same inputs should consistently yield the same outputs.

[126] *E.g.*, Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519 (2018); Alex P. Miller, *Want Less-Biased Decisions? Use Algorithms*, HARV. BUS. REV. (Jul. 26, 2018), https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms. *But see* Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1696–1707 (2020) (considering these arguments and detailing "how bias may still creep into algorithmic decision-making systems").

over again,[127] and they are replicable. It is cheaper and faster to copy an oncology algorithm a thousand times than to train a thousand new oncologists.[128]

However, the weaknesses of algorithms are substantial. Any code-based system will be riddled with bugs—inevitable programming errors that cause unexpected and sometimes unwanted results. The more complex the system, the more likely it is that there will be accidents, as unforeseen interactions may create or exacerbate any single discrete error.[129] Algorithmic systems also introduce new vulnerabilities, insofar as they can be poisoned,[130] hacked, gamed, or otherwise exploited.[131]

Algorithms can be deeply weird or surprising.[132] They don't "think" like humans do (in fact, they don't think at all[133]). Algorithms rely on proxies—for both outputs (what constitutes a "good" employee?) and inputs (how do you measure success at work?).[134] Proxies, whether chosen by human programmers or derived from the data, can be incorrect, value-

[127] DoD GUIDE, *supra* note 111, at 93 (observing that machines excel at "[p]erforming routine, repetitive, or very precise operations").

[128] Kaminski & Urban, *supra* note 72, at 1968–69; Ajunwa, *supra* note 126, at 1734 ("[T]he fact remains that automated hiring is a cost-saving measure.").

[129] *Cf.* CHARLES PERROW, NORMAL ACCIDENTS: LIVING WITH HIGH-RISK TECHNOLOGIES (1999); Bryan H. Choi, *Crashworthy Code*, 94 WASH. L. REV. 39, 64 (2019)("Software complexity grows at an exponential rate, meaning that as the program size increases at a linear rate, the amount of computation needed to prove its correctness grows asymptotically toward infinity. While testing can locate some errors on a piecemeal basis, it cannot comb the entire universe of possible settings (or 'machine-states') that the software might encounter in the wild.").

[130] Paddy Smith, *Data Poisoning: A New Front in the AI Cyber War*, AI MAG. (Oct. 8, 2020), https://aimagazine.com/data-and-analytics/data-poisoning-new-front-ai-cyber-war ("Corrupting the training data leads to algorithmic missteps that are amplified by ongoing data crunching using poor parametric specifications. Data poisoning exploits this weakness by deliberately polluting the training data to mislead the machine learning algorithm and render the output either obfuscatory or harmful.").

[131] MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION 17–18 (2018), https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf (discussing data poisoning attacks, adversarial examples, and the exploitation of goals).

[132] *See generally* JANELLE SHANE, YOU LOOK LIKE A THING AND I LOVE YOU: HOW ARTIFICIAL INTELLIGENCE WORKS AND WHY IT'S MAKING THE WORLD A WEIRDER PLACE (2019) (detailing the weirdness of machine-learning outputs); *see also* WOODS, DEKKER, COOK, SARTER & JOHANNESEN, BEHIND HUMAN ERROR, *supra* note 114, at 216–19.

[133] Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. STATE UNIV. L. REV. 1305, 1308 (2019) ("The reality is that today's AI systems are decidedly not intelligent thinking machines in any meaningful sense.").

[134] *See generally* CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION 20–21 (2016); Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. STATE UNIV. L. REV. 1305, 1337 (2019).

laden, normatively undesirable, and even illegal.[135] And because of both the "black box" nature of some algorithms and the fact that proxies are often hard to detect, the use of algorithms can cloak normatively undesirable or illegal decisions in the garb of mathematical objectivity.[136]

Currently, algorithms are brittle: they may perform well in tasks and situations that are similar to those for and in which they were developed, but even the most advanced and flexible artificial intelligence systems quickly fail with even minor variations in the task or its context.[137] Algorithms thus lack the flexibility and adaptability of human decisionmakers. Algorithms also lack tacit knowledge, mentioned above, and thus miss crucial unarticulated (even inarticulable) aspects of human expert decisionmaking.[138]

Artificial intelligence is especially dependent both on its initial training data and data fed into the model: any errors, biases, or inadequacies will affect the system's structure and outputs. Reproducing biases, both in training data in particular and in society writ large, is a significant and much-discussed problem.[139] An algorithm might be "overfitted" to its training data, such that it produces highly accurate results with respect to that data set but performs poorly on new data, failing to accurately distinguish between relevant data and noise.[140] AI is also subject to the "long tail problem": Since training data will inevitably have more data on common scenarios than uncommon ones, edge cases are particularly hard for algorithms.[141]

It may go without saying that algorithms don't do norms or ethics well. Any concept that is contested or hard to articulate will be hard to

[135] Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 8, at 691–93 (on proxies and masking).

[136] Ajunwa, *supra* note 126, at 1686.

[137] Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. STATE UNIV. L. REV. 1305, 1332 (2019).

[138] FRANK PASQUALE, THE NEW LAWS OF ROBOTICS 24 (2020); Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. STATE UNIV. L. REV. 1305, 1325 (2019) ("[F]or many problem areas there is no easy way to identify or capture the relevant knowledge. In some cases, key concepts or abstractions cannot be meaningfully encoded in a computer-understandable form.").

[139] *E.g.*, NPR, *Joy Buolamwini: How Do Biased Algorithms Damage Marginalized Communities?* (Oct. 30, 2020), https://www.npr.org/transcripts/929204946; Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 8.

[140] *Overfitting*, IBM (March 3, 2021), https://www.ibm.com/cloud/learn/overfitting.

[141] Sasha Harrison, *How to Tame the Long Tail in Machine Learning*, SCALE (June 29, 2021), https://scale.com/blog/taming-long-tail.

translate into code[142]—even apparently "easy" rules like speed limits are subject to a host of coding decisions.[143] Algorithms that try to "learn" ethics from human behavior can import the nastier elements along with the good.[144]

Algorithms can be "black boxes" in ways that pose challenges for our current legal system—whether due to legal protections, deliberate secrecy, or as an inherent function of their design by reason of structure (neural nets) or complexity (just crunching more data than humans can monitor).[145] Whatever one's thoughts about the opacity of the human mind,[146] society has developed ways of querying it for purposes of assigning liability. Algorithms obscure human intent and responsibility.[147] Algorithmic failures can be difficult to identify and assess after the fact, such that some have proposed building in particular technological tools for establishing accountability.[148]

C. Human-in-the-Loop Decisionmaking Systems: Contributions from Engineering and Examples from Law

Humans are often placed in the loop based on these or similar assumptions about the respective strengths and weaknesses of humans and machines; it's a logical step. But MABA-MABA allocation has known flaws. It risks focusing on the individual human or machine components of a system without understanding how they interact with, hamper, or amplify each other.[149]

[142] Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. STATE UNIV. L. REV. 1305, 1332–33 (2019) ("AI tends to work poorly, or not at all, in areas that are conceptual, abstract, value-laden, open-ended, policy-or judgment-oriented; require common sense or intuition; involve persuasion or arbitrary conversation; or involve engagement with the meaning of real-world humanistic concepts, such as societal norms, social constructs, or social institutions.").

[143] *See* Shay, Hartzog, Nelson & Conti, *supra* note 15.

[144] For a recent example, see Matthew Gault, *Ethical AI Trained on Reddit Posts Said Genocide Is OK If It Makes People Happy*, VICE (Nov. 3, 2021), https://www.vice.com/en/article/v7dg8m/ethical-ai-trained-on-reddit-posts-said-genocide-is-okay-if-it-makes-people-happy.

[145] FRANK PASQUALE, THE BLACK BOX SOCIETY (2015); Price, *Regulating Black-Box Medicine*, *supra* note 18.

[146] *See*, *e.g.*, Huq, *A Right to a Human Decision*, *supra* note 2, at 640–46,

[147] Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 8, at 692–93; Ajunwa, *supra* note 126, at 1692–1707.

[148] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felton, Joel R. REidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).

[149] Roth, *supra* note 16, at 1297 ("Researchers have written in the systems engineering context about the dangers of so-called "MABA-MABA" lists. Instead, man-

"Human-centered" design evolved in partial response to these challenges.[150] A number of disciplines, including cybernetics, human factors engineering, and cognitive systems engineering, developed to address large-scale complex systems that involve passing off tasks between humans and machines. These disciplines made observations about common errors and developed a host of underlying principles, some of which we discuss below. Virtually none of this research has been considered by legal academics, policymakers, or practitioners,[151] let alone incorporated into the law of the loop.[152]

Ideally, a human-in-the-loop system would combine the best of both worlds: human flexibility could cushion algorithmic brittleness, algorithmic speed could swiftly resolve easy issues while leaving space for slower humans to weigh in on the harder ones, and algorithmic consistency and human contextuality would balance each other in appropriate equipoise. For some, such hybrid systems are the goal. Frank Pasquale argues that Intelligence Augmentation (IA), in which AI is used not to replace but to augment human capacities, "results in better service and outcomes than either artificial or human intelligence working alone."[153] AI assistants help us navigate the internet[154]; improve our personal shopping experience[155]; and offer medical advice to patients.[156] Human-machine "centaur" chess teams perform better than either human or algorithmic players acting alone; presumably, by presenting more options or by freeing humans from mundane or time-consuming tasks, algorithms may similarly assist humans to make more informed

machine interface designers should focus on what men and machines can do when enhanced by the other, and then ask, 'how do we make them get along better?' . . . Accordingly, we could think of our system as one of "trial by cyborg"— bionic experts, juries, and judges, rendering enhanced, humane justice."').

[150] Jones, *Ironies of Automation*, *supra* note 16, at 110.

[151] *But see* Roth, *supra* note 16 at 1296–98 (discussing this in the context of "trial by cyborg").

[152] The notable exceptions have largely been in regulations of automated transportation systems and nuclear reactors—cyberphysical systems with human operators that can crash or otherwise kill people. *See infra* Part V.

[153] PASQUALE, NEW LAWS OF ROBOTICS, *supra* note 138, at 13. *See also id.* at 29 ("A better frame is 'What sociotechnical mix of humans and robotics best promotes social and individual goals and values?'").

[154] *See, e.g.*, *How Do Search Engines Use Artificial Intelligence?*, LMGTFY, https://lmgtfy.com/?q=how+do+search+
engines+use+artifical+intelligence%3F&s=g (last visited Feb. 5, 2022).

[155] Rory Van Loo, *Digital Market Perfection*, 117 MICH. L. REV. 815, 817–22 (2019) (discussing the up- and downsides of the current and imminent proliferation of automated personal shoppers).

[156] Claudia E. Haupt, *Artificial Professional Advice, 18* YALE J. HEALTH POL'Y L. & ETHICS (2019).

evaluations or better concentrate on the elements of a decision that require distinctly human judgment in other contexts.[157]

But a hybrid system can all too easily foster the worst of both worlds, where human slowness roadblocks algorithmic speed, human bias undermines algorithmic consistency, or algorithmic speed and inflexibility impairs humans' ability to make informed, contextual decisions. Empirically, humans in the loop are often ineffective. Ben Green catalogs the variety of failures unique to this context,[158] including rubber-stamping humans that don't really oversee decisions,[159] the prevalence of "automation bias" that leads humans to defer overmuch to machines,[160] the deterioration of human skills based on automation known as "skill fade,"[161] incorporating or deviating from algorithmic advice in biased ways,[162] and the basic, tautological challenge of relying on humans to monitor the performance of systems designed to improve on human performance.[163] And sometimes, failure may be simply a mismatch of timing and biological unavailability; at a crucial moment, the human in the loop may be a human in the loo.

