Humans in the Loop

*Rebecca Crootof**
*Margot E. Kaminski***
*W. Nicholson Price II****

From lethal drones to cancer diagnostics, humans are increasingly working with complex and artificially intelligent algorithms to make decisions which affect human lives, raising questions about how best to regulate these "human-in-the-loop" systems. We make four contributions to the discourse.

First, contrary to the popular narrative, law is already profoundly and often problematically involved in governing human-in-the-loop systems: it regularly affects whether humans are retained in or removed from the loop. Second, we identify "the MABA-MABA trap," which occurs when policymakers attempt to address concerns about algorithmic incapacities by inserting a human into a decisionmaking process. Regardless of whether the law governing these systems is old or new, inadvertent or intentional, it rarely accounts for the fact that human-machine systems are more than the sum of their parts: they

* Rebecca Crootof is Associate Professor of Law, University of Richmond School of Law.

** Margot E. Kaminski is Associate Professor of Law, Colorado Law School, and Director, Privacy Initiative, the Silicon Flatirons Center.

*** W. Nicholson Price II 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, Rabea Benhalim, Ana Bracic, Christopher Corts, Sue Glueck, Claudia Haupt, Fiona Illi, Izzy Longstaff, Andrea Matwyshn, Emily McReynolds, Aileen Neilson, Paul Ohm, Nadav Orien-Peer, Gabriel Rauterberg, Blake Reid, Nikkita Rivera, Andrew Selbst, Lawrence Solum, Sloan Speck, Elizabeth Stalfort, and Harry Surden. This paper benefited from generous feedback at the AALS 2020 Annual Meeting, the ETH Zurich and the Universities of Zurich and St. Gallen Workshop on the Law & Economics of Innovation, the Legal Scholars Roundtable on Artificial Intelligence, the Northwestern, Penn, and Stanford Law Schools' Junior Faculty Forum for Law and STEM, the Privacy Law Scholars Conference, 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, the Techlaw Junior Scholars Workshop, and the We Robot conference. We thank Nathan Fuller, Abbi Lynch, Phoebe Roque, Rylee Snively, and Angela Theodoropoulos for excellent research assistance. Nicholson Price's work was supported by the Novo Nordisk Foundation (Grant No. NNF17SA0027784). From each of us individually: All mistakes are my co-authors'.

raise their own problems and require their own distinct regulatory interventions.

But how to regulate for success? Our third contribution is to highlight the panoply of roles humans might be expected to play, to assist regulators in understanding and choosing among the options. For our fourth contribution, we draw on legal case studies and synthesize lessons from human factors engineering to suggest regulatory alternatives to the MABA-MABA approach. Namely, rather than carelessly placing a human in the loop, policymakers should regulate the human-in-the-loop system.

INTRODUCTION

Artificially intelligent algorithms are being integrated into decisionmaking processes at mind-boggling speed and scale.[1] Governments use artificial intelligence ("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 employing AI decision assistants to collect

[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 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*, at 39, annex I, COM (2021) 206 final (Apr. 21, 2021), (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*, 86 BROOK. L. REV. 791, 792 (2021) ("This article seeks to capture the state of the art in 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) [hereinafter Huq, *A Right to a Human Decision*] ("I focus on the direct state applications of machine-learning tools to individuals for the purpose of allocating benefits or burdens."); Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. 797, 800 (2021) (noting the trend in state and federal public benefits agencies towards incorporating automated systems); Hamsa Bastani, Kimon Drakopoulos, Vishal Gupta, Ioannis Vlachogiannis, Christos Hadjichristodoulou, Pagona Lagiou, Gkikas Magiorkinis, Dimitrios Paraskevis & Sotirios Tsiodras, *Efficient and Targeted COVID-19 Border Testing via Reinforcement Learning*, 599 NATURE 108, 108–09 (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, Brandon K. Fornwalt, Gino J. Mongelluzzo, Jonathan D. Suever, Brandon D. Geise, Aalpen A. Patel & Gregory J. Moore, *Advanced Machine Learning in Action: Identification of Intracranial Hemorrhage on Computed Tomography Scans of the Head with Clinical Workflow Integration*, NPJ DIGIT. MED., Apr. 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. INFORMATICS e15182 (2020) [hereinafter Sendak et. al., *Real-World Integration*] (discussing Duke Health's Sepsis Watch program).

[5] *See, e.g.*, Rebecca Robbins & 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/ [https://perma.cc/4MPC-UPUR] ("At a growing number of prominent hospitals and clinics around the country, clinicians are turning to AI-powered decision support tools — many of them unproven — to help predict whether hospitalized patients are likely to develop complications or deteriorate, whether they're at risk of readmission, and whether they're likely to die soon.").

[6] *See, e.g.*, PAUL SCHARRE, CTR. FOR A NEW AM. SEC., AUTONOMOUS WEAPONS AND OPERATIONAL RISK 46 (2016), https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf [https://perma.cc/2C5M-73YM] (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.

information, crunch data, assess threats, and recommend strategic moves or specific targets.[7] In these and a host of other fields—agriculture, commerce, advertising, 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, whether certain weighty or irreversible decisions should be delegated to nonhuman entities at all.[8] Without weighing in on these normative questions, we focus on the regulatory response.[9] Regulators frequently respond to these concerns by either explicitly requiring or implicitly encouraging placing a "human in the loop"—which we define

11, 2021), https://www.theatlantic.com/technology/archive/2021/09/i-weapons-are-third-revolution-warfare/620013/ [https://perma.cc/V2UX-CTV8] (raising concerns that AI advancements "will accelerate the near-term future of autonomous weapons" and bring major downsides); Gerrit D. Vynck, *The U.S. Says Humans Will Always Be in Control of AI Weapons. But the Age of Autonomous War Is Already Here*, WASH. POST. (July 7, 2021, 10:00 AM), https://www.washingtonpost.com/technology/2021/07/07/ai-weapons-us-military/ [https://perma.cc/EE6Q-X82B] (noting that weapons systems with autonomous capabilities are being developed).

[7] *See, e.g.*, Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 NAT'L SEC. L. & POL'Y 1, 4 (2019) (predicting that states will use machine learning to aid decisions about whether to use force against or inside another state); Ashley S. Deeks, *Predicting Enemies*, 104 VA. L. REV. 1529, 1530–31 (2018) (noting that leaders are encouraging the use of machine learning to improve military capabilities and decisionmaking).

[8] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016) [hereinafter Barcos & Selbst, *Big Data's Disparate Impact*] (raising awareness about data mining as a potential source of unintentional employment discrimination); Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519 (2018) (proposing that Title VII's antistereotyping theory be employed to ensure that algorithms are applied to counteract workplace discrimination); Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249 (2017) (arguing for preserving human judgment); Ben Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, 45 COMPUT. L. & SEC. REV., July 2022, at 1 (proposing a shift from human to institutional oversight for regulating governmental algorithms); Alex P. Miller, *Want Less-Biased Decisions? Use Algorithms*, HARV. BUS. REV., July 26, 2018, https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms [https://perma.cc/6U9Z-VR6J] (arguing for replacing humans with machines); Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1, 6, 44–54 (2019) [hereinafter Pasquale, *A Rule of Persons, Not Machines*] (suggesting a complementary approach for automation may better realize rule-of-law values than a substitutive approach); FRANK PASQUALE, NEW LAWS OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI (2020) [hereinafter PASQUALE, NEW LAWS OF ROBOTICS] (arguing for retaining humans and supporting them with machines); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 804–05 (2014) (critiquing the trend towards "evidence-based sentencing" in criminal sentencing on constitutional and policy grounds).

[9] We do not take a collective stance on these questions, both because our analysis highlights the context-specific nature of the inquiry, see *infra* Part V, and because we generally can't agree with each other, see *infra* note 13.

as an individual involved in a single, particular algorithmic decision.[10] In the United States, for example, over forty government policies now require human oversight or involvement in various algorithmic decisionmaking processes.[11]

But regulators often deploy humans sloppily, in ways that set up both the human and the greater human-machine system to fail. As algorithms are integrated into more and more decisionmaking processes, policymakers need better guidance on how to use law to foster productive hybrid decisionmaking.

Informed by our disparate areas of expertise[12] and starkly different baseline assumptions regarding the utility and drawbacks of human and algorithmic decisionmaking,[13] we make four generalizable contributions, applicable to regulating human-in-the-loop systems across contexts. We advocate neither for nor against delegating a decision to an algorithm or human; instead, we argue that if policymakers are going to attempt to regulate hybrid systems, they need to do a better job. As we detail, there are pitfalls to be avoided and strategies for crafting more effective regulations.

Our first contribution is identifying that there already is a "law of the human in the loop"—law that influences the inclusion and capabilities of humans who affect what would otherwise be purely algorithmic decisions.[14] Somewhat surprisingly, as of yet there has been

[10] 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. But it is far from the only possible one—and it is a problematic one. *See infra* Part I. As we discuss, more expansive definitions better highlight the myriad ways humans affect algorithmic systems. *See infra* Parts I & V.

[11] 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). *See id.*

[12] 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.

[13] For a "Pollyannaish, techno-utopian" take (per Rebecca Crootof), see W. Nicholson Price II, *Medical AI and Contextual Bias*, 33 HARV. J.L. & TECH. 65, 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 (2016), concluding that autonomous weapon systems will inevitably make mistakes, with horrific consequences; and Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529, 1538 (2019) [hereinafter Kaminski, *Binary Governance*], arguing that while human decisionmaking can be flawed, algorithmic decisionmaking is not necessarily a better replacement and relies, too, on often-flawed human decisions.

[14] The Draft E.U. AI Act, *supra* note 1, which would regulate AI systems across sectors, has been widely hailed as "the first-ever legal framework on artificial intelligence." Eve Gaumond, *Artificial Intelligence Act: What Is the European Approach for AI?*, LAWFARE (June 4, 2021, 11:50

no comprehensive evaluation of the law of the loop.[15] To the extent there is scholarship discussing relevant existing law, it has been largely

AM), https://www.lawfareblog.com/artificial-intelligence-act-what-european-approach-ai [https://perma.cc/Y3V4-7AGD].

[15] There is a robust and growing body of legal scholarship discussing different elements of regulating algorithmic decisionmaking, which 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, 18–19 (2014) (arguing for procedural regularity and oversight to ensure fairness and accuracy in artificially intelligent scoring systems); Pasquale, *A Rule of Persons, Not Machines*, *supra* note 8, at 18–20 (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) (arguing that new technologies have the potential to institutionalize social injustice); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018) (examining how automated systems disproportionately affect the poor); Calo & Citron, *supra* note 2, at 804 (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, 235 (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, 246–47 (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 REGUL. 547 (2016) (discussing private automation complaint-resolution mechanisms). Other scholarship evaluates related social and governance considerations. *See* Kaminski, *Binary Governance*, *supra* note 13 (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). But these forward-looking works on algorithmic decisionmaking and AI rarely acknowledge how law already shapes human involvement in a decisionmaking process.

A smaller 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 BERKELEY 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 State*, 105 CORNELL L. REV. 1875, 1908–10 (2020) [hereinafter Huq, *Constitutional Rights in the Machine-Learning State*] (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) [hereinafter Jones, *Ironies of Automation*] (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 siloed by field and focused on a particular topic or technology, such as international humanitarian law and autonomous weapon systems,[16] health or privacy law and black-box medicine,[17] or administrative law and automated benefit disbursement.[18] These siloed approaches are useful for exploring how law might respond to specific issues, but they are inherently limited in their impact and utility.