There might well be many times and places for humans in the loop.

[157] *Cf.* Paul Scharre, *Centaur Warfighting: The False Choice of Humans vs. Automation,* 30 TEMP. INT'L & COMP. L.J. 151, 154–55 (2016) (discussing the various roles humans play in target selection and engagement—essential operator, moral agent, and fail-safe—and arguing that automated assistants could allow human operators to focus on the latter two); Thomas Newdick, *AI-COntrolled F-16s Are Now Working As A Team in DARPA's Virtual Dogfights*, DRIVE (Mar. 22, 2021), https://www.thedrive.com/the-war-zone/39899/darpa-now-has-ai-controlled-f-16s-working-as-a-team-in-virtual-dogfights (discussing the benefits of AI/human teams).

[158] Green, *supra* note 8, at 14–18.

[159] Michael Veale & Lillian Edwards, *Clarity, Surprises, and Further Question in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling*, 34 COMP. L. & SEC. REV. 398, 400 (2017).

[160] Raja Parasuraman & Dietrich H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381, 390–97 (2010); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1271–72. (2008)

[161] Jones, *Ironies of Automation, supra* note 16, at 112 ("Automation leads to the deterioration of human operator skill, which needs to be more sophisticated to deal with novel and unique situations."); Lisanne Bainbridge, *Ironies of Automation*, 19 AUTOMATICA 775, 775–79 (1983); PETER FUSSEY & DARAGH MURRAY, POLICING USE OF LIVE FACIAL RECOGNITION IN THE UNITED KINGDOM, IN REGULATING BIOMETRICS: GLOBAL APPROACHES AND URGENT QUESTIONS (Amba Kak, ed.) 78–85 (2020), https://ainowinstitute.org/regulatingbiometrics.pdf.

[162] Megan T. Stevenson & Jennifer L. Doleac, Algorithmic Risk Assessment in the Hands of Humans (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440; Ben Green & Yiling Chen, Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments, Proc. Conf. Fairness, Accountability, & Transparency 90, 90–99 (2019), https://doi.org/10.1145/3351095.3372869.

[163] Green, *supra* note 8, at 14.

But regulators don't often address known problems, nor engage with known principles of system design. When a human is placed in the loop carelessly, there is a high likelihood that the human will be disempowered, ineffective, or even create or compound system errors. We call instead for a deliberate and systemic approach to human intervention, one in which the regulatory intervention fosters an environment where the human in the loop is empowered and able to accomplish their intended goals.

1. Notable Examples of Hybrid Failures

These failures are not hypothetical: There is a long if recent history of complex human-machine systems gone awry, from the Three Mile Island disaster to the Boeing 737 MAX crashes.[164] While it can be tempting to blame the humans involved, deploying a hybrid system "changes the nature of the errors that occur" from discrete human or machine error to system error.[165] Focusing on human-in-the-loop or human operator error can obscure bigger system design problems, in how errors can cascade, or how the system passes off tasks and designs interfaces between human and machine.[166]

The story of Three Mile Island, while dated, is an example of the kinds of systemic problems that can arise in human-machine systems. In 1979, the Three Mile Island Unit 2 nuclear reactor partially melted down.[167] At first glance, the Three Mile Island meltdown might look like human error: the human staff ultimately took actions that uncovered the reactor core,[168] and absent the staff's mistaken actions, the meltdown wouldn't have happened. But the reality was more complicated; prior to the staff's action, a mechanical failure had led to automated shutdowns of part of the system, compounded by *another* mechanical failure, compounded by erroneous instruments (showing a valve as closed that was stuck open) and missing instruments that could have shown whether the core was covered with water.[169] The accident caused a paradigm shift in the industry's understanding of the risks of human-machine systems, which recognized that the problem wasn't (solely) the human or (solely)

[164] WOODS, DEKKER, COOK, SARTER & JOHANNESEN, BEHIND HUMAN ERROR, *supra* note 114, at 1 (listing examples of failures of complex systems).

[165] Jones, *Ironies of Automation*, *supra* note 16, at 112.

[166] HANDBOOK OF HUMAN FACTORS AND ERGONOMICS 3 (Gavriel Salvendy ed., 2012); WOODS, DEKKER, COOK, SARTER & JOHANNESEN, BEHIND HUMAN ERROR, *supra* note 114, at 3.

[167] *Backgrounder on the Three Mile Island Accident*, NUCLEAR REGUL. COMM'N (June 21, 2018), https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.html.

[168] *Id.*

[169] *Id.*

the machine.[170] It was the interactions between the two, compounded by poor interface design, interface failures, and a lack of planning or training for this particular kind of emergency.

Lesson fully learned? Alas, no. It remains tempting to blame the human in the loop, even when they have been set up to fail.[171]

Consider the Boeing 737 Max aircraft.[172] In 2018 and 2019, two Boeing 737 Max planes crashed, killing 346 people. [173] Boeing initially blamed the flights' pilots.[174] Investigations then uncovered a design flaw in Boeing's automated flight-control system, known as Maneuvering Characteristics Augmentation System (MCAS).[175] Just as significantly, regulators uncovered systemic human and organizational failings: Boeing (and specific Boeing employees) had knowingly misled the Federal Aviation Administration (FAA) about MCAS, which resulted in pilot-training materials that lacked information about the system.[176]

[170] Woods, Dekker, Cook, Sarter & Johannesen, Behind Human Error, *supra* note 114, at 197.

[171] *Id.* at xi ("[A] lot of people concluded that the accident was caused by 'operator error,' by which they meant that the man who entered the wrong number had made an error, and that was all one needed to know. . . . The enlightened people said the failures had been made by the organization, which is to say by people such as managers and designers. Thereupon the startled management people cried, 'But we didn't enter the inappropriate numbers.' 'No, but you created the poor conditions for the entering of the numbers,' said the enlightened people.").

[173] *See, e.g.*, David Schaper, *Congressional Inquiry Faults Boeing and FAA Failures For Deadly 737 Max Plane Crashes*, NPR (Sept. 16, 2020), https://www.npr.org/2020/09/16/913426448/congressional-inquiry-faults-boeing-and-faa-fail ures-for-deadly-737-max-plane-cr; Peter A. Defazio, Rick Larsen & Staff of H. Comm. on Transportation & Infrastructure, 116th Cong., Final Comm. Report on the Design, Development & Certification of the Boeing 737 Max, (2020), https://transportation.house.gov/imo/media/doc/2020.09.15%20FINAL%20737%20MAX%2 0Report%20for%20Public%20Release.pdf; Boeing Charged with 737 Max Fraud Conspiracy and Agrees to Pay over $2.5 Billion, Dep't Justice (Jan. 7, 2021), https://www.justice.gov/opa/pr/boeing-charged-737-max-fraud-conspiracy-and-agrees-pay-over-25-billion.

[174] Douglas MacMillan, *'Our Daughter Died in Vain': What Boeing Learns from Plane Crashes*, Wash. Post (Oct. 28, 2019), https://www.washingtonpost.com/business/2019/10/28/our-daughter-died-vain-what-boeing-learns-plane-crashes/.

[175] Scott Neuman, *Indonesia Report: Pilots, Ground Crew Share Blame with Boeing for Lion Air Crash*, NPR (Oct. 25, 2019, 5:20 AM), https://www.npr.org/2019/10/25/773291951/pilots-ground-crew-share-blame-for-lion-air-737-max-crash -indonesian-report-says.

[176] Dep't Justice, *supra* note 172; Julie Johnsson, *Ex-Boeing Pilot Charged with Fraud in 737 Max*, Bloomberg, (Oct. 15, 2021, 6:36 PM), https://www.bloomberg.com/news/articles/2021-10-14/u-s-charges-ex-boeing-pilot-in-first-max-criminal-prosecution.

They uncovered interface design failings, too: Boeing had failed to design an effective interface between the MCAS and pilots.[177] The fatal flaw in the Boeing 737 Max was not (just) pilot error, and not (just) a faulty automated system, but also the failure (or really, fraud) by Boeing in not designing and training for effective interactions between human and machine.[178]

The human tendency to blame the human in the loop for accidents—as opposed to the humans who designed or fielded or failed to correct a loopy system—also manifests in the military context. Take the U.S.S. John McCain accident, the U.S. Navy's worst accident at sea in the past forty years. On August 21, 2017, the destroyer collided with another vessel, killing ten sailors, injuring forty-eight others, and sustaining hundreds of millions of dollars in damage to the ship.[179] After a new navigation system had proved prone to errors, the captain chose to employ it in manual mode; unknown to him, this removed various safeguards. The new configuration let different helmsmen unknowingly transfer steering control.[180] Although there was a notification regarding which station had steering control, the size and font type were so small that neither of two helmsman realized that the wrong station was steering the ship; ultimately, during an unrecognized transfer and mixup, the ship changed directions unexpectedly and appeared to be unsteerable, leading to the collision.[181] What result? To this day, "no one responsible for the development or deployment of the technology has faced any known consequences for the McCain disaster."[182] Quite the contrary: the Navy investigated, found the human captain at fault and charged him with homicide, and has committed nearly half a billion dollars to building and installing a modified version of the system on its destroyers over the next decade.[183]

Choosing which tasks to automate versus what to allocate to humans

[177] Neuman, *supra* note 175; *see also* Defazio, Larsen & Staff of H. Comm. on Transportation & Infrastructure, 116th Cong., *supra* note 172, at 90 ("Boeing initially considered adding an MCAS light on the flight control panel that would have illuminated in the event that MCAS failed to activate. The presence of an MCAS fail light on the flight control panel would have notified pilots of the presence of MCAS on the 737 MAX. Ultimately, however, Boeing rejected that idea.").

[178] Dep't Justice, *supra* note 172.

[179] T. Christian Miller, Megan Rose, Robert Faturechi & Agnes Chang, *Collision Course*, PROPUBLICA (Dec. 20, 2019).

[180] *Id.*

[181] *Id.*

[182] *Id.*

[183] *Id.* In response to fleet surveys, the variable touchscreens will be replaced with common physical throttle-and-wheel systems. *U.S. Navy to Ditch Touch Screen Ship Controls*, BBC NEWS (Aug. 12, 2019), https://www.bbc.com/news/technology-49319450.

is thus far more complicated than MABA-MABA would suggest.[184] Automation isn't a costless substitute for human decisionmaking: its use alters human roles and functions, sometimes unpredictably.[185] Some of these new roles tax humans with tasks that run into known human weaknesses, such as sustaining vigilance.[186] Others lean on humans to do more of a different kind of work, such as ensuring the algorithm isn't missing necessary parameters for accurate problem solving.[187] Not only do people adapt to the technology; they also adapt use of the technology to their changed and changing practices.[188] In short, "[t]he question for successful automation is not 'who has control over what or how much.' It is 'how do we get along together.'"[189]

2. Engineering Lessons for Success

Just as human factors and cognitive systems engineering identifies a set of errors that arise in hybrid systems, it also offers some lessons for success with regard to technological design, human facilitation, and organizational support.[190]

a. Technological design

Human-in-the-loop systems must be designed to promote effective interaction between the human and the system. At the very least, designers must consider how information will be transferred, how responsive the system will be to human input, and how the system handles failure.[191]

Well-designed interfaces are critical in hybrid systems. A poorly designed or inadequate interface creates opportunities for critical information to be garbled or lost in translation. For example, during the Three Mile Island meltdown, there were no sensors to detect how much

[184] Sidney W.A. Dekker & David D. Woods, *MABA-MABA or Abracadabra? Progress on Human-Automation Coordination*, 4 COGNITION, TECH. & WORK 5 (2002).