In contrast, comparing systems across contexts yields other benefits, in that it highlights the many ways that law already governs human-in-the-loop systems.[19] Sometimes, the law requires human decisionmakers.[20] Other times, legal systems may indirectly incentivize keeping or placing a human in the automated decisionmaking process.[21] Or law may discourage or prohibit retaining human influence.[22]

Identifying the existing law of the loop allows us to better see its problems.[23] Namely, it often operates haphazardly and inadvertently—

harder ones—while overseeing fallible automated systems); Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245 (2016) (discussing the role of AI in criminal adjudication).

[16] *See, e.g.*, Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837, 1861 & n.79 (2015) [hereinafter Crootof, *Killer Robots Are Here*] (discussing how autonomous weapon systems might be most effectively regulated).

[17] *See, e.g.*, W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421 (2017) [hereinafter Price, *Regulating Black-Box Medicine*] (suggesting a collaborative approach to govern medical algorithms).

[18] *See, e.g.*, Calo & Citron, *supra* note 2, at 800–02.

[19] 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, for example, *Draft E.U. AI Act*, *supra* note 1; agency documents that are treated as authoritative or quasi-binding by relevant stakeholders, see, for example, 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, for example, Int'l Comm. Red Cross, *International Committee of the Red Cross (ICRC) Position on Autonomous Weapon Systems: ICRC Position and Background Paper*, 915 INT'L REV. RED CROSS 1335, proposing regulations for autonomous weapon systems. We endeavor throughout to be clear about the nature of the "law" we consider.

[20] For example, the Draft E.U. AI Act, *supra* note 1, sometimes explicitly requires human oversight. *See infra* Section II.A.1. Meanwhile, the 1968 Convention on Road Traffic implicitly assumes a (presumably human) driver. *See infra* Section II.A.2.

[21] 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* Section II.B.1.

[22] 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* Section II.C.1.

[23] 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. *See* Rebecca Crootof & BJ Ard, *Structuring Techlaw*, 34 HARV. J.L. & TECH. 347 (2021) [hereinafter Crootof & Ard, *Structuring Techlaw*]; Meg Leta Jones, *Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw*, 2018 U. ILL. J.L. TECH. & POL'Y 249, 256 ("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*

it hasn't been designed to succeed. The law of the loop typically doesn't identify *why* a human is or is not in the loop; clarify what role(s) the human is supposed to play; account for the human's needs, skills, or frailties; or anticipate the ways in which working in tandem with a machine will channel and influence that human's behavior.

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 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, will bear the brunt of liability.[24] Tort rules, though not intended to affect the design or operation of the hybrid system, profoundly shape its dynamics.[25] 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.

Our second significant contribution is to emphasize that human-machine systems are different from—and often more complicated than—the sum of their parts. We identify and describe "the MABA-MABA trap," a common but underdiscussed error that arises in attempts to regulate human-in-the-loop systems. Based on what we know about what "Men Are Better At" and what "Machines Are Better At,"[26] policymakers often assume that adding a human to a machine system will result in the best of both worlds. There's a seductive simplicity to this "slap a human in it" approach.[27] First, the human in the loop is a concrete and identifiable entity, and thus a familiar regulatory target. Second, this tactic is supported by persuasive truths:

First Amendment Law, 51 U.C. DAVIS L. REV. 589, 615 (2017) (contending that technological disruption can lead to changes in legal doctrine and raises a "pacing problem").

[24] Ryan Calo, *Robots in American Law* 36 (U. Wash. Sch. L., Legal Stud. Rsch. Paper No. 04, 2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737598 [https://perma.cc/9P4Q-V5JM] (observing that judges have a tendency to attribute liability to the person "in the loop" over a robotic system).

[25] *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.

[26] Jones, *Ironies of Automation*, *supra* note 15, at 104–06.

[27] *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) [hereinafter Jones, *The Right to a Human in the Loop*] (describing the E.U.'s use "of the human in the loop as a regulatory tool").

at present, for example, machines are better at repetitive tasks, while humans are better at complex contextual analysis. But the MABA-MABA approach is grounded on an assumption that is intuitive, wrong, and dangerous: that human-machine systems represent the best of both worlds and don't introduce new issues of their own.

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? Require that control be transferred to the human when there's unidentifiable debris.[28] In fact, Tesla did just this: in sixteen known instances, the software turned off autopilot less than one second before impact, transferring control over to a human driver.[29]

But simply adding a human is not the fix it appears to be. Instead, it often facilitates (sometimes deadly) accidents and sets the human up for failure and blame—which may be precisely what Tesla intended.[30] Rather than marrying the best of humans and machines, hybrid human-machine systems can exacerbate the worst of each, while adding new sources of error.[31] Humans may have inadequate training or expertise to perform their tasks well, interface issues may lead to information being lost in translation, and handoffs may be bungled—all of which increase the risk of complex failure cascades. However well-meaning regulators might be, careless law has consequences. Further, a MABA-MABA response distracts policymakers from more effective—albeit more complicated and difficult—forms of regulation.[32]

To regulate human-in-the-loop systems well, one must first identify what the human is intended to do. Our third contribution is to detail the panoply of roles that humans in the loop might be expected

[28] *See* Jones, *Ironies of Automation Law*, supra note 15, at 104–06.

[29] The National Highway Traffic Safety Administration ("NHTSA") said it had discovered sixteen separate instances where Autopilot "aborted vehicle control less than one second prior to the first impact," suggesting the driver was not prepared to assume full control over the vehicle. Christiaan Hetzner, *Elon Musk's Regulatory Woes Mount as U.S. Moves Closer to Recalling Tesla's Self-Driving Software*, FORTUNE (June 10, 2022, 4:42 PM), https://fortune.com/2022/06/10/elon-musk-tesla-nhtsa-investigation-traffic-safety-autonomous-fsd-fatal-probe/ [https://perma.cc/UR3U-6JKS].

[30] *Id.* ("CEO Elon Musk has often claimed that accidents cannot be the fault of the company, as data it extracted invariably showed Autopilot was not active in the moment of the collision.").

[31] We are not the first to note the dangers of a simplistic and blindered approach to human-in-the-loop regulation. *See* Roth, *supra* note 15, at 1296–97 (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).

[32] 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.

to play. This typology should be useful to policymakers interested in clarifying the stakes of governance debates, to regulators in articulating their aims, to implementers in understanding the purposes of regulations, and to evaluators in assessing regulations' success.

For our fourth contribution, we provide guidance for policymakers crafting regulations. Namely, rather than simply placing a human in the loop, policymakers should be specific about what humans in the loop are there to do (recognizing that some roles may conflict), take context into account, and learn from human factors engineering and related fields about how best to regulate hybrid systems *as systems*.

Returning one last time to autonomous vehicles: regulators might want humans to be in control when there is unidentified debris in the road, in order to correct errors the vehicle might otherwise make. For the human operator to play that corrective role well, a system must be designed to anticipate that the individual may be distracted or tired. While humans may be better at responding to unexpected circumstances, we're certainly not better at doing so milliseconds before a crash! Accordingly, policymakers could require designers to comply with human factors engineering best practices, which might include building in alerts and sufficient time for a human operator to assess the situation before receiving control of the vehicle.[33]

Here's what's coming: In Part I, we adopt a limited definition of the "human in the loop"[34] that reflects lawmakers' problematically narrow focus on the 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 inappropriately. In Part III, we shift to issues that regulators wishing to craft new rules must consider, beginning with the MABA-MABA trap. After acknowledging the common understandings of what humans are better at and what machines are better at, we discuss the special challenges of regulating hybrid systems. In Part IV, we suggest that if regulators choose to put a human in the loop, they need to articulate why they are doing so. To assist in this endeavor, we offer a nonexhaustive typology of possible human roles, including corrective, resilience, justificatory, dignitary, accountability, stand-in, friction, "warm body," and interface roles. Part V synthesizes lessons

[33] NHTSA took Tesla to task precisely for these sorts of failures. Hetzner, *supra* note 29. In escalating the investigation into Tesla's software towards a recall, NHTSA specifically cited human factors research when noting its concerns that "Autopilot and associated Tesla systems may . . . undermin[e] the effectiveness of the driver's supervision." *Id.*; *see infra* Part V (discussing the use of human factors research in regulatory systems).

[34] *See infra* note 42 (illustrating the experience of being a human in a loop).

from human factors engineering to offer actionable suggestions for crafting regulations going forward, then employs two case studies—the draft E.U. AI Act and law enforcement use of facial recognition technologies—to showcase how our process could improve the law of the loop.

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 appears to be. There is no straightforward, one-size-fits-all solution. Instead, as demonstrated by the regulation of 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"

This Part begins by defining a "human in the loop" in line with the definition regulators often implicitly employ. It then argues that this is the wrong framing for regulating human-machine systems.

A. Introducing the Definition . . .

For the purposes of analyzing existing law, we define a "human in the loop" as an individual who is involved in a single, particular decision made in conjunction with an algorithm.

Human-in-the-loop systems may take a variety of forms. They include ones where an individual human decisionmaker has the discretion to use an algorithmic system to reach a particular decision in a particular instance, such as the doctor who chooses whether or not to use an AI diagnostic tool when treating a patient. Human-in-the-loop systems also encompass ones where an individual and algorithm pass off tasks or perform tasks in concert, such as the pilot who performs some tasks manually while relying on an autopilot for others. Less obviously, these systems include ones where an individual alters an algorithm mid-determination, such as the lawyer who reconfigures the search parameters of an e-discovery tool. They include when an individual determines whether or how to implement an algorithmically informed conclusion, such as the commander who decides against engaging an algorithm's recommended target.[35] Arguably, even a system that enables or requires immediate human review of an

[35] Our definition of a "human in the loop" thus includes more actors than those which focus only on humans engaged in oversight or review. *See, e.g.*, Brennan-Marquez et al., *supra* note 15, at 749; Green, *supra* note 8, at 2 n.1.

automated decision before implementation (that is, individualized contestation) is a human-in-the-loop system.[36]

It might be helpful to contrast these human-in-the-loop systems with ones where humans are "off" the loop.[37] Let's say a company decides to use an AI system or other algorithm to screen all job applicants.[38] That process could be entirely automated. After a candidate submits her resume and uploads a video answering written questions, an algorithm could scan the resume for particular terms and the interview for particular personality characteristics such as "the willingness to work hard and persevere," then reject any candidates who don't meet certain criteria.[39] If no human reviews the process, evaluates the decisions, or provides any other form of oversight, humans are "off" the loop. But if a human has the ability to intervene in an individual decision—to change it, approve it, or immediately implement it—then there is a human "in" the loop.

Our definition is purposively narrow: we don't address many other relevant individuals or the relationship between humans and levels of automation.[40] Instead, we limit our focus to the individual

[36] Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1989 (2021) (describing such systems as allowing for individualized contestation); *see* Huq, *A Right to a Human Decision*, *supra* note 2.

[37] 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. As we define it, however, the "human in the loop" includes folks that might otherwise be considered a "human on the loop."

[38] *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 [https://perma.cc/K8LR-9YHA]; Nathan Mondragon, *Creating AI-Driven Pre-Hire Assessments*, HIREVUE (June 6, 2021), https://www.hirevue.com/blog/hiring/creating-ai-driven-pre-employment-assessments [https://perma.cc/2GAC-5PHD] (heavily emphasizing AI's use as a tool rather than a substitute for people).

[39] Ridiculous, right? Still, see Mondragon, *supra* note 38, discussing the current popularity of such systems.