[185] *Id.* ("[A]utomation does not replace a human weakness. It creates new human strengths and weaknesses—often in unanticipated ways.") (citing Bainbridge, *supra* note 161).

[186] *Id.* at 5 (citing D.E. BROADBENT, PERCEPTION AND COMMUNICATION (1958)).

[187] *Id.* at 5 ("It also exacerbates the system's reliance on the human strength to deal with the parametrization problem (automation does not have access to all relevant world parameters for accurate problem solving in all possible contexts).").

[188] *Id.* at 6.

[189] *Id.* at 7.

[190] *See infra* Part XX.

[191] *Id.* at 7; *see also* Robin R. Murphy & David D. Woods, *Beyond Asimov: The Three Laws of Responsible Robotics*, IEEE INTELLIGENT SYSTEMS July–Aug. 2009, at 17–18.

water covered the nuclear reactor, and an instrument erroneously communicated that an open valve was closed.[192] Similarly, Boeing failed to design an effective interface between pilots and the MCAS system. Human factors engineering suggests a need to focus on human cognitive strengths—such as pattern recognition and responsiveness to change—in designing effective informational interfaces between human and machine.[193]

Machines must also be built to respond to useful human interaction.[194] What good is an informational interface if a human has no way to intervene? An algorithmic system might also need to be built for different responses to different humans in different roles. For example, a diagnostic algorithm might be built to differently process and respond to input by a doctor who notes an error during application to an individual patient; input by the head of a medical practice who determines that there has been a pattern of errors over multiple uses; and input by the hospital administration trying to determine hospital policy over when the algorithm should be used and by whom.[195] When there is a handoff, technology needs to be designed to smoothly transfer control to humans when doing so serves the goals of the system, both during ordinary use and in emergency settings.[196]

Technology should be designed to be resilient.[197] Failure, or at least encountering unpredicted events, is inevitable for complex systems.[198] Principles from resilience engineering include recording crash events using a black box for purposes of accident diagnosis and future

[192] *See supra* Part III.C; *see also* Elish, *Moral Crumple Zones*, *supra* note 78, at 40–42 (detailing how human-in-the-loop interface design issues contributed to the Three Mile Island and Air France Flight 447 disasters).

[193] Dekker & Woods, *MABA-MABA*, *supra* note 184, at 7.

[194] Murphy & Woods, *Beyond Asimov, supra* note 191, at 17–18.

[195] *See, e.g.*, Deirdre K. Mulligan, Daniel Kluttz & Nitin Kohli, *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*, *in* AFTER THE DIGITAL TORNADO (Kevin Werbach, ed., 2020).

[196] Murphy & Woods, *Beyond Asimov*, *supra* note 191, at 18 ("[B]umpy transfers of control have been noted as a basic difficulty in human interaction with automation that can contribute to failures.").

[197] Murphy & Woods, *Beyond Asimov*, *supra* note 191, at 17 ("Even if a specific disturbance is unpredictable in detail, the fact that there will be disturbances is virtually guaranteed, and designing for resilience in the face of these is fundamental."). For an incorporation of resilience principles into a legal framework, see Gary E. Marchant & Yvonne A. Stevens, *Resilience: A New Tool in the Risk Governance Toolbox for Emerging Technologies*, 51 U.C. DAVIS L. REV. 233 (2017).

[198] Waldemar Karwowski, *A Review of Human Factors Challenges of Complex Adaptive Systems*, 54 HUMAN FACTORS 983, 985 ("Perrow (1984) proposed that all accidents be viewed as 'normal' events in the sense that, given the complexity of system characteristics, multiple and unexpected interactions of failures are inevitable.").

debugging, designing a safe failure mode (such as "return home" or "stop movement"),[199] and designing for smooth transfer of control.[200] Sometimes a safe failure mode may require passing control back to a human; however, other times, such as when the reaction time required is very fast, it may situationally be safer to rely on an automated failure mode by the machine.[201]

b. Human dynamics

This brings us to the humans. As the Boeing 737-Max investigations show, it can be crucial to properly train humans in the loop, ensuring that they have not only relevant knowledge but also situational awareness: the ability to draw on a particular piece of knowledge in a particular situation when interacting with the machine.[202] Boeing should have required flight simulator training for use of its planes equipped with MCAS. It did not. (And now it does—after a set of flight simulator tests showed pilots used the wrong procedures for handling MCAS-related emergencies.[203]) Other known cognitive factors, such as the human tendency to use heuristics or oversimplify, have implications for training materials and practices,[204] especially for training non-experts to play a role in the loop. Even a factor as seemingly intrinsic as attention span can be influenced by training, technological design, and organizational policies.[205] Jones observes that in practice "the more advanced and reliable the automation, the more important the human operator must be."[206] Training can enable that operation.

Both interface design and training can affect automation bias, in which the human user of a system tends to overinflate the likelihood that the system is correct. Some tout cultivating undertrust as a solution to automation bias—that is, trying to engender *less* trust than might

[199] Choi, *supra* note 129; Murphy & Woods, *Beyond Asimov*, *supra* note 191, at 17.

[200] Murphy & Woods, *supra* note 191, at 18.

[201] *Id.* at 18 ("[W]hen conditions require very short reaction times, a pilot may not be allowed to override some commands generated by algorithms that attempt to provide envelope protection for the aircraft.").

[202] *Id.* at 12.

[203] Natalie Kitroeff & David Gelles, *In Reversal, Boeing Recommends 737 Max Simulator Training for Pilots*, NY TIMES (Jan. 8, 2020), https://www.nytimes.com/2020/01/07/business/boeing-737-max-simulator-training.html.

[204] WOODS, DEKKER, COOK, SARTER & JOHANNESEN, BEHIND HUMAN ERROR, *supra* note 114, at 13. *See also id.* at 19 (discussing training for flexibility under pressure).

[205] *Id.* at 14 ("[T]his control of attentional focus can be seen as a skillful activity that can be developed through training or supported (or undermined) by the design of artifacts and intelligent machine agents.") (citations omitted).

[206] Jones, *Ironies of Automation*, *supra* note 16, at 91.

otherwise be warranted.[207] Certainly, some degree of undertrust is useful: famously, undertrust may have averted what might have otherwise become World War III. In 1983, then-Lieutenant Colonel Stanislav Petrov of the Soviet Air Defence Forces decided that the Soviet early warning systems' report of launched U.S. missiles was probably a false alarm.[208] In contravention of his orders, he decided against responding with force—then sweated for twenty minutes before finding out he was right.[209] But one of the aims of introducing machine intelligence into a decisionmaking system is to identify relevant but counterintuitive factors—if human beings sometimes follow and sometimes disregard these counterintuitive results, the discrepancy in operator actions may result in more inconsistency than a purely human system[210] or decreased performance when humans overrule correct but counterintuitive algorithmic decisions.[211] Further, undertrust can also contribute to accidents. Due to its repeated glitches, the USS John McCain captain quickly learned not to trust the new navigation system, and so he often used it in backup manual mode.[212] Unfortunately, this practice created new risks, as the manual mode disabled built-in safeguards, the lack of which made the *McCain* disaster possible.

c. Organizational Dynamics

We close with a few observations about organizations—the blunt end of the human-machine system, outside of the loop that policymakers usually target. In complex human-machine systems, tales of the "isolated blunders of individuals mask the deeper story—a story of multiple contributors that create the conditions that lead to operator errors."[213]

[207] *See, e.g.*, Karen Hao, *We Need to Design Distrust into AI Systems to Make Them Safer*, MIT TECH. REV. (May 13, 2021), https://www.technologyreview.com/2021/05/13/1024874/ai-ayanna-howard-trust-robots/.

[208] Marc Bennetts, *Soviet Officer Who Averted Cold War Nuclear Disaster Dies Aged 77*, GUARDIAN (Sept. 18, 2017), https://www.theguardian.com/world/2017/sep/18/sovietofficer-who-averted-cold-war-nuclear-disaster-dies-aged-77.

[209] *Id.*

[210] Crootof, *Cyborg Justice*, *supra* note 15, at 244.

[211] Price, Gerke & Cohen, *Potential Liability for Physicians Using Medical AI*, *supra* note 106, at 1765.

[212] Miller, Rose, Faturechi & Chang, *supra* note 179. The phenomenon of "alert fatigue" is also well characterized in medicine. *See, e.g.,* Jessica S. Ancker *et al.*, *Effects of Workload, Work Complexity, and Repeated Alerts on Alert Fatigue in a Clinical Decision Support System*, 17 BMC MED. INFORMATICS & DECISION MAK. 36 (2017).

[213] David D. Woods & Richard I. Cook, *Perspectives on Human Error: Hindsight Biases and Local Rationality*, *in* 1 HANDBOOK OF APPLIED COGNITION 141, 143–44 (Francis T. Durso, ed., 1999) ("Rather than being the main instigators of an accident, operators tend to be the inheritors of system defects. . . . Their part is that of adding the final

That is, when individual humans in the loop fail, it is often because there are underlying problems with both technological design and organizational dynamics.[214] Organizational design and individual training and resourcing can make considerable differences in making the human in the loop more effective at whatever her task is—and making the system more resilient as a whole. For example, resilience engineering suggests that organizations should have in place a plan for failure, with checklists for contingencies.[215] Organizations shape human-machine interactions in many other ways; they select and empower (or not) humans for the loop, assign workloads, and help frame the decisions.[216]

Perhaps most significantly, organizations can play a large role in placing individual humans in the loop in a "goal conflict" or double bind. Failures in complex systems often occur when a human gets trapped in juggling hard-to-reconcile goals (such as safety versus cost).[217] "[C]onstraints imposed by organizational or social context can create or exacerbate competition between goals."[218]

For example, many states decided, in response to the opioid crisis, to institute prescription drug monitoring programs, implement them through an algorithm, and require physicians and pharmacists to consult the algorithm when prescribing controlled substances or filling such a prescription.[219] The company that created the algorithm, Appriss, was "adamant that a NarxCare score is not meant to supplant a doctor's diagnosis."[220] That is, Appriss insisted that doctors remain in the loop. However, doctors are in practice placed in a double bind: overruling the algorithm means risking prosecution for overprescribing or for prescribing to a patient deemed high risk. This means that—despite the fact that researchers worry about significant flaws in the screening tool, including flagging cancer patients as doctor-shoppers and building in bias against women—the humans in the loop are unlikely to do much to

garnish to a lethal brew whose ingredients have already been long in the cooking.") (quoting JAMES REASON, HUMAN ERROR 173 (1990)).

[214] *Id.*

[215] *Id.*

[216] Miller, Rose, Faturechi & Chang, *supra* note 179, at 15.

[217] *Id.* at 21 (observing that "goal conflicts played a role in the accident evolution, especially when they place practitioners in double binds").

[218] *Id.* at 22.

[219] Maia Szalavitz, *The Pain Was Unbearable. So Why Did Doctors Turn Her Away?,* WIRED (Aug. 11, 2021), https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/; *see also* Jennifer D. Oliva, *Dosing Discrimination: Regulating PDMP Risk Scores*, 110 CAL. L. REV. (forthcoming 2022) at 211, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3768774 (describing PDMP algorithms and physician reliance on them)

[220] Szalavitz, *supra* note 219.

correct it because of organizational dynamics.