[40] 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 ("SAE")'s six 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 Automated Vehicles for Safety: The Road to Full Automation,* NAT'L HIGHWAY TRAFFIC SAFETY ADMIN., https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety (last visited Oct. 4, 2022, 8:36 PM) [https://perma.cc/MP79-SNVK]; *Automated Driving Systems*, NAT'L HIGHWAY TRAFFIC SAFETY ADMIN., https://www.nhtsa.gov/vehicle-manufacturers/automated-driving-systems (last visited Feb. 8, 2023) [https://perma.cc/TH9A-6H6Z]. 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-five automated systems will not have humans in the loop, insofar as they can operate without human intervention. Other SAE level-five automated systems may, such as in circumstances where humans face legal incentives to exercise their option to control the car.

involved in a particular decision because lawmakers tend to focus on such individuals.[41] The human in the loop[42] is a concrete and identifiable entity and thus a frequent regulatory target.

Note that nothing in our definition requires 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] In contrast, our definition here encompasses humans who are unable to effectively achieve a regulator's aims. Again, this is purposive: as we survey below, the law of the loop regularly places humans into automated decisionmaking processes without ensuring that these individuals are able to successfully play their roles.[45]

B. . . . but the Definition is Misleading . . .

We've introduced a definition of the "human in the loop" that serves our particular purposes: to identify and analyze the existing law

[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. ANN. 42/1-20 (West 2019) (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. ANN. § 10-3-1104.9(1)(b) (West 2021) (prohibiting insurers from using discriminatory algorithms 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, see *infra* Part V).

[42] *See infra* note 395 (illustrating the experience of being a human in a loop).

[43] Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189 (2019) [hereinafter Kaminski, *Right to Explanation, Explained*] (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); *see also* Crootof, *Killer Robots Are Here*, *supra* note 16, at 1861 & n.79 (arguing that autonomous weapon systems with nominal human involvement should be considered "effectively autonomous weapon systems" rather than semi-autonomous weapon systems).

[44] Brennan-Marquez, Levy, and Susser require the human to have "some degree of meaningful influence" to be considered in the loop, *supra* note 15, 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, *supra* note 8, at 8–11 (contrasting the differentiation of "oversight" policies into those "emphasizing human discretion" versus those "requiring 'meaningful' human input").

[45] *See infra* Part II.

of humans in the loop. At the same time, however, we believe this definition is inherently misleading.

Namely, by focusing only on an individual decision, this framing obscures the fact that humans are *everywhere* in automated decisionmaking systems.[46] There is no such thing as a completely independent machine, let alone a machine decisionmaker. From a systems perspective, a machine cannot exist nor operate without humans. Humans choose and affect the design of such systems. Humans select training data and inputs. Humans ask the question the system answers. Humans implement the conclusion. Humans conduct ex post evaluations of decisionmaking processes. It's humans all the way down. But these and other myriad influential forms of human involvement fade into the unregulated background when we focus overmuch on a discrete application of a system or its particular results.

We are not the first to note this. Meg Jones observes that there is no such thing as a purely algorithmic decision because every system must include a human somewhere, if only to address inevitable failures.[47] Similarly, Andrew Selbst, danah boyd, Sorelle Friedler, Suresh Venkatasubramanian, and Janet Vertesi note that all "technical systems are subsystems" within social and human contexts.[48] Human-in- or human-off-the-loop systems are part of expansive sociotechnical systems which always include humans.[49]

Further, whether a human appears to satisfy the narrow definition of a "human in the loop"—and therefore to be a visible regulatory target—also depends on how regulators define the scope of a system's tasks. Consider the elevator: if the task is described as physically moving the elevator up and down in response to a request for a particular floor, then there is no human in the loop—we no longer have elevator operators.[50] But if the "loop" includes choosing what floor to go to and deciding when to call and command the elevator, the human

[46] *See, e.g.*, KATE CRAWFORD, THE ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE 9 (2021) (discussing how AI is conceptually often disembodied from the systems of extraction and power relations between humans that undergird its use); Hannah Bloch-Wehba, *Algorithmic Governance from the Bottom Up*, 48 BYU L. REV. 69, 119–23 (2022) (discussing how oversight in the design stage includes transparency requirements and external stakeholder involvement); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 655 (2017).

[47] Jones, *Ironies of Automation*, *supra* note 15, at 84; *id.* at 104 (describing this as humans having a "permanent role in the loop").

[48] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian & Janet Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, *in* FAT* '19: PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 59, 59 (2019), https://dl.acm.org/doi/10.1145/3287560.3287598 [https://perma.cc/9YF5-FLLB].

[49] *Id.* at 60.

[50] Jones, *Ironies of Automation*, *supra* note 15, at 108.

user still does most of the information processing and decisionmaking, relegating the elevator to the limited role of implementation.[51] If policymakers wanted to ensure that there was a human in the loop of elevator operating, they would first need to define what constitutes the relevant "loop" of elevator operating. Under one framing, there's no problem: there's a human operator who calls the elevator and tells it where to go. Under the other framing, the elevator is entirely automated and elevator operators who once were in the loop are now out of it.

Were we (and regulators) to employ a broader definition of "human in the loop" or a wider frame for what constitutes a system's tasks, an autonomous vehicle's software would be more easily recognized as part of a system infused with humans: the human engineers who design the software, program its confidence thresholds, and frame its risk tolerances; the human insurance agents who decide the software is safer than human drivers in some circumstances; the human employers in a transport company who decide to mandate that all truck drivers deploy the software; the human inspectors who evaluate crashes and suggest design improvements; the human city officials who make decisions influencing driving infrastructure, such as street direction, signage, bike lanes, and traffic flow; and even the human former truck drivers who band together to form a union or to leave for another company that better respects their expertise and autonomy.[52] As individuals and as embedded in organizations, each of these humans constitutes a human in the loop of a complex sociotechnical system—and a potential target for regulators.

Put plainly, humans are everywhere, whether in the loop, on the loop, off the loop, or hidden from view. When we frame the loop as only a particular decision and the machine as "not human," we ignore other influential humans and human organizations. This has important implications for the scope and aim of regulations.[53]

C. . . . and Has Problematic Regulatory Implications

Regulators often implicitly employ a narrow definition of what constitutes a human in the loop and thereby limit the framing of a system's tasks, with various negative consequences.

[51] *Id.*

[52] For more on truck drivers and automation, see KAREN LEVY, DATA DRIVEN: TRUCKERS, TECHNOLOGY, AND THE NEW WORKPLACE SURVEILLANCE (2022).

[53] As we argue below, rather than focus on the individual in the decisional loop, regulators should regulate human-in-the-loop systems. *See infra* Part V.

First, obscuring the role of the various influential humans limits the apparent available types of regulatory interventions. If regulators only see the human involved in a particular decision as a relevant regulatory subject, it is easy to presume that they should focus their efforts on regulating that individual. But individuals within a system can only have a limited impact on that system's output: they are constrained by their capabilities, resources, and what the system allows them to do.

When debating potential regulations for autonomous weapon systems, for example, the International Committee of the Red Cross noted that any regulatory definition should focus on "autonomy in the critical functions of *selecting and attacking targets*."[54] 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 directly by humans.[55] In doing so, it dramatically circumscribed the relevant loop to target selection and engagement, which in turn circumscribed who would be considered humans in the (lethal) loop. Under this definition and framing, humans involved in the creation, training, deployment, or use of target-recommendation systems would not be considered relevant to the discussion, regardless of the fact that target-recommendation systems can have lethal impacts.

Second, a narrow approach limits the effectiveness and impact of the regulation. The E.U.'s General Data Protection Regulation ("GDPR") regulates certain "automated individual decision-making" by prohibiting "solely automated" decisions with significant effects.[56] Reuben Binns and Michael Veale point out that this enshrines a huge potential loophole.[57] Because Article 22 regulates the loop only when it is linked to a final decision with significant effects, it potentially misses out on (1) automated and often invisible backend systems that significantly inform or constrain later human decisions and (2) earlier, solely automated decisions in a chain of decisionmaking. That is, the relevant loop for the GDPR's Article 22 is only the very last step.[58] As a

[54] *Towards Limits on Autonomy in Weapon Systems*, INT'L COMM. RED CROSS (Apr. 9, 2018), https://www.icrc.org/en/document/towards-limits-autonomous-weapons [https://perma.cc/S2NF-ASCL] (emphasis added).

[55] *Id.*

[56] GDPR, *supra* note 15, art. 22; *see also* Kaminski, *Right to Explanation, Explained*, *supra* note 43; Kaminski, *Binary Governance*, *supra* note 13.

[57] Reuben Binns & Michael Veale, *Is That Your Final Decision? Multi-stage Profiling, Selective Effects, and Article 22 of the GDPR*, 11 INT'L DATA PRIV. L. 319 (2021).

[58] *See also* Mireille Hildebrandt, *The Disconnect Between 'Upstream' Automation and Legal Protection Against Automated Decision Making*, JOTWELL (Apr. 7, 2022), https://cyber.jotwell.com/the-disconnect-between-upstream-automation-and-legal-protection-

consequence, many forms of automated decisionmaking fall outside of the requirements of Article 22 (though, to be clear, not outside of the scope of the GDPR as a whole).

Third, a narrow approach allows policymakers to appear to be doing something at the expense of engaging in more useful regulatory interventions. By focusing too closely on regulating the human in the loop, regulators miss the opportunity to regulate human-machine systems and the organizations that design and deploy them.[59]

There are times when this narrowness is useful—there is utility in focusing in on a particular regulatory subject at the expense of related others—but all too often, narrow definitions and framings are adopted implicitly and unthinkingly, without awareness of their attendant costs.

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 of humans in algorithmic systems.[60] Obviously, law can drive whether humans are in the loop at all by requiring or forbidding their presence. Less obviously, law often creates incentives that encourage or discourage human inclusion.

Collecting these different examples together highlights common problems. Namely, law's influence is frequently incidental and path-dependent[61] rather than grounded on thoughtful evaluations of the human's intended role and how law might facilitate their success.[62] Some laws place a human in the loop carelessly. 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.

against-automated-decision-making/ [https://perma.cc/73RP-VGQG] (reviewing Binns & Veale, *supra* note 57).

[59] *See infra* Part V.

[60] 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.

[61] *See infra* Part IV.

[62] 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.

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 decisionmaking. Ben Green has compiled over forty policies that prescribe human oversight of algorithms employed in governmental decisionmaking.[63] 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.[64]

For example, the newly proposed E.U. AI law ("the draft AI Act") obliges providers of "high risk" AI systems to design and develop them so that "they can be effectively overseen by natural persons during the period in which the AI system is in use."[65] "High risk" AI providers are also required to build systems with "appropriate human-machine interface tools."[66] They must enable humans in the loop to understand the capacities of the system "and be able to duly monitor its operation," which entails correctly interpreting its output, detecting and addressing anomalies, rejecting bad outputs, and stopping its operation when necessary.[67] Providers must also design the system so that humans in the loop remain aware of the human tendency to defer to a

[63] Green, *supra* note 8, at 4.

[64] *Id.* at 9–14.

[65] *Draft E.U. AI Act*, *supra* note 1, art. 14(1) ("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."). "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. *Id.* at annex III. The European Commission can add to this list.

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 (2021). Title II of the draft E.U. AI Act regulates "unacceptable risks"; Title III regulates "high risks"; Title IV regulates limited risks; and Title IX regulates minimal risks. *Draft E.U. AI Act*, *supra* note 1.