3. Human Factors in the Law

While there has been little discussion of human factors engineering in policy debates on algorithmic decisionmaking, it has had important impacts in other legal fields. In fact, U.S. law already incorporates human factors research into the governance of some highly specified kinds of complex human-machine systems. We review three examples: railroad safety, nuclear power, and medical devices.

First, a few caveats: All three of our examples come from safety-critical systems, where expert government agencies regulate heavily because failures often result in physical injury or death. Such heavy and costly regulation may be arguably less warranted when consequences are not life-threatening. Our examples thus do not illustrate the array of ways in which regulatory systems might be designed differently, such as through legislation rather than regulation, through evolving doctrine, or through informal coordination. Also, our examples do not all address automated systems. They do, however, deal with complex human-machine systems that have long presented analogous concerns about human-machine interfaces, inadequate operator training, and failure cascades.[221]

a. Railroads

Our first example comes from railroads. The Federal Railroad Agency (FRA) has promulgated detailed regulation of the design of complex human-machine systems that draws explicitly and heavily on human factors engineering. It employs a combination of substantive standards and risk mitigation practices, or as the regulations describe it, both "criteria and processes."[222] It incorporates both permissive and mandatory human factors design criteria for consideration in designing and implementing certain signal and train control systems, including automated systems.[223]

[221] Jones, *Ironies of Automation*, *supra* note 16, at 92–93 (discussing, among others, railroad law as an example of the law of automation).

[222] 49 C.F.R. pt. 236, app. C(a): Safety Assurance Criteria and Processes (2021).

[223] 49 C.F.R. pt. 229, app. F: Recommended Practices for Design and Safety Analysis (2021); 49 C.F.R. pt. 236, app. E: Human-Machine Interface (HMI) Design (2021). *See also* 49 C.F.R. § 236.905: Railroad Safety Program Plan (RSPP)(b)(3) ("The RSPP must require a description of the process used during product development to identify human factors issues and develop design requirements which address those issues."); 49 C.F.R. § 236.907: Product Safety Plan (PSP)(a)(11) ("The PSP must include the following: (11) A

For example, the FRA requires product designers of Positive Train Control Systems, designed to automatically stop a train to prevent an accident, to address the "human factor engineering principle."[224] Product designers must consider human factors in designing a machine's interface with its human operator, including an operator's limited ability to process large amounts of information,[225] limited long term memory,[226] and expectation that there will be "consistent relationships between actions and results."[227] Designers must keep the human operator "in-the-loop" so that she does not lose situational awareness and is not overly reliant on the machine.[228] Designers must consider several rather detailed methods for keeping the human "in-the-loop": warning a human operator before she must take action; requiring a human operator to remain "in-the-loop" for a minimum of thirty minutes at a time; and providing "timely feedback to an operator regarding the system's automated actions, the reasons for such actions, and the effects of the operator's manual actions on the system."[229] Additionally, interface design shouldn't itself distract the operator from safety-related duties.[230]

Rather than assuming human perfection, or alternatively blaming humans for foreseeable problems, the FRA requires product designers to design complex technological systems around the persons operating them to maximize system success. The regulations are detailed, addressing the design of displays and controls and the design of communications to the human operator, including the operator's physical characteristics, education, and cognitive capacity.[231] Designers must "locate displays as close as possible to the controls that affect them"; "arrange controls

human factors analysis, including a complete description of all human-machine interfaces . . . .").

[224] 49 C.F.R. pt. 236, app. E(a). That is, designers "must sufficiently incorporate human factors engineering that is appropriate to the complexity of the product; the educational, mental, and physical capabilities of the intended operators and maintainers; the degree of required human interaction with the component; and the environment in which the product will be used." *Id.* at app. B(5).

[225] 49 C.F.R. pt. 236, app. E(c)(3) ("HMI design must… minimize an operator's information processing load," including by providing information "in a format or representation that minimizes the time required to understand and act" (a substantive standard) and conducting tests of such decision aids (a process)).

[226] *Id.* at (c)(4).

[227] *Id.* at (c)(2).

[228] *Id.* at (c)(1).

[229] *Id.*

[230] *Id.* at (c)(1)(v).

[231] *Id.* at (e)(9)("[D]esign display and controls to reflect specific gender and physical limitations of the intended operators."); *id.* at (f)(5) ("Where text is needed, use short, simple sentences or phrases with wording that an operator will understand and appropriate to the educational and cognitive capabilities of the intended operator[.]").

according to their expected order of use"; "group similar controls together"; and "design controls to allow easy recovery from error."[232] Communications must display only information necessary to the operator; emphasize its relative importance; display time-critical information in the center of the field of view and non-time-critical information in the lower right hand corner; use no more than seven colors; and show warnings designed to match the level of risk.[233]

The FRA addresses not just design but also training[234] and organizational policies.[235] The FRA also establishes risk-management processes, suggesting that companies assess complex human-machine systems—as systems—for risks of failure.[236] For example, companies should keep hazard logs and document assumptions about human performance so they can later be evaluated for accuracy.[237]

FRA regulations provide an example of how a regulator can consider systems-level human factors in regulating complex human-machine systems and create tremendously detailed rules for system designers and operators.

b. Nuclear Reactors

The Nuclear Regulatory Commission (NRC) also incorporates human factors in regulation of nuclear reactors, but in nothing like the FRA's detail. The NRC refers to human factors engineering in regulation but leaves the details to external documents, including related Guidelines,[238] the Three Mile Island (TMI) Action Plan Report conducted in response to

[232] *Id.* at (e)(1)-(9).

[233] *Id.* at (f)(1)-(13).

[234] *See, e.g.*, 49 C.F.R. § 228.411(b): Training (discussing training to reduce fatigue).

[235] 49 C.F.R. pt. 228, app. D: Guidance on Fatigue Management Plans.

[236] 49 C.F.R. pt. 229, app. F(f): Recommended Practices for Design and Safety Analysis ("An MTTHE [Mean Time to Hazardous Events] value should be calculated for each subsystem or component, or both, indicating the safety-critical behavior of the integrated hardware/software subsystem or component, or both. The human factor impact should be included in the assessment, whenever applicable . . . . ").

[237] *Id.* at (4)(i) ("[D]ocument any assumptions regarding human performance. The documentation should be in a form that facilitates later comparisons with in-service experience.").

[238] *See* U.S. NUCLEAR REGULATORY COMMISSION, HUMAN-SYSTEM INTERFACE DESIGN REVIEW GUIDELINES (2020), https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0700/r3/index.html; *see also* U.S. NUCLEAR REGULATORY COMMISSION, CLARIFICATION OF TMI ACTION PLAN REQUIREMENTS 3–51 (1980), https://www.nrc.gov/docs/ML0514/ML051400209.pdf.

the Three Mile Island disaster,[239] and specific company documents incorporated into certification standards by reference.[240]

Since the Three Mile Island disaster, certain applicants for reactor construction permits must provide the NRC with "a control room design that reflects state-of-the-art human factor principles."[241] These applicants are also required to demonstrate that they will establish a program for improving plant procedures, including "emergency procedures, reliability analyses, *human factors engineering*, crisis management, operator training, and coordination with . . . industry efforts."[242]

Rather than spell out relevant considerations in the regulations or appendices, the regulations reference the NRC's TMI Action Plan Report and subsequent related NRC Guidelines.[243] The NRC's 563-page Human-System Interface Design Review Guidelines, for example, provide in great detail guidelines for everything from interface displays and user-interface interaction to alarm systems and the design of workstations.[244] The guidelines go beyond human-machine interface design to reference operator training programs and processes for verification and validation.[245]

c. Medical Devices

Third and finally, FDA incorporates human factors engineering into the certification of certain Class II medical devices, which are required by regulation to perform human factors testing and analysis to "validate that the device design and labeling are sufficient for effective use by the

[239] U.S. NUCLEAR REGULATORY COMMISSION, NRC ACTION PLAN DEVELOPED AS A RESULT OF THE TMI-2 ACCIDENT (1980), https://www.nrc.gov/docs/ML0724/ML072470524.pdf; U.S. NUCLEAR REGULATORY COMMISSION, RESOLUTION OF GENERIC SAFETY ISSUES, TMI ACTION PLAN ITEMS (2012), https://tmi2kml.inl.gov/Documents/2c-L2-NUREG/NUREG-0933,%20Resolution%20of%20Generic%20Safety%20Issues,%20Section%201,%20TMI%20Action%20Plan%20Items%20(2011-12).pdf

[240] *See, e.g.,* 10 C.F.R. pt. 52, app. E(2)(a): Design Certification Rule for the ESBWR Design (incorporating documents provided in its GE-Hitachi Nuclear Energy application for certification of the Economic Simplified Boiling–Water Reactor (ESBWR) design).

[241] 10 C.F.R. § 50.34(f)(2)(iii).

[242] 10 C.F.R.§ 50.34(f)(2)(ii) (emphasis added).

[243] *Id.* (referencing U.S. NUCLEAR REGULATORY COMMISSION, *supra* note 238; U.S. NUCLEAR REGULATORY COMMISSION, *supra* note 239, at 3–51).

[244] U.S. NUCLEAR REGULATORY COMMISSION, HUMAN-SYSTEM INTERFACE DESIGN REVIEW GUIDELINES (2017), https://www.nrc.gov/docs/ML2016/ML20162A214.pdf. Section 9-1 on Automation Systems in particular contains detailed guidelines on Automation Displays (9.1), Alerts (9.2), Interaction and Control (9.3), Adaptive Automation (9.6) and more. *Id.*

[245] *Id.* at 9–8; B-16–B-17.

intended user."[246]

Like the NRC, FDA has not promulgated regulation specifying what such testing and analysis must entail. Instead, it has issued and revised guidance, most recently the Guidance on Applying Human Factors and Usability Engineering to Medical Devices.[247] This Guidance discusses human factors engineering as an aspect of risk management, directing companies to consider the role of device users, the use environment, and the device user interface.[248] It discusses processes such as validation testing and actual use testing.[249] Like other regulators, FDA directs companies to consider user training as an aspect of device operation and risk mitigation.[250]

Compared to both the FRA and NRC rules and guidance, the FDA Guidance contains more generalities and fewer specifics about interface design. For example, rather than tell designers how many colors should be used or where important information should show up on a screen, the Guidance states that interface design should be "logical and intuitive to use."[251] It counsels designers to look to graphic interfaces, elements that provide information to the user, and the logic of system interaction, but it doesn't tell designers more specifically what to do.[252] It emphasizes more strongly the use environment and variations among device users—presumably because both vary more for medical devices than for trains or nuclear reactors.[253]

* * *

These three examples from U.S. law showcase three possible models of how to regulate the design of a complex system with a human in the

[246] *See, e.g.*, 21 C.F.R. § 870.5200**:** External cardiac compressor. Similar requirements apply to 21 C.F.R. § 870.5210: Cardiopulmonary resuscitation (CPR) aid; 21 C.F.R. § 870.1415: Coronary vascular physiologic simulation software device; and elsewhere.

[247] FDA, APPLYING HUMAN FACTORS AND USABILITY ENGINEERING TO MEDICAL DEVICES: GUIDANCE FOR INDUSTRY AND FDA STAFF (2016), https://www.fda.gov/media/80481/download.

[248] *Id.* at 4, 7–11.

[249] *Id.* at 21, 28.

[250] *Id.* at 8, 24.

[251] *Id.* at 11.

[252] *Id.* at 10.