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

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

machine, without any guidance on how to do so.[68] The Act's preamble suggests the human in the loop should "have the necessary competence, training and authority to carry out that role"—one of the few times extant law governing human-in-the-loop systems acknowledges this need—however, nothing in the Act itself requires this.[69]

The draft AI Act additionally creates a sui generis humans-in-the-loop requirement for some biometric systems: it mandates sign-off by two humans before biometric identification can be used in certain contexts.[70] This requirement positions the humans as end-of-the-loop gatekeepers to prevent action on the basis of an inaccurate algorithmic identification.

A new Colorado law governing the use of facial recognition by government agencies places a human in the loop when a government agency uses facial recognition "to make decisions that produce legal effects concerning individuals."[71] Those decisions must be "subject to meaningful human review,"[72] defined as "review or oversight by one or more individuals who are trained . . . and who have the authority to alter a decision under review."[73]

In addition to existing law, some explicitly call for new law requiring a human in the loop. For instance, a civil society–led movement is campaigning for regulations that would require human involvement in decisions to employ lethal force in war.[74]

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

[69] *Id.* at recital 48.

[70] *Id.* art. 14(5). 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." *Id.*; *see also* Veale & Zuiderveen Borgesius, *supra* note 65, 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.").

[71] COLO. REV. STAT. § 24-18-303 (2022). *See generally* SB 22-113, 73d Gen. Assemb., Reg. Sess. (2022) (codified at COLO. REV. STAT. §§ 24-18-301 to 309, and elsewhere).

[72] COLO. REV. STAT. § 24-18-303.

[73] *Id.* § 24-18-301(III)(9). The law briefly outlines training requirements for anybody who operates a facial recognition system ("FRS") or processes personal data obtained from such a system, in addition to requiring (without specificity) that relevant individuals be specifically trained for meaningful human review. *Id.* § 24-18-305. In addition to training individuals to conduct a meaningful human review, training requirements apply more broadly to include periodic trainings for all individuals who operate an FRS or process personal data obtained from an FRS, and it must include, at minimum, coverage of the capabilities and limitations of the service, and procedures to interpret and act on the output of the service. *Id.*

[74] *See, e.g.*, Int'l Comm. of the Red Cross, *supra* note 19, at 8–9.

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.[75] If applied to autonomous vehicles, the provision could be interpreted to require a human driver in the loop. Similarly, medical prescriptions may only be written by a specified set of professionals.[76] 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. Any number of laws might similarly require 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 practitioners and scholars intentionally read existing rules to contain an implicit human-in-the-loop requirement. Consider the debate around regulating autonomous weapons systems: in contrast to those who argue for creating new rules, some advocates for a ban claim that this technology is already prohibited under (their particular interpretations of) existing law.[77]

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

[75] *See* U.N. Convention on Road Traffic art. 8, Sept. 19, 1949, 125 U.N.T.S. 22; U.N. 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).

[76] *See, e.g.*, *Who Can Prescribe and Administer Prescriptions in Washington State*, WASH. STATE DEP'T HEALTH, https://www.doh.wa.gov/LicensesPermitsandCertificates/ProfessionsNewReneworUpdate/PharmacyCommission/WhoCanPrescribeandAdministerPrescriptions#Prescribe (last visited Sept. 21, 2022) [https://perma.cc/M6RT-8K29].

[77] *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," but also calling for new regulations); 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 [https://perma.cc/97T2-J6PR] (arguing that fully autonomous weapon systems could violate the right to life without being able to be held accountable); Jeffrey Kahn, *"Protection and Empire": The Martens Clause, State Sovereignty, and Individual Rights*, 56 VA. J. INT'L L. 1 (2016) (arguing that the Martens Clause to the 1899 Hague Convention Respecting the Laws and Customs of War on Land prohibits the adoption and use of autonomous weapon systems).

arbitrage fosters human involvement when keeping a human in the loop avoids costly regulations. Second, liability rules may encourage the inclusion of a human 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 as well as to other technologies and practices.[78]

1. Regulatory Arbitrage

Some laws encourage regulatory arbitrage when retaining a human in the loop allows a system's developers or users to avoid more onerous regulation. This often manifests 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 jurisdiction.[79] Software in certain categories is defined as a "medical device," such that marketing it requires going through FDA 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 processes,[80] 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"[81] by a human "health care professional" who is able to review its output before use, then the software is likely

[78] *Cf.* Crootof & Ard, *Structuring Techlaw*, *supra* note 23 (discussing the respective benefits of tech-neutral and tech-specific rules).

[79] Kind of. *See* Barbara J. Evans & Frank A. Pasquale, *Product Liability Suits for FDA-Regulated AI/ML Software*, *in* THE FUTURE OF MEDICAL DEVICE REGULATION: INNOVATION AND PROTECTION 22, 23 (I. Glenn Cohen, Timo Minssen, W. Nicholson Price II, Christopher Robertson & Carmel Shachar eds., 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.

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

[81] FDA, DOCKET NO. FDA-2017-D-6569, CLINICAL DECISION SUPPORT SOFTWARE: GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF 4 (2022).

not a medical device subject to FDA approval.[82] This 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 [clinical decision support software] and was not a diagnostic medical device."[83]

Similarly, the E.U. GDPR imposes additional obligations on the use of "solely automated" decisionmaking with significant effects.[84] These potentially costly requirements[85] do not apply to automated

[82] 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, § 3060, 130 Stat. 1033, 1130–31 (2016) (codified at 21 U.S.C. § 360j(o)) (emphasis added). *See* W. Nicholson Price II, Rachel E. Sachs & Rebecca S. Eisenberg, *New Innovation Models in Medical AI*, 99 WASH. U. L. REV. 1121, 1146 (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 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.

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

[84] GDPR, *supra* note 41, at 46 (art. 22).

[85] *See id.*; Art. 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*, at 20, 32, WP 251rev.01 (Feb. 6, 2018), https://ec.europa.eu/newsroom/article29/items/612053 [https://perma.cc/N366-3MQK] [hereinafter GDPR Guidelines] (suggesting impact assessments are required for some systems and audits are best practices); *see also* Kaminski, *Binary Governance*, *supra* note 13, at 1595. The GDPR's Article 2 also contains a right to "human intervention" in an algorithmic decision with significant effects. *See* Huq, *A Right to a Human Decision, supra* note 2, at 622–23:

> 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. *See* Kaminski & Urban, *supra* note 36, at 1989 ("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 PROT. L. REV. 433 (2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3305272 [https://perma.cc/G5U5-QRRU] (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.").

decisions that are made with significant human involvement,[86] and some argue they do not apply to automated decisions made with even nominal human involvement.[87] Thus, companies are 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 harms when things go wrong (or even when things go right).[88] 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.[89] Tort, criminal, or administrative liability may follow.

So long as liability rules 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

[86] Decisions where the human is a rubber stamp, however, are likely covered. GDPR Guidelines, *supra* note 85, 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.").

[87] 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. *See* Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 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." *See* Kaminski, *The Right to Explanation, Explained*, *supra* note 43, at 199–201; *see also* GDPR Guidelines, *supra* note 85, 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.

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

[89] *See, e.g.*, Rebecca J. Hamilton, *Platform-Enabled Crimes: Pluralizing Accountability When Social Media Companies Enable Perpetrators to Commit Atrocities*, 63 B.C. L. REV. 1349, 1951–52 (2022) (noting how Facebook's "algorithmically-curated newsfeed . . . pushed incitement against the Rohingya [in Myanmar] to a new level.").

affecting the actual decision.[90] 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";[91] they could as easily be referred to as the legal crumple zone. Take a diagnostic algorithm. Today, medical practitioners have the 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 legal landscape is still uncertain). The possibility of offloading liability encourages developers to avoid creating algorithms that provide their own autonomous diagnosis.[92]

Notably, this incentive structure carries no 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. Including a human undermines such claims.

Successful legal challenges to U.S. government use of algorithms have largely been procedural in nature.[93] There is a growing body of case law where algorithmic decisions were invalidated on procedural due process grounds.[94] For example, states—prompted by federal government monetary incentives—have adopted various Value Added

[90] Calo, *supra* note 24, at 36.

[91] Madeleine Clare Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAGING SCI., TECH. & SOC'Y 40, 42 (2019) [hereinafter Elish, *Moral Crumple Zones*].

[92] 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. Michael D. Abràmoff, Danny Tobey & Danton S. Char, *Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process*, 214 AM. J. OPTHALMOLOGY 134, 139 (2020).

[93] Huq, *Constitutional Rights in the Machine-Learning State*, *supra* note 15, at 1880:

> 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.

(footnote omitted).

[94] *See, e.g.*, Barry v. Lyon, 834 F.3d 706, 718–20 (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).

Model algorithmic systems for evaluating public school teachers.[95] These systems' recommendations affected teacher merit pay, the award or revocation of teacher tenure, and even teacher terminations.[96] In a Texas lawsuit, teachers successfully claimed that the lack of transparency regarding how the algorithm reached its conclusions constituted a due process violation.[97]

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 in any U.S. governmental 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, such as a fiduciary duty; and via liability rules which require meeting standards of care that can only be provided by algorithms.

1. Silence

In the face of background incentives, legal silence can discourage the presence of humans in many algorithmic systems. Efficiency gains are their own motivation in most contexts; performing better, or performing at lower cost and using fewer resources, are typically good results from the system designer's perspective.[98] Certainly, 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. But whenever there is some benefit to be gained from algorithmic efficiency or speed, law's silence will foster eliminating human influence.

[95] *See* 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 [https://perma.cc/M87Z-CW2D].

[96] *Id.*

[97] Houston Fed'n of Tchrs., Loc. 2415 v. Houston Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1179 (S.D. Tex. 2017) ("[W]ithout access to SAS's proprietary information—the value-added equations, computer source codes, decision rules, and assumptions—EVAAS scores will remain a mysterious 'black box,' impervious to challenge.").

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

If there is a benefit to making decisions with superhuman speed, efficiency pressures discourage involving humans. Take stock trading: algorithms complete the vast majority of trades, making decisions which take relentless advantage of arbitrage opportunities in millionths of a second—speeds no human could even imaginably reach.[99] For those with the capacity to use high-frequency trading algorithms, putting humans in the loop of quotidian 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.[100] In general, platforms cannot be held liable for taking down legal content.[101] This liability shield, found in Section 230 of the Communications Act,[102] gives platforms enormous leeway to filter or not filter content as they deem best. Between January 2020 and March 2020, for example, YouTube removed over 6.1 million videos allegedly in violation of its Community Guidelines,[103] which include categories such as spam, child safety, nudity or sexual, and violent or graphic.[104] Of these, 5,711,586 videos were removed through automated flagging.[105] Using humans to take down videos would be prohibitively expensive, making it impossible for YouTube or any other significant platform to create a curated online environment. Thus, in the absence of a legal requirement, platforms are disincentivized from including humans in the content moderation loop.

[99] Merritt B. Fox, Lawrence R. Glosten & Gabriel V. Rauterberg, *High-Frequency Trading and the New Stock Market: Sense and Nonsense*, 29 J. APPLIED CORP. FIN. 30, 30–32, 38 (2017). 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*, WEEK (Jan. 11, 2015), https://theweek.com/articles/493238/wall-streets-secret-advantage-highspeed-trading [https://perma.cc/Y5V8-M2ZL].

[100] L. Cici, *Hours of Video Uploaded to YouTube Every Minute as of February 2020*, STATISTA (Apr. 4, 2022), https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/ [https://perma.cc/R2WG-PW3Y].

[101] *See* 47 U.S.C. § 230.

[102] For more on why this isn't actually section 230 of the Communications Decency Act, as popularly known, see Blake E. Reid, *Section 230 of . . . What?*, BLAKE E. REID (Sept. 4, 2020), https://blakereid.org/section-230-of-what/ [https://perma.cc/QA8X-JHSY].