[253] *Id.* at 10 ("The lighting level might be low or high, making it hard to see device displays or controls. The noise level might be high, making it hard to hear device operation feedback . . . . The room could contain multiple models of the same device, component or accessory, making it difficult to identify and select the correct one. . . . The device might be used in a moving vehicle, subjecting the device and the user to jostling and vibration that could make it difficult for the user to read a display or perform fine motor movements.").

loop. Regulators could promulgate formal and detailed regulations dictating precise training requirements and the design of user interfaces. They could require consideration of human factors research as part of a licensing regime and issue accompanying detailed guidance on licensing standards. Or they could promulgate regulation as part of a licensing regime and issue even more generalized guidance on basic design principles, addressing a wider variety of users and environments. Each approach might be appropriate for a different regulatory environment.[254] And these models are certainly not the only way to address the human in the loop.

What is abundantly clear, however, is that merely declaring "there must be a human" will NOT set systems up for success or avoid failure. Instead, regulators are most effective when they detail the purpose of having a human in the loop—in these examples, to promote safety by correcting errors—and construct a regulatory regime to serve that end.

IV. THE ROLE OF THE HUMAN IN THE LOOP

The human in the loop is a tempting regulatory target, not least because they are an identifiable entity. But a myopic, MABA-MABA focus may obscure the larger, more important regulatory question animating calls to retain human involvement in decisionmaking processes. Namely, what do we want humans in a loop to *do*?

We identify six reasons to include a human in the loop: Humans may play (1) corrective roles to improve system performance, including error, situational, and bias correction; (2) justificatory roles to increase the system's legitimacy by providing reasoning for decisions; (3) dignitary roles to protect the dignity of the humans affected by the decision; (4) accountability roles to allocate liability or censure; (5) interface roles to link the systems to human users; and (6) "warm body" roles to preserve human jobs. None of these roles are mutually exclusive; to the contrary, humans in the loop may often fill multiple roles simultaneously.

A. Corrective Roles

Perhaps the most straightforward justification for humans in the loop

[254] To the extent that each of these approaches envision industry involvement, they are each potentially subject to agency capture, but that dynamic is outside our scope. *See* Nicholas Bagley & Richard L. Revesz, *Centralized Oversight of the Regulatory State*, 106 COLUM. L. REV. 1260, 1284–92 (2006) (describing and critiquing various versions of agency capture theory).

is corrective: Due to their as-yet-unique strengths, humans can improve the decisionmaking systems' accuracy.[255] Corrective roles come in at least three flavors: mine-run error correction, where the algorithm's decision is factually wrong; situational tailoring, where the algorithm's nominally correct determination is inaccurate in a particular context; and bias correction, where the algorithm's conclusion may be statistically accurate from the data it has been trained on, but it nonetheless reflects a systemic bias that runs counter to social values.

Of course, what constitutes a "right" decision may vary across contexts and evaluators. And, as discussed above, humans and algorithms are better at achieving different types of "right" decisions—humans are better at contextual analysis, while algorithms can consider more factors and better ensure that like cases are treated alike. But accuracy—in all its difficult-to-measure complexity—looms in the background of all "corrective" roles.[256]

1. Error Correction

Algorithms may be fast and cheap, but they make mistakes. Especially in the earlier stages of algorithmic development and use, humans are frequently involved in simply checking the algorithm's

[255] Accuracy is a critical factor in evaluating the utility of any decisionmaking system. However, an emphasis on accuracy brings its own complexity. False and true positives and negatives often differ in seriousness, and the frequencies of different errors are linked. For example, classifier performance can be characterized in terms of false positives, true positives, false negatives, and true negatives. If you want to catch more cases (e.g., identify more cancerous lesions among skin photos), you are looking to increase the rate of true positives. But this will also typically increase the rate of false positives (e.g., lesions classified as cancerous that are actually benign). False and true negatives are similarly linked, and the two-by-two frequency grid constitutes the aptly and delightfully named "confusion matrix." For a handy explainer, see Rachel Lea Ballantyne Draelos, *Measuring Performance: The Confusion Matrix*, GLASS BOX: MACHINE LEARNING AND MEDICINE (Feb. 17, 2019), https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/.

[256] The same is true in due process literature writ broad: accuracy is often cited as a or even the primary goal of affording due process rights. Thus when scholars discuss due process and algorithmic regulation, accuracy is a natural focus or goal. Kaminski & Urban, *supra* note 72, at *17 ("The Due Process Clause of the Fifth Amendment requires that '[n]o person shall . . . be deprived of life, liberty, or property, without due process of law.' In practice this requires notice and an opportunity to be heard 'appropriate to the nature of the case.' But why? A common answer is an instrumentalist one: to ensure accuracy. The Supreme Court has stated more than once that '[t]he function of legal process . . . is to minimize the risk of erroneous decisions.'") (internal citations omitted); Huq, *A Right to a Human Decision*, *supra* note 2, at 653–54.

results.[257]

As a cautionary tale of using algorithms *without* human error correction, consider the Michigan Integrated Data Automated System, also known as MIDAS.[258] The Michigan Unemployment Insurance Agency relied on MIDAS to identify and address welfare fraud; the system flagged individuals of fraud, sent automated questionnaires to frequently-unmonitored mailboxes, charged them with fraud, and (absent a response) began to garnish tax refunds and wages—all without any human involvement from agency staff.[259] Unfortunately, the system was prone to error; a later human audit found a 93% *error* rate.[260] In contrast, when there was a human reviewer, a mere (!) 44% of alleged frauds were found to be erroneous.[261] After a class-action lawsuit, the state committed to involving humans in every determination of fraud—that is, to ensure that there is always a human in the loop to catch and fix an algorithm's errors.[262]

2. Situational Correction

Alternatively, humans may improve system's outputs by tailoring an algorithm's recommendations based on population-level data to individual circumstances. As noted above, an algorithm's inherent brittleness and possible ineptness in addressing long tail events may result in inaccurate determinations in specific circumstances. A system that calculates the risk of a particular medical treatment, for example, may assume the availability of blood transfusions should it be needed—a perfectly reasonable assumption in most cases. But if the patient is a practicing Jehovah's Witness and morally opposed to blood transfusions or if the system is being used in an environment where blood transfusions are unavailable, that assumption would no longer hold. A human physician who knows the relevant contextual facts would (ideally) question the algorithmic system's risk estimation and adjust the

[257] Humans' role as error correctors often overlaps with their accountability role, as their supervisory position renders them the last entity able to affect an outcome. *See supra* Part II.B.4.

[258] Stephanie Wykstra, *Government's Use of Algorithm Serves Up False Fraud Charges*, UNDARK (June 1, 2020), https://undark.org/2020/06/01/michigan-unemployment-fraud-algorithm/.

[259] *Id.*

[260] David Eggert, *Michigan Reverses 44,000 Jobless Fraud Cases, Refunds $21M,* AP NEWS (Aug. 11, 2017), https://apnews.com/article/dc3370d57e264448b67f75ceb63ad120.

[261] *Id.*

[262] Maurice & Jane Sugar Law Center for Economic & Social Justice v. Arwood & Moffett-Massey, *Stipulated Order of Dismissal*, E.D. Mich., Case No. 2:15-cv-11449 (Feb. 2, 2017), https://www.bwlawonline.com/wp-content/uploads/2017/02/Zynda-ORD-2017-02-02-Robo-Fraud-Settlement-and-Dismissal.pdf.

algorithm's output or the physician's own behavior accordingly.

Human tailoring to improve outputs will be particularly important when a decisionmaking system is intended to prioritize individualized fairness over efficiency, "like-treated-alike" fairness, or other aims. Algorithmic-like criminal sentencing guidelines are efficient, but given that they do not take all mitigating factors into account, judges sometimes adjust their results at sentencing.[263]

Certainly, specific tailoring is easy to take too far; every individual circumstance is different, but that is not a justification for overturning generally applicable recommendations in all circumstances. If individual tailoring is the default, algorithmic systems lose the fairness benefits of treating like cases alike, the efficiency benefits of generally applicable recommendations, and—should the human introduce error—the accuracy benefits of high-performing algorithms.

3. Bias Correction

Humans may be also expected to correct algorithmic bias, which may manifest as prejudicial or inaccurate results.[264]Although some algorithmic systems were developed with the intention of providing an unprejudiced alternative to biased human decisions, research has persistently shown that many algorithmic systems are themselves deeply biased: they incorporate biases from their designers, from insufficient or unequally collected datasets, and from datasets that accurately reflect biases in reality.[265] In addition to biased results due to biased training

[263] *See* Susan R. Klein, *Movements in the Discretionary Authority of Federal District Court Judges over the Last 50 Years*, 50 LOY. U. CHI. L.J. 933, 957–58 (2019) ("The Court returned federal district judges much of their pre-1984 sentencing discretion in *United States v. Booker*. This decision generates more of an impact with each passing year. Judges are feeling freer to ignore the guidelines, almost always sentencing below the now-advisory range."). *Cf.* State v. Loomis, 2016 WI 68, ¶ 92, 371 Wis. 2d 235, 881 N.W.2d 749 ("COMPAS risk assessment may be used to 'enhance a judge's evaluation, weighing, and application of the other sentencing evidence in the formulation of an individualized sentencing program appropriate for each defendant.'. . . . 'COMPAS is merely one tool available to a court at the time of sentencing and a court is free to rely on portions of the assessment while rejecting other portions.'" (internal citations omitted).

[264] Kaminski, *Binary Governance*, *supra* note 12, at 1541. This goal is difficult, not least because what constitutes problematic "bias"—and thus what is needed to correct it—is contested. Selbst & Barocas, *Big Data's Disparate Impact*, *supra* note 8, at 714–15; Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 916–17 (2017).

[265] BENJAMIN, *supra* note 15, at 11 ("[B]ias enters through the backdoor of design optimization in which the humans who create the algorithms are hidden from view."); Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 8, at 677–92 ("Not only can

sets and designs, algorithmic decisionmaking systems also introduce "technical bias"—systemic inaccuracies that result from attempts to translate complex realities into crunchable code.[266] Accordingly, humans may be included in the loop to identify and counteract observed algorithmic biases. For example, AI-enabled facial recognition raises bias concerns because it has been shown to have a higher rate of inaccuracy for black women than, say, white men.[267] The draft AI Act would require two humans in the loop to verify AI identifications before they can be acted on.[268]

B. Justificatory Roles

Humans may also be included within a loop to justify decisions. Justification is often a crucial element of legitimacy; offering reasons for a decision help make it palatable to those affected by it.[269] For instance, it may be particularly important to the affected party to be provided a

data mining inherit prior prejudice through the mislabeling of examples, it can also reflect current prejudice through the ongoing behavior of users taken as inputs to data mining."); Huq, *A Right to a Human Decision*, *supra* note 2, at 647 ("[T]raining data, moreover, is generally not produced by an algorithm. It is a function of human action. As a result, it can replicate the biases and blind spots of the individuals who created it."); Lehr & Ohm, *supra* note 36, at 668 ("Inaccuracy and bias are paid much attention, and they can indeed be traced back in part to poor data and variable specifications."); Ngozi Okidegbe, *Discredited Data*, 107 CORNELL L. REV. (forthcoming 2022) (arguing that the data built from certain sources—namely, carceral knowledge sources—will necessarily be biased).

[266] Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM TRANSACTIONS ON INFO. SYS. 330, 333–36 (1996) (discussing how AI decisionmaking systems reach biased results due to a combination of (1) preexisting bias, due to biased training data sets and biased system design; (2) technical bias, which is caused by a system's limitations, including the loss of context and simplified formulations that attend any attempt to translate reality into code; and (3) emergent bias, which results from user interactions); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION 20 (2016) ("[M]odels are, by their very nature, simplifications. No model can include all of the real world's complexity or the nuance of human communication. Inevitably, some important information gets left out.").