[103] *YouTube Community Guidelines Enforcement,* GOOGLE, https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2020Q1;exclude_automated:all&lu=total_removed_videos (last visited Oct. 9, 2022) [https://perma.cc/CJ4D-QL2Y] (Removed Videos by the Numbers).

[104] *Id.* (Removed Videos by Removal Reason).

[105] *Id.* (Removed Videos by the Numbers).

By remaining silent, law permits other forces to push humans out of decisionmaking systems.[106] But that is a policy choice; lawmakers needn't take that tack. In the face of the huge amount of effort expended in enabling high-frequency trading,[107] one could plausibly imagine a world where the U.S. Securities and Exchange Commission required that a human sign off on every trade. If a society prioritized quality content over speed and quantity, social media companies could be required to engage in more particularized review of approval decisions. Absent such requirements, in any context where algorithms are thought to do things more cheaply, more efficiently, or more precisely than a human, there will be pressure to keep humans out of the loop.

2. Fiduciary Duties

Fiduciary duties to maximize returns—such as a CEO's duty to corporate shareholders to manage corporate assets or a stock broker's duty to customers of best execution—may strengthen the efficiency effect. Fiduciary duties may add legal heft to existing incentives by (at least nominally) requiring fiduciaries to pursue better performance outcomes 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[108]—but the soft law implications matter.

Fiduciary duties strengthen commercial norms that promote better performance over values such as dignity or fairness.[109] They can

[106] *Cf.* Joel R. Reidenberg, *Lex Informatica*: *The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553 (1998) (discussing how different regulatory modalities influence each other); LAWRENCE LESSIG, CODE: VERSION 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 able to turn back time and reset the market; today, financial markets include various "tripwires" that stop trading when algorithms start acting unusually. *See* Bob Pisani, *"Flash Crash" 5 Years Later: What Have We Learned?*, CNBC (May 5, 2015, 2:07 PM), https://www.cnbc.com/2015/05/05/flash-crash-5-years-later-what-have-we-learned.html [https://perma.cc/JUA4-UN7E]. 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.

[107] *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").

[108] *See, e.g.*, City of Warren Gen. Emps.' Ret. Sys. v. Roche, No. CV 2019-0740-PAF, 2020 WL 7023896, at *20 (Del. Ch. Nov. 30, 2020) ("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." (quoting Morrison v. Berry, No. CV 12808-VCG, 2019 WL 7369431, at *25 (Del. Ch. Dec. 31, 2019))).

[109] *See* Quadrant Structured Prods. 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

also serve as a buffer to public critique for seemingly coldhearted decisions. While Martin Shkreli was lambasted for trotting out fiduciary duties to defend his price gouging on lifesaving medications,[110] 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.[111]

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, regulated entities will automate decisionmaking processes. This manifests in multiple contexts. The National Security Agency reportedly minimized human oversight of surveilled material, reasoning that if no human was involved, reviewing and classifying gathered data wouldn't constitute a "search" implicating the Fourth Amendment.[112] Scholars have argued that automated computer analysis of personal data online doesn't violate privacy laws in the way human involvement would.[113] YouTube purportedly automates the flagging and takedown of copyrighted material in part because involving a human would trigger a higher expectation of a complex (and costly) fair use analysis under current copyright law.[114] In each case, keeping humans out of the loop helps keep regulators out too.

informed basis to maximize the value of the corporation for the benefit of its residual claimants, the ultimate beneficiaries of the firm's value."

[110] Dan Diamond, *Martin Shkreli Admits He Messed Up: He Should've Raised Prices Even Higher*, FORBES (Dec. 3, 2015, 12:55 PM), https://www.forbes.com/sites/dandiamond/2015/12/03/what-martin-shkreli-says-now-i-shouldve-raised-prices-higher/?sh=55536d471362 [https://perma.cc/CS9A-7NAY].

[111] *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 [https://perma.cc/6LC6-NJMP] (connecting AI replacement of workers with shareholder incentives).

[112] Richard Posner, *Our Domestic Intelligence Crisis*, WASH. POST. (Dec. 21, 2005), https://www.washingtonpost.com/archive/opinions/2005/12/21/our-domestic-intelligence-crisis/a2b4234d-ba78-4ba1-a350-90e7fbb4e5bb/ [https://perma.cc/4R65-YSSB] ("[M]achine collection and processing of data cannot, as such, invade privacy.").

[113] *See* Bruce E. Boyden, *Can a Computer Intercept Your Email?*, 34 CARDOZO L. REV. 669, 717 (2012) ("[A]utomated processing does not by itself pose any threat to privacy. . . . The [Wiretap] Act has always required at least the prospect of human review . . . ."); 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 . . . .").

[114] Katharine Trendacosta, *Unfiltered: How YouTube's Content ID Discourages Fair Use and Dictates What We See Online*, ELEC. FRONTIER FOUND. (Dec. 10, 2020), https://www.eff.org/wp/

4. Liability Rules

In addition to encouraging human inclusion,[115] liability rules can also function to discourage including a human in the loop, especially when liability is tied to performance standards that are best achieved without human error or slowness.

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 medical algorithmic systems.[116] They are concerned that once these systems reach a high enough level of performance, their use—and physician deference to their recommendations—will become the standard of care. If so, human physicians who deviate from these standards would be liable for consequent injury.[117] To be clear, this argument doesn't posit that humans will be forced out of 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 their deference increases and their skills atrophy. While this potential outcome appears to be some way off,[118] it highlights how background liability rules could gradually operate to affirmatively push humans out of the loop.

D. Prohibiting

If a rulemaker determines that machine decisionmaking will always be preferable to human decisionmaking in a particular context, they may use law to explicitly prohibit human involvement. Likely due to the relative newness of algorithmic decisionmakers and familiarity with human ones, to the best of our knowledge there are as yet no bans on including a human in the loop, though at least one scholar has

unfiltered-how-youtubes-content-id-discourages-fair-use-and-dictates-what-we-see-online [https://perma.cc/Y92E-3BA5].

[115] *See supra* Section II.B.2.

[116] 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, 72–73 (2019).

[117] *Id.* at 61–63.

[118] 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 JAMA 1765 (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) [hereinafter Price, Gerke & Cohen, *Potential Liability for Physicians Using Medical AI*] (finding that physicians who deviate from nonstandard of care AI recommendations are not yet viewed as liable for resulting injuries).

proposed them.[119] Given interest in curtailing the errors and discrepancies associated with human discretion, however, it is easy to imagine them being enacted in the near future.

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.[120] (Over time, however, case law reintroduced significant judicial discretion, including for deviations from the guidelines.) Similarly, workers' compensation tables assign predetermined amounts for injuries that occur in the course of employment without allowing for individualized tweaks.[121] It is not a big leap to imagine legal requirements that human judges use algorithms in sentencing or damage awards.[122]

* * *

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, as those seeking to create new rules must consider how they might build upon or be undermined by these broader background legal regimes.

Having highlighted the existence of a complex, multilayered law of the loop, the remainder of this Article shifts focus to contemporary and forward-looking regulation—that is, how best to craft new law for human-in-the-loop systems.[123]

[119] *See* Orly Lobel, The Law of AI for Good 41 (Sept. 26, 2022) (unpublished manuscript) (on file with authors) (arguing that "under certain circumstances . . . there should be a prohibition on humans entering the loop when such entrance would diminish the benefits of automation and risk error and bias").

[120] *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.

[121] *See, e.g.*, TENN. CODE ANN. § 50-6-207 (2022) (providing the compensation schedule under Tennessee's Worker Compensation Law).

[122] We offer these hypotheticals as merely that; we certainly would not endorse such requirements!

[123] 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.

III. WHAT REGULATORS GET WRONG: THE "MABA-MABA TRAP"

A central normative question in the discourse is whether human decisionmakers should ever or always be replaced by machines.[124] But this question—while academically, politically, and morally fascinating—is deeply misleading for regulators. By focusing on who (or what) is making a decision, the question obscures the fact that human-in-the-loop systems are distinct entities, capable of being regulated as such.

Policymakers appear to think of human-machine systems as the sum of their parts. But hybrid systems are distinct entities. To help explain why, in this Part we introduce legally minded audiences to what we call 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 what "Machines Are Better At."[125] The original 1951 "Fitts list"—one of the first articulations of humans' and machines' respective skills—identified, for example, that machines are better at performing repetitive and routine tasks, while humans are better at improvising.[126]

A MABA-MABA regulatory approach 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. The first two Sections of this Part provide a summary of commonly observed strengths and weaknesses of both human and machine decisionmaking.[127] (Readers versed in these concepts should feel free to skip ahead.) Thanks to MABA-MABA, humans may be—understandably!—placed in the loop,

[124] *See supra* note 8 and accompanying text.

[125] Jones, *Ironies of Automation*, *supra* note 15, at 105.

[126] 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 10–11 (1951), https://apps.dtic.mil/sti/pdfs/ADB815893.pdf [https://perma.cc/W3UD-MGPW]. 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., MIL-HDBK-763, HUMAN ENGINEERING PROCEDURES GUIDE 93 (1987) [hereinafter DOD GUIDE]; Jones, *Ironies of Automation*, *supra* note 15, at 105 fig.2 (explaining that 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, *Why the Fitts List Has Persisted Throughout the History of Function Allocation*, 16 COGNITION, TECH. & WORK 1 (2014).

[127] For a fairly exhaustive and sometimes similar categorization of the pros and cons of humans versus machines in the legal context, see Cary Coglianese & Alicia Lai, *Algorithm vs. Algorithm*, 71 DUKE L.J. 1281, 1309 (2022): "Digital algorithms are able to perform a variety of tasks better than humans can. . . . This is not to deny that humans will remain better at some tasks than will digital algorithms."

whether by designers or policymakers, based on assumptions about their respective capabilities.

But merely inserting a human in the loop does not necessarily result in the best of both human and machine. Instead, adding or maintaining a human in the loop of an automated system creates a new entity: a hybrid system. And hybrid systems create or exacerbate well-known problems—problems policymakers currently remain unaware of or largely ignore.[128]

Getting humans to work well with algorithmic systems is far more difficult than it may first appear, as evidenced by the fact that entire subfields of engineering and computer science focus on these issues.[129] MABA-MABA's attractive simplicity is thus a trap for regulators. In the final Section of this Part, we introduce the special challenges of hybrid systems—complexities which are familiar in engineering and computer science but, with a few notable exceptions, are rarely recognized by legal scholars or policymakers.[130]

A. Human Decisionmaking

Regulators presumably put humans in the loop because they think they will do something there. What, precisely, are their assumptions about human decisionmaking and the ways in which it differs from machines? A vast and expanding literature explores precisely how humans make decisions;[131] here, we simply highlight a few commonly recognized 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. Some believe that only humans are capable of moral judgment; others who more fully embrace a computational theory

[128] Roth, *supra* note 15, at 1296–98 (criticizing MABA-MABA and calling instead for looking to systems engineering for ideas on how to better design human-machine systems).

[129] *See, e.g.*, DAVID D. WOODS, SIDNEY DEKKER, RICHARD COOK, LEILA JOHANNESEN & NADINE SARTER, BEHIND HUMAN ERROR (2d ed. 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).

[130] *See infra* Part V (discussing the regulation of railroads, nuclear reactors, and medical devices).