[267] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. 77 (2018).

[268] Draft E.U. AI Act, *supra* note 1, Art. 14(5).

[269] Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 THEORETICAL INQUIRIES L. 83, 113; Tom R. Tyler, *Psychological Perspectives on Legitimacy and Legitimation*, 57 ANN. REV. PSYCHOLOGY 375, 376 (2006) ("Legitimation refers to the characteristic of being legitimized by being placed within a framework through which something is viewed as right and proper. So, for example, a set of beliefs can explain or make sense of a social system in ways that provide a rationale for the appropriateness or reasonableness of differences in authority, power, status, or wealth. This has the consequence of encouraging people to accept those differences.").

justification for the length of a prison sentence, the refusal to grant parole, or a hiring decision. Justification may also provide some transparency about how decisions are reached or allow for subsequent contestation.

But algorithmic systems often cannot supply satisfying reasons for their determinations; indeed, it is sometimes impossible even for those who design or regularly use certain algorithms to explain how they reach their conclusions. In some deep learning models, for instance, the algorithm's decisionmaking process may be too complex to explain or literally uninterrogable by human agents.[270] In addition, even where a purely algorithmic system *can* provide a reason, that algorithmic reason may not be sufficient to legitimate the decision in the minds of the decision's subject. Humans, on the other hand, can give reasons for their decisions, and including a human in the loop can enable the entire hybrid system to provide more satisfactory justifications.

This possible effect is not entirely hypothetical; a 2021 empirical study found that, as AI involvement in a legal decision increased, the perceived legitimacy of that decision decreased.[271] A human in the loop could potentially make a decision appear more legitimate, regardless of whether or not they provide accurate or salubrious justifications. Ideally, of course, the human in the loop would comprehend, interpret, and explain the algorithm's bases for recommendation. Otherwise, the human's explanation is little more than a post-hoc rationalization. However, some humans in the loop may play this deceptive, albeit still justificatory, role.[272]

[270] Price, *Regulating Black-Box Medicine*, *supra* note 18; Jenna Burrell, *How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y, Jan. 2016, at 1, 9 ("With greater computational resources, and many terabytes of data to mine (now often collected opportunistically from the digital traces of users' activities), the number of possible features to include in a classifier rapidly grows way beyond what can be easily grasped by a reasoning human.").

[271] Kirsten Martin & Ari Ezra Waldman, *Perceptions of the Legitimacy of Automated Decision-Making* 23 (working manuscript) (on file with authors), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3964900. However, whether the data for the decision was gathered specifically for a particular decision or aggregated by a third party was far more influential than the nature of the decisionmaker. *Id.* at 4–5, 21.

[272] *See* Brennan-Marquez, Levy & Susser, *supra* note 16, at 754 ("In some cases, the skeuomorphic human is not a Siri-esque humanoid interface, but a real flesh-and-blood person—albeit one who lacks any meaningful ability to influence the relevant decision-making process. In these cases, the human is effectively no more than an ornamental aspect of the system's interface.").

C. Dignitary Roles

Sentencing a criminal to a certain length of imprisonment, firing an employee, denying benefits to an individual with a contestable disability, determining which parent has custody over a child, and selecting military targets are all decisions that may profoundly impact someone's life, and having a human involved in that decision may help maintain the dignity of the individual affected by it. Accordingly, some argue that subjecting humans to purely algorithmic decisions violates their dignity, insofar as it treats humans as objects, fails to respect them as individuals rather than as members of groups, or does not adequately respect their autonomy.[273] Humans in the loop help obviate this concern by ensuring that there is a human element to the decisionmaking process. Indeed, some have characterized having a human in a decisionmaking loop as a "fundamental right."[274]

D. Accountability Roles

Some fear that humans will delegate difficult decisions to algorithms out of a desire to duck responsibility for undesirable outcomes;[275]

[273] *See* Kaminski, *Binary Governance*, *supra* note 12, at 1542–45 (summarizing and classifying these three dignitary arguments regarding algorithms); Jones, *The Right to a Human in the Loop*, *supra* note 28, at 230 (2017) ("Protecting American personhood has meant subjecting individuals and groups to as much accuracy, fairness and objectivity—computational neutrality—as possible.); Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 SETON HALL L. REV. 995, 1016–17 (2017) ("[W]hen faced with crucial decisions, a human should be treated with the dignity of having a human decision-maker address his or her personal matter.").

[274] *See* Jones, *The Right to a Human in the Loop*, *supra* note 28, at 230 (describing European "insistence on the categorization of . . . a human in the loop as a fundamental right"); EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS, GETTING THE FUTURE RIGHT: ARTIFICIAL INTELLIGENCE AND FUNDAMENTAL RIGHTS 60 ("Using AI-driven technologies broadly implicates the duty to respect human dignity, the foundation of all fundamental rights guaranteed by the Charter [of Fundamental Rights of the EU]. . . . AI-driven processing of personal data must be carried out in a manner that respects human dignity. This puts the human at the centre of all discussions and actions related to AI. Rather than the technology, the 'human being' creating and affected by the new technology needs to be the focus.").

[275] *Cf.* Rebecca J. Krystosek, *The Algorithm Made Me Do It and Other Bad Excuses: Upholding Traditional Liability Principles For Algorithm-Caused Harm*, MINN. L. REV. BLOG. (May 17, 2017), https://minnesotalawreview.org/2017/05/17/the-algorithm-made-me-do-it-and-other-bad-excuses/#post-2431 ("[H]owever else the law might shift to accommodate the proliferation of algorithms, legal liability should not be avoidable merely because an algorithm caused the harm, rather than a person."); Shailin Thomas, *Artificial Intelligence and Medical Liability (Part II)*, HARV. L. PETRIE-FLOM CTR.: BILL HEALTH BLOG (Feb. 10, 2017) http://blogs.harvard.edu/ billofhealth/2017/02/10/artificial-intelligence-and-medical-liability-part-ii ("[B]y decreasing the degree of discretion

accordingly, sometimes humans are included in the loop to ensure that someone is legally liable, morally responsible, or otherwise accountable for the system's decisions.[276] More cynically, sometimes the human is there to be the fall guy for an organization or for the algorithm's developer.

If the human in the loop has the power, information, judgment, and time to make the final decision in the human-algorithmic system, then the human can legitimately be held responsible. Consider clinical decision support software which makes recommendations to physicians, but specifies that every recommendation simply presents information which should be taken into account by the physician. As the system is envisaged, the physician maintains the final decisional authority—and consequently the moral and legal responsibility—for the final decision.

But responsibility can also be assigned to a human in a hybrid system who has no meaningful authority or ability to affect outcomes. M.C. Elish and Tim Hwang describe the concept of a human "liability sponge," where humans in the loop may "soak up" the legal and moral liability around a negative incident, including bearing the weight of tort liability, professional sanctions, or other opprobria.[277]

All humans in the accountability role—both legitimate ones and "liability sponges"—may simultaneously serve as a "moral crumple zone."[278] Both soak up liability, but the moral crumple zone explicitly does so to protect another entity: "[w]hile the crumple zone in a car is meant to protect the human driver, the moral crumple zone protects the integrity of the technological system, at the expense of the nearest

physicians exercise in diagnosis and treatment, medical algorithms could reduce the viability of negligence claims against health care providers."); Price*, Regulating Black-Box Medicine*, *supra* note 18, at 457 n.188 ("If an algorithm is unknown or impossible to disclose, under what context can physicians be liable for decisions relying on that algorithm? Is knowledge of the reliability of the algorithm sufficient to immunize against such liability?").

[276] *E.g.,* Bettina Berendt & Soren Preibusch, *Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop-and Under the Looking Glass*, 5 BIG DATA 135 (2017).

[277] M.C. Elish & Tim Hwang, *Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation*, Comparative Studies in Intelligent Systems—Working Paper #1 V2, 15 (May 18, 2015), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2720477.

[278] Elish, *Moral Crumple Zones, supra* note 78, at 41 ("Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a highly complex and automated system may become simply a component—accidentally or intentionally—that bears the brunt of the moral and legal responsibilities when the overall system malfunctions.").

human operator."[279] Not only does the human in the loop protect the system itself from censure, they also shield a host of remote decisionmakers who contributed to or may even have been better able to prevent the accident: the humans who designed, programmed, manufactured, purchased, or deployed the system.[280] U.S. judges, for example, regularly attribute tort liability for accidents involving robots to a person in the loop, rather than to a robotic system or relevant remote decisionmakers.[281]

E. "Warm Body" Roles

Concerns about technology displacing humans have long existed.[282] In response, humans are sometimes included in a loop to protect their jobs. Unlike other roles, humans can fulfill this one merely by being present; whether, how, or how well they contribute to the ultimate result is largely irrelevant.

Amidst ongoing debates about whether AI is going to replace certain types of physicians,[283] for instance, it is entirely predictable that the American Medical Association, the largest association of physicians in the United States, emphasizes the use of augmented intelligence rather than artificial intelligence. "Augmented intelligence" is "a conceptualization of artificial intelligence that focuses on AI's assistive role, emphasizing that its design enhances human intelligence rather than replaces it."[284] Doctors value their jobs;[285] ensuring a role for

[279] *Id.* at 40.

[280] This structure might seem deeply cynical, but the law actively facilitates the creation of moral crumple zones. *See supra* Part III.B.2.

[281] Calo, *supra* note 25, at 36; *see also* Kyle Graham, *Of Frightened Horses and Autonomous Vehicles: Tort Law and Its Assimilation of Innovations*, 52 SANTA CLARA L. REV. 1241, 1260-66 (2012) (discussing examples where early accidents involving cars or airplanes were often attributed to user error, rather than to the fact that steering devices unexpectedly detached or engines failed).

[282] Rather than being anti-technology, the much-maligned original "Luddites" were opposed to the ill-treatment of under-skilled laborers facilitated by the Industrial Revolution, as well as the tech-fostered reduction of overall employment. *E.g.* Cory Doctorow, *Science Fiction is a Luddite Literature*, LOCUS MAG., Jan. 3, 2022.

[283] *See, e.g.*, Sara Reardon, *Rise of Robot Radiologists*, 576 NATURE S54, S58 (2019) ("In the short term, AI algorithms are more likely to assist doctors than replace them."); Roxana Guilford-Blake, *Wait. Will AI Replace Radiologists After All?*, RADIOLOGY BUS. (Feb. 18, 2020), https://www.radiologybusiness.com/topics/artificial-intelligence/wait-will-ai-replace-radiologists-after-all (cataloging different viewpoints on the likelihood of AI replacing many radiologists).

[284] *Artificial Intelligence in Medicine*, AMA, https://www.ama-assn.org/amaone/augmented-intelligence-ai (last visited Sept. 14, 2021).

[285] Well, some do. Increasing rates of burnout in the medical profession are a

themselves within algorithmic systems is one way to protect them. Lawyers (and legal academics!) similarly emphasize the importance of keeping human lawyers involved in legal processes rather than relying fully on AI.[286] And fighter pilots push back hard against the idea that they can be replaced by drones.[287]

Frank Pasquale makes the point more broadly, arguing that a foundational principle of robotics should be that "[r]obotic systems and AI should complement professionals, not replace them."[288] In addition to corrective justifications, he argues that we must purposefully retain meaningful work for humans because it is important to both individual self-worth and community governance.[289] Pasquale emphasizes that the better role for AI is human "intelligence augmentation" rather than replacement, noting "the critical distinction between technology that replaces people and technology that helps them do their jobs better."[290] Further, Pasquale notes that our entire economic system depends on not fully automating human jobs: while human decisionmakers are expensive, those expenses ultimately power consumption, which in turn powers the economy.