[131] *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); STANFORD ENCYCLOPEDIA OF PHIL., DECISION THEORY (2020); Leigh Buchan & Andrew O'Connell, *A Brief History of Decision Making,* HARV. BUS. REV., Jan. 2006, https://hbr.org/2006/01/a-brief-history-of-decision-making [https://perma.cc/EQ6D-TZ38].

of mind might not. Relatedly, human experts have "tacit knowledge"—knowledge that can't always readily be translated into code.[132]

Humans are flexible decisionmakers who can choose to deviate from strict rules and exercise discretion.[133] This flexibility allows us to make contextual decisions, including when a rule must be bent, when to incorporate factors a machine might not have or might categorize as out of scope, and when to make an analysis at a different level of generality to achieve a preferable result.[134]

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 shifts from a criminal to an antitrust case. This ability to reason across tasks and settings 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.

Finally, while our internal decisionmaking processes may be opaque, humans can be interrogated and give reasons for their decisions—though the extent to which such reasons may be post hoc rationalizations is hotly debated.[135]

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.[136] Some of these inconsistencies are due to the fact

[132] PASQUALE, NEW LAWS OF ROBOTICS, *supra* note 8, at 24 ("We know more than we can explain." (citing philosopher Hubert L. Dreyfus's theories of tacit knowledge)).

[133] Of course, the amount of discretion a human decisionmaker can exercise depends on the organizational structure, available resources and other capacity enhancers, and relative power.

[134] Kaminski, *Binary Governance*, *supra* note 13, 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).

[135] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL 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."); *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 . . . ." (emphasis omitted)).

[136] *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. 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).

that we are biased—we are subject both to human decisionmaking biases, like saliency or recency,[137] as well as 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. We get tired. We get bored. We get hungry. We get injured, and we fall ill. We (desperately) need vacations and mental health breaks.

B. Algorithmic Decisionmaking

Algorithms[138] are capable of incredible feats. They 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).[139] Algorithms can store and process vast amounts of information—which, among other things, can be edited or deleted in a way human memory cannot.[140]

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

[137] *See* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES 11 (Daniel Kahneman, Paul Slovic & Amos Tversky eds., 1982) (noting that, in addition to "familiarity," "salience" may also affect the "retrievability of instances").

[138] Not all algorithms are the same. *See* Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305, 1310 (2019) ("[M]ost successful artificial technological approaches fall into two broad categories: (1) machine learning and (2) logical rules and knowledge representation."). Some models reflect a programmer's attempt to model the real world by coding particular rules. *Id.* at 1316. Other models use data analytics to detect and apply patterns across large data sets. *Id.* at 1311. Still others represent hybrids of these approaches. *Id.* at 1319. Our point is that what we say here of algorithms may be true of some and not of others, depending on the kind of system at issue.

[139] *E.g.*, Shannon E. French & Lisa N. Lindsay, *Artificial Intelligence in Military Decision-Making: Avoiding Ethical and Strategic Perils with an Option-Generator Model*, *in* EMERGING MILITARY TECHNOLOGIES: ETHICAL AND LEGAL PERSPECTIVES 53, 54 (2022).

[140] DOD GUIDE, *supra* note 126, at 93; *see also* Jones, *Ironies of Automation*, *supra* note 15, at 105.

variation.[141] This has led some to argue that algorithms discriminate less than human decisionmakers.[142]

Algorithms scale. They do not get bored making the same decision over and over and over again,[143] and they are replicable. It is cheaper and faster to copy an oncology algorithm a thousand times than to train a thousand new oncologists.[144]

After dwelling for any amount of time on human frailties, it's easy for algorithms' relative strengths to dazzle. Why, then, do regulators put humans in the loop? Because algorithmic decisionmaking has its weaknesses too, and regulators likely believe they can create a decisional system comprised of the best of each.

Currently, algorithms are brittle: they may perform well in tasks and situations that are similar to those for which 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 context.[145]

Algorithms can be deeply weird or surprising.[146] An algorithm trained to pilot planes in a flight simulator learned to crash the plane immediately to achieve an error-free (and therefore perfect) score.[147]

[141] 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 JAMA 305, 305 (2020). Nevertheless, once a developed or trained model has been implemented, the same inputs should consistently yield the same outputs.

[142] *E.g.*, Bornstein, *supra* note 8; Miller, *supra* note 8. *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").

[143] DOD GUIDE, *supra* note 126, at 93 tbl.VII (observing that machines excel at "[p]erforming routine, repetitive, or very precise operations").

[144] *See* Kaminski & Urban, *supra* note 36, at 1968–69 (explaining efficiencies of algorithms and AI); Ajunwa, *supra* note 142, at 1734 ("[T]he fact remains that automated hiring is a cost-saving measure.").

[145] Surden, *supra* note 138, at 1332.

[146] *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); Janelle Shane, AI WEIRDNESS, www.aiweirdness.com (last visited Sept. 28, 2022) [https://perma.cc/YV47-N9MD] (collecting ongoing stories in a blog); *see also* WOODS ET AL., *supra* note 129, at 216–19 (noting that accidents are opportunities for learning and change—but only if they are acknowledged as unexpected incidents).

[147] Janelle Shane, *When Algorithms Surprise Us*, AI WEIRDNESS (April 13, 2018), https://www.aiweirdness.com/when-algorithms-surprise-us-18-04-13/ [https://perma.cc/5Q56-UXHN]:

> In one of the more chilling examples, there was an algorithm that was supposed to figure out how to apply a minimum force to a plane landing on an aircraft carrier. Instead, it discovered that if it applied a *huge* force, it would overflow the program's

And a large-language AI system that generated a convincing conversation that it was sentient also generated convincing conversations that it was secretly a squirrel, a Tyrannosaurus Rex, and a self-aware Magic 8 ball.[148] Algorithms don't "think" like humans do (in fact, they don't "think" at all).[149] Algorithms rely on proxies—for both inputs (what does a "good" employee do?) and outputs (what constitutes a "good" employee?).[150] Proxies, whether chosen by human programmers or derived from the data, can be incorrect, value laden, normatively undesirable, and even illegal.[151] 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.[152]

Artificial intelligence is especially dependent on both 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.[153] 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 information and noise.[154] 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.[155]

Further, any code-based system will be riddled with bugs—inevitable programming errors that cause unexpected and sometimes

memory and would register instead as a very *small* force. The pilot would die but, hey, perfect score.

[148] Janelle Shane, *Interview with a Squirrel*, AI WEIRDNESS (June 16, 2022), https://www.aiweirdness.com/interview-with-a-squirrel/ [https://perma.cc/LBH8-7GXQ].

[149] Surden, *supra* note 138, at 1308 ("The reality is that today's AI systems are decidedly not intelligent thinking machines in any meaningful sense.").

[150] *See* CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 20–21 (2016) (explaining that algorithms use proxies to draw statistical correlations between different behaviors); Surden, *supra* note 138, at 1337 (arguing that the use of proxies allows AI to produce intelligent results without intelligence).

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

[152] Ajunwa, *supra* note 142, at 1686.

[153] *E.g.*, Manoush Zomorodi, *Joy Buolamwini: How Do Biased Algorithms Damage Marginalized Communities?*, NPR (Oct. 30, 2020), https://www.npr.org/transcripts/929204946 [https://perma.cc/64SH-GH4R]; Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 8.

[154] *Overfitting*, IBM (Mar. 3, 2021), https://www.ibm.com/cloud/learn/overfitting [https://perma.cc/48H9-57A2].

[155] Sasha Harrison, *How to Tame the Long Tail in Machine Learning*, SCALE (June 29, 2021), https://scale.com/blog/taming-long-tail [https://perma.cc/7QSK-TWHV].

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.[156] Algorithmic systems also introduce new vulnerabilities, insofar as they can be poisoned,[157] hacked, gamed, or otherwise exploited.[158]

Algorithms don't do norms or ethics well. Any concept that is contested or hard to articulate will be hard to translate into code[159]—even apparently "easy" rules like speed limits are subject to a host of coding decisions.[160] Relatedly, algorithms lack the social conditioning and tacit knowledge that humans have and thus miss crucial unarticulated (even inarticulable) aspects of human decisionmaking.[161] Meanwhile, algorithms that try to "learn" ethics or social norms from human behavior can import the nastier elements along with the good.[162]

Algorithms can be "black boxes" in ways that pose challenges for our current legal system. This may be due to legal protections (like trade secrets law), deliberate secrecy (such as confidential training datasets), or as an inherent function of their design by reason of

[156] *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.

[157] 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 [https://perma.cc/YJJ4-K9WD] ("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.").

[158] 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 [https://perma.cc/M6YV-BBQF] (discussing data poisoning attacks, adversarial examples, and the exploitation of goals).

[159] Surden, *supra* note 138, at 1322–23 ("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.").

[160] *See* Shay et al., *supra* note 15 (discussing the many choices the programmers in the experiment faced when deciding how to code a speeding violation).

[161] PASQUALE, NEW LAWS OF ROBOTICS, *supra* note 8; Surden, *supra* note 138, at 1325 ("[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.").

[162] 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, 10:58 AM), https://www.vice.com/en/article/v7dg8m/ethical-ai-trained-on-reddit-posts-said-genocide-is-okay-if-it-makes-people-happy [https://perma.cc/3BBE-3R8Y].

structure (as occurs with neural nets) or complexity (which arises when algorithms are used to crunch more data than humans can monitor).[163]

Whatever one's thoughts about the opacity of the human mind,[164] society has developed ways of querying decisionmakers and identifying reasoning errors. But algorithmic failures can be difficult to identify and assess after the fact, such that some have proposed building technological tools for establishing accountability.[165] Relatedly, algorithms often obscure human intent and responsibility,[166] making humans harder to interrogate even when they are involved.

C. The Trap

Policymakers often place humans in the loop based on assumptions about the respective strengths and weaknesses of humans and machines; on the face of things, 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's weaknesses.[167]

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

[163] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Price, *Regulating Black-Box Medicine*, *supra* note 17.

[164] *See, e.g.*, Huq, *A Right to a Human Decision*, *supra* note 2, at 640–46 (discussing the opacity of human and machine minds).

[165] *See* Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 695–704 (2017) (advocating for employing computational techniques that would increase algorithmic accountability).

[166] Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 8, at 692–93; Ajunwa, *supra* note 142, at 1692–1707.

[167] Roth, *supra* note 15, at 1296–97:

> Researchers have written in the systems engineering context about the dangers of so-called "MABA-MABA" lists. Instead, man-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?"

alone."[168] AI assistants help us navigate the internet;[169] improve our personal shopping experience;[170] and offer medical advice to patients.[171] Human-machine "centaur" chess teams for at least some time performed 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 evaluations or better concentrate on the elements of a decision that require distinctly human judgment in various contexts.[172]

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 impair 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,[173] including rubber-stamping humans that don't really oversee decisions,[174] the prevalence of "automation bias" that leads humans to

[168] PASQUALE, NEW LAWS OF ROBOTICS, *supra* note 8, 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?' ").

[169] *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 Sept. 27, 2022) [https://perma.cc/U57Z-4ZGM].

[170] 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).

[171] Claudia E. Haupt, *Artificial Professional Advice,* 18 YALE J. HEALTH POL'Y L. & ETHICS 55, 67–70 (2019).

[172] *Cf.* Paul Scharre, *Centaur Warfighting: The False Choice of Humans vs. Automation*, 30 TEMP. INT'L & COMP. L.J. 151, 154–56 (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, 9:55 PM), https://www.thedrive.com/the-war-zone/39899/darpa-now-has-ai-controlled-f-16s-working-as-a-team-in-virtual-dogfights [https://perma.cc/Z32V-VVSL] (discussing the benefits of AI-human teams).