One notable feature of the warm body role is that it prioritizes the value of the human in the loop, rather than humans on which the algorithmic system acts. Protectionism to save the jobs of doctors may be great—but not if the protected doctors injure more patients through their presence in the loop. Similarly, keeping human truckers driving will prevent automated trucks from decimating the trucking workforce—but could result in more accidents and costlier shipping. These outcomes are not necessarily *driven* by protectionism, but protectionism may obscure other goals that focus more on the performance of the system or its

substantial problem, and some, at least, hope that the addition of AI to medicine may create more space for human-centered interactions. ERIC TOPOL, DEEP MEDICINE: HOW ARTIFICIAL INTELLIGENCE CAN MAKE HEALTHCARE HUMAN AGAIN 18 (2019) ("Now, the highest-ever proportion of doctors and nurses are experiencing burnout and depression owing to their ability to provide real care to patients. . . . The greatest opportunity offered by AI is not reducing errors or workloads, or even curing cancer: it is the opportunity to restore the precious and time-honored connection and trust—the human touch—between patients and doctors.").

[286] *See, e.g.*, Jerry Levine, *Lawyers Can Be More 'Human' with the Help of AI. Here's How.*, ABOVE THE LAW (Sept. 23, 2021), https://abovethelaw.com/legal-innovation-center/2021/09/23/lawyers-can-be-more-human-with-the-help-of-ai-heres-how/.

[287] Hasard Lee, *F-35 Pilot: Forget Drones, the Skies Still Belong to Fighter Pilots*, SANDBOXX (June 14, 2021) https://www.sandboxx.us/blog/f-35-pilot-forget-drones-the-skies-still-belong-to-fighter-pilots/.

[288] PASQUALE, THE NEW LAWS OF ROBOTICS, *supra* note 138, at 3.

[289] *Id.* at 4.

[290] *Id.* at 12–13.

impact. On a broader scale, protectionism is likely to entrench the interests of those already empowered and involved in system design at the expense of non-incumbents and other stakeholders.

F. Interface Roles

Humans can also serve an interface role, helping users interact with an algorithmic system. Sometimes, it's just easier, cheaper, or faster to retain/insert a human link than to create a user-friendly interface.[291] For example, the human-facing customer service representative or tax advisor can input information into a specialized algorithm on behalf of another, suggest alternatives at decision points, and translate the system's jargon and conclusions.[292] Similarly, a physician may translate ambiguous patient-reported symptoms into formal medical terms for an algorithm.

Humans in these interface roles may not necessarily be "in" a loop; they may simply enter information into or report the results of an algorithmic system. A physician delivering momentous news, despite that news being purely reached via algorithmic means, may add an important human element to an algorithmic determination.[293] As Brennan-Marquez, Levy, and Susser argue, the perception that this human interface is "in the loop," even if that perception is inaccurate, may itself affect those impacted by the decision by making the system more intuitive to use or the results more palatable to swallow.[294] As a result, we consider this role something of an edge case; humans playing an interface role may actually be "in the loop" or they may merely appear that way.

V. RECOMMENDATIONS

What should one make of all of this? We do not pretend to offer a complete set of solutions here; instead, we offer three recommendations for those thinking about how the law might improve human-in-the-loop systems.

First, clarity about the roles of humans in the loop is crucial; interventions that add humans into loops or regulate human-involved systems will be haphazard so long as they lack a clear sense of what they are trying to do. Second, context matters tremendously for

[291] Brennan-Marquez, Levy & Susser, *supra* note 16, at 754–55.
[292] *Id.* at 754 (discussing the DMV clerk example).
[293] *Id.* at 755.
[294] *Id.*

implementation. While we have painted a broad picture of the law of the loop and related considerations, putting lessons into practice will require careful attention to the specific fields at issue. Generalities, standing alone, are at best little more than platitudes; at worst, they risk becoming influential but normatively problematic rules.[295] Third, governance should be systemic and attend to engineering principles; focusing too narrowly on just the human in the loop will frequently lead to failure. In this Part, we apply the principles gleaned from human factors engineering and related regulation above to the draft AI Act, and find, perhaps surprisingly, that the Act could be improved by taking cues from existing U.S. regulations.

Stepping back, policymakers must be aware that they are not creating law in a vacuum; while they may be the first to draft a rule that specifically targets a particular system, they must be aware that they are doing so against a backdrop of extant, generalist law that will also affect the human in the loop.[296]

A. Clarity of Roles

A key step in regulating human-in-the-loop systems is deceptively straightforward: when requiring human involvement, legislators, regulators, and other rulemakers should clarify what role(s) the human is supposed to play. Without understanding the desired role, designing systems for success, creating metrics to track that success, and evaluating success becomes substantially more difficult.[297] Conversely, identifying the intended role(s) fosters systems and organizational design that ensures the human in the loop has needed authorities and capabilities.

1. Why Clarify?

Ideally, rulemakers would explicitly state the human's intended role in the loop, but roles may also be inferred. For example, the stated goals of the draft AI Act's human oversight requirement for high risk systems are to "prevent[] or minimis[e] the risks to health, safety or fundamental

[295] *See supra* Part XXX.

[296] *See supra* Part XXX.

[297] *See, e.g.,* NAT'L RESEARCH COUNCIL, HUMAN ENGINEERING FOR AN EFFECTIVE AIR NAVIGATION AND TRAFFIC-CONTROL SYSTEM (Paul M. Fitts, ed. 1951) (prompting the field of function allocation research by announcing: "It appears likely, that for a good many years to come, human beings will have intensive duties in air navigation and traffic control. It is extremely important that sound decisions be made regarding what these duties should be.").

rights."[298] Inasmuch as human oversight is intended to correct errors that affect health or safety, these goals are largely corrective in nature. A central oddity of the Act, however, is that it uses a risk management and product safety framework for also addressing what are normally considered dignitary harms: harms to fundamental rights.[299] Thus the Act's human oversight requirement appears to be motivated by a mixture of corrective and dignitary goals.

Explicitness of purpose helps clarify what ability and agency a human in the loop must have. If the human in the loop is to serve an error correction role, they must be able to change the system's result. If they are to serve a genuine justificatory role, insight into the machine's decision is necessary—but not the ability to change it. And if the human is to serve as a liability sponge, perhaps their mere powerless presence is enough. Knowing what the human is meant to do is key to enabling their success.

When regulators are uncomfortable mandating that a human-in-the-loop system prioritize a certain role, they can still facilitate clarity by requiring those who design or field such systems be explicit about what *they* expect the humans to do. That is to say, if an automated truck requires a human alert at the wheel, it is useful for everyone involved to know whether that human is meant to correct algorithmic errors (presuming the human is better than the algorithmic system in emergencies), to serve a warm body paycheck role (mandated, presumably, by labor unions), or to serve as a liability sponge (though we suspect system designers will be loath to admit this).[300]

To the extent that designers may resist clarifying roles—or casting them accurately—one could imagine regulators offering carrots or threatening consequences. A well-defended characterization could win the benefits of a regulatory safe harbor, where the system was subject to less oversight or scrutiny. Meanwhile, a refusal to clarify roles or an apparent failure to do so accurately could result in fines or presumptions of bad faith in reviews or litigation.

However achieved, clarity would allow regulated entities to better comply with rules, evaluators to better assess systems, and critics to

[298] Draft E.U. AI Act, *supra* note 1, art. 14(2).

[299] Veale & Zuiderveen Borgesius, *supra* note 53, at 103 ("In data protection law, human oversight typically relates to human dignity. In the Draft E.U. AI Act, human oversight instead relates to minimising risks to health, safety and fundamental rights.").

[300] While we focus on the actions of policymakers, explicitness about human-in-the-loop goals on the part of system designers would also benefit system users and evaluators.

argue about role priority and success.

2. Role Complexities

Identifying a role will often be more complex, not least because multiple roles may be implicated and may require balancing. This balancing of different goals for humans in the loop can result in concerning outcomes: a human might be included in the loop despite a profound performance hit, sacrificing accuracy for dignitary aims (debatably worthwhile) or as a liability sponge (probably problematic). Alternately, humans can be nominally included in the loop in such a way as to have essentially no performance impact—but to appear to fulfill a dignitary, justificatory, or whatever other rationale is being paid lip service. Brennan-Marquez, Levy, and Susser describe this dynamic in *Strange Loops*, where systems appear to have a human involved in decisionmaking, but the human is essentially powerless.[301] Consider the constrained-but-blamable clerk at the Department of Motor Vehicles, who can only rotely respond to complaints of inflexibility with "that's all the computer will let me do."

Another set of problematic interactions arise from faux goals. Warm body roles, for instance, are often cloaked in corrective arguments. Many professionals have a deeply vested interest in making sure their jobs are not automated away, and it is easy to argue that they must retain an error-correcting role. To the extent corrective and warm body roles dovetail—as they often currently do—they may usefully reinforce each other. But such cloaking is problematic for system design, as it limits the scope of possible policy responses. For example, in the absence of a desire to preserve a human job, an alternative response to corrective concerns might be to mandate better performance by the algorithms, at the potential cost of jobs—which would have the added benefit of minimizing the likelihood that the human who *is* in the loop introduces errors. Alternatively, it is beneficial to articulate these "warm body," job-preserving aims when they exist, to foster a transparent debate on the benefits and tradeoffs of retaining a human in the loop for the sake of preserving that human job. In short, whatever the complexities of roles being involved, the rationale(s) for including humans in the loop should be explicit.

B. The Importance of Context

Although the features of humans in the loop have some

[301] Brennan-Marquez, Levy & Susser, *supra* note 16, passim.

commonalities across fields, the importance of different roles varies by context. Dignitary and justificatory rationales are less important (though non-negligible) when the relevant decisions do not impinge the dignity of human beings. For example, incoming missiles shot down by an automated defense system have no particular dignitary claim to a human making the targeting decision.[302] Using AI to optimize telecommunications networks and reduce energy use arguably also has minimal direct impact on human dignity.[303]

In other contexts, however, dignitary and justificatory rationales may weigh heavily. For example, even if sentencing algorithms could be made more accurately predictive of recidivism than a human judge (an immense if!), the dignitary and justificatory value of having a human judge involved might counsel in favor of retaining a central role for humans in sentencing processes.[304] This is arguably true of the judicial system more broadly speaking. In France, for example, automated decision-making is banned in the judicial context, and limited elsewhere in the administrative state.[305]

Context also matters in terms of determining what exactly human intervention brings to the table. The value of a human, for example, in overcoming bias depends on the relative bias of human decisionmakers and algorithms, which will change depending on field, human predilections, and the data available. Some humans are experts. Some have no training. Some are managers, and some are peons. Some work in systems that provide them both authority and reporting infrastructure or even whistleblower protection. Others are deeply embedded in the culture and rationales of a company using automated decisionmaking, such that inserting a human in the loop is very unlikely to depart from outcomes a company might prefer.[306]

[302] *See, e.g.*, Jen Kirby, *Israel's Iron Dome, Explained by an Expert,* VOX (May 14, 2021), https://www.vox.com/22435973/israel-iron-dome-explained (describing Israel's automated air defense system).