[173] Green, *supra* note 8, at 14–18; *see also* Marina Chugunova & Daniela Sele, *We and It: An Interdisciplinary Review of the Experimental Evidence on How Humans Interact with Machines*, 99 J. BEHAV. & EXPERIMENTAL ECON. 1, 2–3 (2022) (reviewing human-computer interactions); Christoph Engel & Nina Grgić-Hlača, *Machine Advice with a Warning About Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court*, 13 J. LEGAL ANALYSIS 284, 286 (2021) (experimentally evaluating the effects of algorithmic accuracy warnings and finding limited effects).

[174] Michael Veale & Lilian Edwards, *Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling*, 34 COMPUT. L. & SEC. REV. 398, 400 (2018).

defer overmuch to machines,[175] the automation-associated deterioration of human abilities known as "skill fade,"[176] incorporating or deviating from algorithmic advice in biased ways,[177] and the basic tautological challenge of relying on humans to monitor the performance of systems designed to improve on human performance.[178] 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 are many times and places for humans in the loop. But regulators don't often address known problems, nor engage with known principles of hybrid human-machine 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.

Hybrid system failures are not hypothetical: there is a long history of complex systems gone awry.[179] 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.[180] Focusing on human-in-the-loop error can obscure bigger system design problems, such as how the system's interface fosters confusion and how discrete errors can cascade.[181]

The story of Three Mile Island, while somewhat dated, is still a chilling example of the kinds of systemic problems that can arise in human-machine systems. In 1979, the Three Mile Island Unit 2 nuclear

[175] Raja Parasuraman & Dietrich H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381, 390–98 (2010); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1271–72 (2008).

[176] Jones, *Ironies of Automation*, *supra* note 15, 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 Uses of Live Facial Recognition in the United Kingdom*, *in* REGULATING BIOMETRICS: GLOBAL APPROACHES AND URGENT QUESTIONS 78, 78–85 (Amba Kak ed., Sept. 2020), https://ainowinstitute.org/regulatingbiometrics.pdf [https://perma.cc/R2B5-5LJH].

[177] Megan T. Stevenson & Jennifer L. Doleac, *Algorithmic Risk Assessment in the Hands of Humans*, IZA INST. OF LAB. ECON. (Dec. 2019), https://docs.iza.org/dp12853.pdf [https://perma.cc/Z4XU-YHWF]; Ben Green & Yiling Chen, *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments*, *in* PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 90, 90–99 (Jan. 2019), https://dl.acm.org/doi/pdf/10.1145/3351095.3372869 [https://perma.cc/L34P-BXZL].

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

[179] WOODS ET AL., *supra* note 129, at 1–2 (listing examples of failures of complex systems).

[180] Jones, *Ironies of Automation*, *supra* note 15, at 112.

[181] Waldemar Karwowski, *The Discipline of Human Factors and Ergonomics*, *in* HANDBOOK OF HUMAN FACTORS AND ERGONOMICS 3 (Gavriel Salvendy ed., 2012); WOODS ET AL., *supra* note 129, at 3.

reactor partially melted down.[182] At first glance, the meltdown might look like human error: the human staff mistakenly took actions that uncovered the reactor core,[183] and absent these actions, the meltdown wouldn't have happened. But the reality was more complicated. The machine side also failed, both in its operations and its interface(s) with humans. The human staff uncovered the reactor core in response to mechanical failures compounded by erroneous instruments that showed a valve closed that was actually open; meanwhile, there were no instruments to measure and show whether the core was covered with water.[184]

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) the machine.[185] It was the interactions between the two, exacerbated 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.[186]

Consider the Boeing 737 Max aircraft. In 2018 and 2019, two Boeing 737 Max planes crashed, killing 346 people.[187] Boeing initially

[182] *Backgrounder on the Three Mile Island Accident*, U.S. NUCLEAR REGUL. COMM'N, https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.html (last updated June 21, 2018) [https://perma.cc/T66E-DPKZ].

[183] *Id.*

[184] *Id.*

[185] WOODS ET AL., *supra* note 129, at 197.

[186] *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.

[187] *See, e.g.*, David Schaper, *Congressional Inquiry Faults Boeing and FAA Failures for Deadly 737 Max Plane Crashes*, NPR (Sept. 16, 2020, 5:46 AM), https://www.npr.org/2020/09/16/913426448/congressional-inquiry-faults-boeing-and-faa-fail ures-for-deadly-737-max-plane-cr [https://perma.cc/H9AU-KZVA]; STAFF OF H. COMM. ON TRANSP. & 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%20Report%20for%20Public%20Release.p df [https://perma.cc/TG8Z-EKGP]; *Boeing Charged with 737 Max Fraud Conspiracy and Agrees to Pay over $2.5 Billion*, U.S. DEP'T OF JUST. (Jan. 7, 2021), https://www.justice.gov/opa/pr/boeing-charged-737-max-fraud-conspiracy-and-agrees-pay-over-25-billion [https://perma.cc/GZ2Q-SC7W].

blamed the flights' pilots.[188] Investigations then uncovered a design flaw in Boeing's automated flight-control system, known as Maneuvering Characteristics Augmentation System ("MCAS").[189] Just as significantly, regulators found 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 adequate information about the system.[190] They discovered interface design failings too: Boeing had failed to design a system that effectively transferred information between MCAS and the pilots.[191] The fatal flaw in the Boeing 737 Max was not (just) pilot error, and not (just) a faulty automated system, but also failures to design the system or to train the humans for effective human-machine interactions.[192]

The tendency to blame the human in the loop for accidents—as opposed to the humans who designed or fielded or failed to correct a flawed system—also manifests in the military context. Take the USS *John McCain* accident, the U.S. Navy's worst accident at sea in the past forty years.[193] 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.[194] After a new navigation system had proved prone to errors, the captain chose to employ it in manual mode.[195] Unknown to him, this removed various safeguards and allowed different helmsmen to unknowingly transfer

[188] 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/ [https://perma.cc/RSQ3-9TRV].

[189] 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 [https://perma.cc/M2EU-GAQH].

[190] U.S. DEP'T OF JUST., *supra* note 187; Julie Johnsson, *Ex-Boeing Pilot Charged with Fraud in 737 Max Probe*, BLOOMBERG (Oct. 15, 2021, 10:36 PM), https://www.bloomberg.com/news/articles/2021-10-14/u-s-charges-ex-boeing-pilot-in-first-max-criminal-prosecution [https://perma.cc/TZC8-JDMU].

[191] *See* Neuman, *supra* note 189 (discussing MCAS software failures); *see also* STAFF OF H. COMM. ON TRANSP. & INFRASTRUCTURE, *supra* note 187, 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.

[192] *See* U.S. DEP'T OF JUST., *supra* note 187 (discussing the lack of information about MCAS in training materials).

[193] T. Christian Miller, Megan Rose, Robert Faturechi & Agnes Chang, *Collision Course*, PROPUBLICA (Dec. 20, 2019), https://features.propublica.org/navy-uss-mccain-crash/navy-installed-touch-screen-steering-ten-sailors-paid-with-their-lives/ [https://perma.cc/L3WS-MZ83].

[194] *See id.*

[195] *See id.*

steering control.[196] 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 seemed to be unsteerable, leading to the collision.[197]

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."[198] Quite the contrary: the Navy investigated, found the human captain at fault, and charged him with homicide—and then committed nearly half a billion dollars to building and installing a modified version of the same problematic navigation system on its destroyers over the next decade.[199]

Choosing which tasks to automate versus what to allocate to humans is far more complicated than the MABA-MABA approach would suggest.[200] Automation isn't a costless substitute for human decisionmaking: its use alters human roles and functions, sometimes unpredictably.[201] Some of these new roles tax humans with tasks that run into known human weaknesses, such as sustaining vigilance.[202] Others lean on humans to do more of a different kind of work than their job once entailed, such as ensuring the algorithm isn't missing necessary parameters for accurate problem solving.[203] Not only do people adapt to technologies they use, they also adapt their use of the technology to their changed and changing practices.[204] In short, "[t]he question for successful automation is not '[w]ho has control over what or how much?' It is '[h]ow do we get along together?' "[205]

[196] *See id.*

[197] *See id.*

[198] *Id.*

[199] *Id.* In response to fleet surveys, the variable touchscreens will be replaced with common physical throttle-and-wheel systems. *See U.S. Navy to Ditch Touch Screen Ship Controls*, BBC NEWS (Aug. 12, 2019), https://www.bbc.com/news/technology-49319450 [https://perma.cc/SQ78-K5PK].

[200] *See* S.W.A. Dekker & D.D. Woods, *MABA-MABA or Abracadabra? Progress on Human-Automation Coordination*, 4 COGNITION, TECH. & WORK 240, 240–41 (2002) (arguing that MABA-MABA methods are overly simplistic).

[201] *Id.* at 241 ("[A]utomation does not replace a human weakness. It creates new human strengths and weaknesses–often in unanticipated ways." (citing Bainbridge, *supra* note 176)).

[202] *See id.* at 241 (citing D.E. BROADBENT, PERCEPTION AND COMMUNICATION (1958)).

[203] *Id.* at 241–42 ("It also exacerbates the system's reliance on the human strength to deal with the parametrisation problem (automation does not have access to all relevant world parameters for accurate problem solving in all possible contexts) . . . .").

[204] *Id.* at 242.

[205] *Id.* at 243.

"Human-centered" design evolved in response to these challenges.[206] A number of fields—including cybernetics, human factors engineering, human-computer interaction, and cognitive systems engineering—developed to address large-scale complex systems that involve passing off tasks between humans and machines. These fields made observations about common errors and developed a host of underlying principles, some of which we discuss below.[207] As best as we can tell, virtually none of this research has been considered by legal academics, policymakers, or practitioners,[208] let alone incorporated into the law of the loop.[209]

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 even policymakers who intend to place a human in the loop rarely articulate what role they want that human to play or what goals they want that human to accomplish. A myopic MABA-MABA focus obscures 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*? If we don't know what the human is intended to do, it's impossible to assess whether a human is improving a system's performance or whether regulation has accomplished its goals by adding a human.

We identify nine possible reasons policymakers might use regulation to include a human in the loop. Our categories are illustrative rather than exhaustive, and they are intended to highlight the variety of (admirable and distasteful) reasons policymakers might wish to encourage human involvement in algorithmic decisionmaking processes. Humans may play (1) corrective roles to improve system performance, including error, situational, and bias correction; (2) resilience roles to act as a failure mode or alternatively stop the whole system from working under an emergency; (3) justificatory roles to increase the system's legitimacy by providing reasoning for decisions; (4) dignitary roles to protect the dignity of the humans affected by the decision; (5) accountability roles to allocate liability or censure;

[206] Jones, *Ironies of Automation*, *supra* note 15, at 110.

[207] *See infra* Part V.

[208] *But see* Roth, *supra* note 15, at 1296–98 (discussing this problem in the context of "trial by cyborg").

[209] 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.

(6) "stand-in" roles to act as proof that something has been done or stand in for other humans and human values; (7) friction roles to slow the pace of automated decisionmaking; (8) "warm body" roles to preserve human jobs; and (9) interface roles to link the systems to human users.[210] None of these roles are mutually exclusive; to the contrary, humans in the loop often fill multiple roles simultaneously.[211] And, importantly, not all of these roles are intended to make a hybrid system more accurate or efficient. A regulator may wish to ensure that there is a human in the loop without regard to (or even with the intention of undermining) performance in the interest of prioritizing other values.