[303] *See, e.g.*, *AI in Networks*, ERICSSON, https://www.ericsson.com/en/ai (last visited January 17, 2022) (explaining how Ericsson, a telecommunications company, incorporates AI in its networks).

[304] *E.g.*, Kiel Brennan-Marquez & Stephen Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137 (2019); Crootof, *Cyborg Justice*, *supra* note 15, at 238–42.

[305] Gianclaudio Malgieri, *Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations*, 35 COMPUT. L. & SEC. REV. 105237 (2019).

[306] *See generally* ARI EZRA WALDMAN, INDUSTRY UNBOUND: THE INSIDE STORY OF PRIVACY, DATA, & CORPORATE POWER (2021) (discussing how corporate culture influences decisionmaking in privacy-related systems); Ari Ezra Waldman, *Privacy Law's False Promise*, 97 WASH. L. REV. 773, 807–15 (2020) (same).

This brings us to our larger point about context: whatever their goals, regulators need to widen the framing from the human in the loop to the system as a whole. Our relatively narrow definition of the human in the loop is driven by the focus of policymakers, but that precisely reflects the problem. Context includes not just the capacities and capabilities of particular humans, but entire organizational infrastructures within which they are embedded—and which are themselves shaped by humans. Humans generate the raw data used by analysts, who collect, curate, organize and otherwise transform it into training and input data. Humans are involved in the design, acquisition, and deployment of algorithmic systems. After a decision is made, individual humans are affected, and (should a route be available) they may appeal it; zooming out, humans are engaged in ex post decision oversight and system evaluation, which may entail responding to appeals, auditing results, and otherwise evaluating performance metrics. Based on these evaluations, humans may engage in structural updates and redesigns. All of these broader structures offer complexities in understanding and opportunities for regulatory intervention.

C. Learning from Engineering

With a clear view of what humans in the loop should be doing, policymakers should step back to consider hybrid systems from a systemic perspective, rather than focusing just on the humans themselves. In some sense, the definition we propound in Part I is inherently problematic: law focuses on the human in the loop for many reasons, but that focus obscures the larger systemic issues that need to be considered for effective governance. Whatever the goal, when a human in the loop isn't effective, it's often not their fault, but a failure of systems design.

1. Lessons for Regulating Human-in-the-Loop Systems

As our examples in Part III illustrate, regulators of safety-critical systems have for a while now concluded that if you're going to put a human in the loop, you need a LOT of implementing regulations to design the system so it isn't set up to fail. Such regulation aims to ensure the technology is designed and built for successful human interactions. It establishes resilience, both technically (through requiring fail-safe modes when applicable) and organizationally (through requiring or encouraging checklists or plans). It addresses human capacity and encourages training, including around issues such as automation bias. It at least nods to organizational factors, including how empowered a person is, what her workplace looks like, and what her incentives are during a crisis. It requires or encourages developers and/or users to test the system before it is used and keep records of its use and inevitable failings to figure out whether it is in practice working and what might be done to fix it, if not.

Unfortunately, regulators of modern algorithmic systems do not seem to have adopted these principles. As described above, many forms of the law of the loop come from the application of older, generally applicable rules. Others put a human in the loop without considering these systemic concerns. Still others play lip service to hybrid challenges but miss important aspects.

Consider the draft AI Act, which appears to comply with many of these principles. For example, the Act requires the providers of high-risk AI systems to design systems specifically for human oversight.[307] The soft law accompanying the Act emphasizes the necessity of educating and training the human providing oversight.[308] The Act, too, in many ways emphasizes systemic risk mitigation through ex ante design accompanied by ongoing monitoring–principles embraced by human factors design. At its core, the Act requires the providers of high-risk AI systems to build systems according to a set of standards[309] and conduct a conformity assessment before placing the systems on the market.[310] (Unlike the licensing systems in the regulations above, however, this conformity assessment is self-administered and self-certified.[311]) The Act tasks providers with performing ex ante risk assessments and risk mitigation.[312] It requires record keeping, with automatic logs for high-risk AI systems.[313] It requires providers to design a plan for post-market monitoring for incidents caused by high-risk AI systems[314] and to report serious incidents.[315]

However, viewed through a comparative lens with regards to the regulations discussed above, the draft AI Act has significant failings. At the core of these failings is that the Act divides regulated entities into providers, who build AI systems, and users, who use them. As a

[307] Draft E.U. AI Act, *supra* note 1, art. 14.

[308] *Id.* Recital 48.

[309] *Id.* arts. 41–42.

[310] *Id.* art. 43.

[311] *Id.* at 14 ("A comprehensive ex-ante conformity assessment through internal checks . . . .").

[312] *Id.* art. 9.

[313] *Id.* art. 12.

[314] *Id.* art. 61.

[315] *Id.* art. 62.

consequence, *nobody is really responsible for the human-machine system as a whole*. The providers must build the system in a particular way, and the users must follow instructions set out by providers, including instructions on human oversight.[316] But as Veale and Borgesius note, "[s]omewhat strangely, no obligations for human oversight flow directly from the Act to a user. In relation to human oversight, users must simply follow the instruction manual."[317] That is: there is no requirement that users of high-risk AI systems train a human tasked with oversight, nor that they prevent overstimulation in her environment, task her with staying in the loop for longer periods of time to prevent over-reliance, give her organizational authority, or design her work schedule to mitigate fatigue.

Moreover, the draft AI Act's human oversight design requirements for providers focus less on understanding and mitigating known human factors such as attention limitation or fatigue, or on designing an effective human-machine system, than on increasing the agency and power of the human in the loop. The Act dictates that high-risk AI systems "shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons."[318] The Act's requirements are broadly functional rather than detailed in nature: such tools shall enable individuals to: "fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation," "remain aware of … 'automation bias,'" "be able to correctly interpret the high-risk AI system's output," "be able to decide… not to use the high-risk AI system" and be able to stop the system.[319] By placing such an emphasis on enhancing the capacities of a human serving an "oversight" role, the Act might paradoxically end up overloading that human, or even setting up that human for blame should the system fail.

There remains, too, a question of whether the Act's reliance on *self*-assessment accompanied by monitoring will be adequate to prevent capture by the regulated industries.[320] By contrast, the safety-critical systems discussed above either involve direct regulation or an ex ante licensing scheme by the relevant authority which can refuse to license a particular product.

[316] *Id.* art. 29.

[317] Veale & Zuiderveen Borgesius, *supra* note 53, at 104.

[318] Draft E.U. AI Act, *supra* note 1, art. 14(1).

[319] *Id.* art. 14(4).

[320] Notably, high-risk AI systems that implicate product safety instead would follow an existing system of third party conformity assessments. *Id.* at 14.

Finally, a core challenge for the draft AI Act—and really, for any law that attempts to use human oversight to protect fundamental human rights—is how to validate and verify that the human in the loop is accomplishing desired goals.[321] How is impact on rights measured? What kind of expertise would humans need to have to be effective in their role in the loop? How could one person's take on contested concepts such as "fairness" and "discrimination" make that process legitimate? Writing regulations to effectuate simple goals (minimizing deaths from train crashes or nuclear meltdowns) is hard enough; how will regulators crystallize the complex governance needed to ensure that hybrid systems effectively and consistently protect dignity or ensure accurate justifications?

2. Complements and Alternatives to a Human in the Loop

Successfully placing a human in the loop is hard. If regulation of safety critical systems is any indication, it may require complex, detailed, even heavy handed regulation. This kind of regulation—its costs, its interventionism—works well only in particular contexts. And in fact many, including us, have argued that AI systems should *not* have such rigid regulation, but are better governed by more flexible, dynamic rules.[322] The goals regulators want to serve by placing a human in the loop may often be better served through other regulatory mechanisms.

Each of us has written at length about tactics and tools for regulating automated decisionmaking in different contexts.[323] These tactics can either replace or compliment the human in the loop.[324] For example, corrective goals such as reducing error and bias may be achieved through ex ante interventions on a systemic level, such as conducting risk assessments or imposing design goals and requirements, and ex post measures, such as auditing and performance metrics.[325] Justificatory

[321] *Id.* art. 14(2) ("Human oversight shall aim at preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter.").

[322] *See, e.g.,* Crootof, *Killer Robots are Here*, *supra* note 17; Kaminski, *Binary Governance*, *supra* note 12; Price, *Regulating Black-Box Medicine, supra* note 18; *see also* Crootof & Ard, *Structuring Techlaw*, *supra* note 24, at 399-413 (discussing the tradeoffs between more- and less-flexible legal designs and language).

[323] Crootof, *Cyborg Justice*, *supra* note 15; Crootof, *Killer Robots Are Here*, *supra* note 17; Crootof, *War Torts*, *supra* note 12; Kaminski & Urban, *supra* note 72; Price, *Regulating Black-Box Medicine*, *supra* note 18.

[324] *Accord* Green, *supra* note 8.

[325] Kaminski, *Binary Governance*, *supra* note 12, at 1595; Margot E. Kaminski &

goals may be addressed ex ante through systemic measures, such as requiring companies to articulate why they are using automated decisionmaking, and ex post, through disclosing such reasons to impacted individuals.[326] Both justificatory and accountability goals can be served by incorporating the voices of impacted stakeholders early in the design process, pre-implementation[327] and forcing actors to internalize the costs of employing such technologies.[328] Meanwhile, dignitary goals are furthered by establishing individual rights to contest certain automated decisions.[329]

Human oversight is no panacea. If anything, it creates new problems to solve. In some circumstances, it may well be worth solving those problems. In others, a human in the loop can be augmented with or supplanted by other, more effective forms of regulation.

CONCLUSION

> *Sometimes, when people make mistakes, they try to fix them. People who make mistakes sometimes try to fix them by putting humans in the loop. But if humans are put in the loop, they might make mistakes too.*[330]

Humans in the loop can play important roles in hybrid algorithmic systems. But humans aren't simply a regulatory or design patch, to be haphazardly inserted as a solution to problems that are really about the way the system as a whole is structured. Humans can fill any number of potentially useful roles, whether corrective, dignitary, or accountability, or something else—but they need to be situated and enabled to succeed in those roles. That requires knowing what those roles are , governing and considering the systems as a whole, and adjusting that governance over time.

Gianclaudio Malgieri, *Algorithmic Impact Assessments Under The GDPR: Producing Multi-Layered Explanations*, 11 INT'L DATA PRIV. L. 125 (2021); Price, *Regulating Black-Box Medicine, supra* note 18; Price, *Medical AI and Contextual Bias*, supra note 12.

[326] Hildebrandt, *supra* note 269, at 114–15.

[327] *See, e.g.*, Ngozi Okidegbe, *When They Hear Us: Race, Algorithms and The Practice of Criminal Law*, 29 KAN. J.L. & PUB. POL'Y 329 (2019); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L. J. 59 (2017) (arguing that "public discourse and input" can improve the construction of risk assessment tools); Kaminski, *Binary Governance*, *supra* note 12, at 1533–34; Kaminski & Malgieri, *Participatory AI Governance* (manuscript on file with authors).

[328] Crootof, *War Torts*, *supra* note 12.

[329] Kaminski & Urban, *Right to Contest*, *supra* note 72.

[330] tl;dr of this paper's abstract, algorithmically generated at http://tldrpapers.com without a human editor in the loop.

How law should deal with humans in the loop of algorithmic systems presents a vast and growing challenge as these systems proliferate. The right answers will require care and attention, informed by the literature on how human/machine systems and issues particular to specific contexts. Our goal here has been to raise the issues and to present considerations for policymakers working to improve governance going forward.