A. Corrective Roles

Perhaps the most straightforward justification for humans in the loop is corrective: due to their as-yet-unique strengths, humans can improve system accuracy.[212] 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 nonetheless reflects a systemic bias that runs counter to social values.[213]

Of course, what constitutes a "right" or "accurate" decision may vary across contexts and evaluators and may be hard to define. And, as

[210] Note that many of these roles reference purposes outside what we define here as the decisional loop itself.

[211] Relatedly, these roles may have unintended side effects. For example, Ben Green argues that—regardless of what role the human is intended to play—human oversight requirements often "legitimize government use[ ] of faulty and controversial algorithms without addressing the fundamental issues with these tools." Green *supra* note 8, at 1, 9.

[212] Accuracy is a critical factor in evaluating the utility of any decisionmaking system. But 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 (for instance, 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 (for instance, 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: MACH. LEARNING & MED. (Feb. 17, 2019), https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/ [https://perma.cc/GA4T-ZN3J].

[213] These forms of correction may be interrelated; improving accuracy may affect the system's distributional impacts, insofar as it alters who is affected by the new, smaller set of errors—and these changes may be predictable or not, yielding different normative implications. We do not explore these complex interrelationships here.

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.[214]

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 results.[215]

As a cautionary tale of using algorithms *without* human error correction, consider the Michigan Integrated Data Automated System ("MIDAS").[216] 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.[217] Unfortunately, the system was prone to error; a later audit found a ninety-three percent *error* rate.[218] In contrast, when there was a human reviewer, a mere (!) forty-four percent of alleged frauds were found to be erroneous.[219] After a class-action lawsuit, the state committed to involving humans in every determination of fraud—

[214] 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 36, at 1990:

> 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.

[215] 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 infra* Part IV.E.

[216] 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/ [https://perma.cc/R5PP-BZ5T].

[217] *Id.*

[218] A MIDAS touch indeed! David Eggert, *Michigan Reverses 44,000 Jobless Fraud Cases, Refunds $21M*, AP NEWS (Aug. 11, 2017), https://apnews.com/article/dc3370d57e264448b67f75ceb63ad120 [https://perma.cc/5EDD-PSCQ].

[219] *Id.*

that is, to ensure that there is always a human in the loop to catch and fix the algorithm's errors.[220]

2. Situational Correction

Alternatively, humans may improve a 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. A system that calculates the risk of a particular medical treatment, for example, may assume the availability of blood transfusions should they 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 algorithm's output or their 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.[221] For example, algorithmic-like criminal sentencing guidelines may be efficient in the sense that they cost fewer resources and less time, but given that they do not take all mitigating factors into account, judges sometimes adjust their results at sentencing.[222]

[220] Maurice & Jane Sugar L. Ctr. for Econ. & Soc. Just. v. Arwood, No. 2:15-cv-11449 (E.D. Mich. Feb. 2, 2017) (dismissed per stipulation), https://www.bwlawonline.com/wp-content/uploads/2017/02/Zynda-ORD-2017-02-02-Robo-Fraud-Settlement-and-Dismissal.pdf [https://perma.cc/ENG8-68JN].

[221] 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.

[222] *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, 881 N.W.2d 749, 768 (Wis. 2016):

> 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."

Situational correction thus can constitute deploying principles of equity when the particularities of human experience outrun our ability (or in this case, a machine's ability) to make general rules. And some believe that only humans, not machines, are able to determine what is fair in outlier cases—what morally "ought" to be done in lieu of the general rule.

3. Bias Correction

Humans may be also expected to correct algorithmic bias, which may manifest as prejudicial or inaccurate results.[223] 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.[224] In addition to biased results due to biased training sets and designs, algorithmic decisionmaking systems also introduce "technical bias"—systemic inaccuracies that result from attempts to translate complex realities into crunchable code.[225]

Accordingly, humans may be included in the loop to identify and counteract observed algorithmic biases. For example, AI-enabled facial

(internal citations omitted).

[223] Kaminski, *Binary Governance*, *supra* note 13, at 1541. This goal is difficult, not least because what constitutes problematic "bias"—and thus what is needed to correct it—is contested. Barocas & Selbst, *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).

[224] BENJAMIN, *supra* note 15, at 5–6 ("[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 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 46, 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. 2007 (2022) (arguing that the data built from certain sources—namely, carceral knowledge sources—will necessarily be biased).

[225] 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, *supra* note 150, at 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.").

recognition raises bias concerns because it has been shown to have a higher rate of inaccuracy for Black women than for White men.[226] Concerns about unequal treatment may have motivated both E.U. and Colorado lawmakers to put a human in the loop when facial recognition systems ("FRS") are used, to serve a bias-correction role by verifying AI identifications.[227]

B. Resilience Roles

Resilience refers to the ability of a complex system to withstand failure by minimizing the harms from bad outcomes.[228] A human in the loop can serve a resilience role by acting as the backstop when an automated system malfunctions or breaks down.[229] For example, human pilots are supposed to be able to take over and fly a plane should autopilot send it into a nosedive. Or a human can recognize that something has gone haywire and use a manual override to stop the crash (think of Homer Simpson and the nuclear power plant meltdown[230]).

C. 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 impacted by it.[231] A core aspect of the legitimacy of legal systems is that people governed by a legal system get to tell their own story, and they are only able to tell their own stories if they know and can thus counter the alternative

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

[227] *See supra* Section II.A.1 (discussing the E.U. and Colorado laws).

[228] 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).

[229] Jones, *Ironies of Automation*, *supra* note 15, at 91 (discussing how humans are always in the loop during inevitable system failures).

[230] *The Simpsons: Homer Defined* (Fox Broadcasting Company television broadcast Oct. 17, 1991).

[231] Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 THEORETICAL INQUIRIES L. 83, 113 (2019); Tom R. Tyler, *Psychological Perspectives on Legitimacy and Legitimation*, 57 ANN. REV. PSYCH. 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.

stories being told.[232] For instance, it may be particularly important to the affected party to be provided a justification for the length of a prison sentence, the refusal to grant parole, or a firing decision, so that the affected party can decide whether to live with the decision or how best to contest it. 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 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.[233] 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 or responsive 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.[234] Given this, 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—for example, as required in French administrative law.[235] Otherwise, the human's explanation is little more than a post hoc rationalization. Some humans in the loop may play an only deceptively justificatory role.[236]

[232] Lawrence B. Solum, *Procedural Justice*, 78 S. CAL. L. REV. 181, 274 (2004).

[233] Price, *Regulating Black-Box Medicine*, *supra* note 17; Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1, 9 (2016) ("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.").

[234] Kirsten Martin & Ari Ezra Waldman, *Governing Algorithmic Decisions: The Role of Decision Importance and Governance on Perceived Legitimacy of Algorithmic Decisions*, 2022 BIG DATA & SOC'Y 1, 9. 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 6.

[235] Gianclaudio Malgieri, *Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations*, COMPUT. L. & SEC. REV., Oct. 2019, art. 105327.

[236] *See* Brennan-Marquez et al., *supra* note 15, 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

D. Dignitary Roles

Some argue that subjecting humans to algorithmic decisions on significant subjects violates human dignity. Indeed, in Europe, some characterize having a human in a decisionmaking loop as a "fundamental right."[237]

Dignity is hard to define and harder to quantify. It is about affording respect to individuals and ensuring that they may exercise their fundamental freedoms, including the right to self-determination. Violations boil down to a fundamental interference in selfhood—an injury that goes to the core of who a person is.

In contrast to respecting and affording human dignity, objectification reduces a person to an inhuman object, rendering them reified, often static.[238] Or, as Tal Zarsky puts it, "individuals should be treated as fellow persons and not mere machines."[239] Lee Bygrave explains that the E.U.'s law restricting automated decisionmaking reflects a central concern with objectification.[240] That is, using an AI system to make a significant decision about a person is inherently objectifying, reducing a person to their "data shadow" and showing an inherent disrespect for their humanity.[241]

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.").

[237] *See* Jones, *The Right to a Human in the Loop*, *supra* note 27, at 230 (describing European "insistence on the categorization of . . . a human in the loop as a fundamental right"); EUR. UNION AGENCY FOR FUNDAMENTAL RTS., GETTING THE FUTURE RIGHT: ARTIFICIAL INTELLIGENCE AND FUNDAMENTAL RIGHTS 60 (2020) ("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."); Kaminski, *Binary Governance*, *supra* note 13, at 1542–45 (summarizing and classifying three dignitary arguments regarding algorithms).

[238] Martha C. Nussbaum, *Objectification*, 24 PHIL. & PUB. AFFS. 249, 256–57 (1995) (arguing that there are seven forms of objectification: instrumentalizing to achieve a further purpose; denying autonomy; treating as inert, as fungible, as violable, as owned by another person; and denying subjectivity).

[239] Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1552. Zarsky, however, disagrees with this stark view of the effects of automated decisionmaking on dignity, calling it "neo-Luddite" and "anachronistic." *Id.*; *see also* JOHNNY RIVERS, *Secret Agent Man*, *on* AND I KNOW YOU WANNA DANCE (Imperial Rec. 1966) ("They've given you a number/And taken away your name").

[240] Lee A. Bygrave, *Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling*, 17 COMPUT. L. & SEC. REP. 17, 18 (2001) (voicing a concern that "the registered data images of persons (their 'data shadows') threaten to usurp the constitutive authority of the physical self despite their relatively attenuated and often misleading nature. . . . [T]his threat brings with it the threat of alienation and a threat to human dignity.").

[241] *See* Jones, *The Right to a Human in the Loop*, *supra* note 27, at 232; 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.").

Concerns over respecting dignity and avoiding objectification tie to closely related concerns about a lack of transparency and due process. Automated decisionmaking arguably reduces individuals' ability to self-constitute; that is, it denies individuals the opportunity to push back or define their individual selves during critical decisions.[242] When decisions are made for you without the opportunity for participation, you have less freedom, and you become trapped in the way you have been constituted by others. Think Kafka's *The Trial*.[243] Mireille Hildebrandt links dignity to due process in the data analytics context, arguing for legal protection for what she calls the "incomputable self."[244]

Some argue that placing a human in the loop helps alleviate dignitary concerns. The strong form of this argument is that adding a human touch renders a decision inherently more humane.[245] That is, the very humanness of the human in the loop inherently reduces a risk of objectification. Placing the human in the loop demonstrates respect for the human subject.

The weaker and perhaps more palatable form of the dignitary argument brings in arguments from other Subsections of this Part.[246] A human in the loop serves some functional role that makes an automated decision less inherently objectifying. Whether that role entails providing more individualized context (situational correction) or affording affected individuals the ability to contest a decision, thus enabling participation and voice (justificatory roles and accountability roles), the importance of the human in the loop is less about their humanity and more about the roles that they serve.[247]

[242] *See* Jones, *The Right to a Human in the Loop*, *supra* note 27, at 230–32; Zarsky, *supra* note 241, at 1016–17.

[243] Dan J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393 (2001) (employing the Kafka metaphor).

[244] Hildebrandt, *supra* note 231, at 86:

> [A]t our essence is that we are incomputable, meaning that any computation of our interactions can be performed in multiple ways—leading to a plurality of potential identities. The need to navigate this plurality is what shapes and nourishes our agency; to deny or reduce this plurality is to diminish our agency.

(emphasis omitted).

Hildebrandt argues that a "practical and actional right to reject computation and/or to be computed in alternative ways . . . underlin[es] the indeterminate nature of each and every individual person and the 'equal respect and concern' that our governments owe each of them." *Id.* at 121.

[245] *See, e.g.*, Bygrave, *supra* note 240.

[246] *See supra* Parts IV.A, IV.C; *infra* Part IV.E; *see also, e.g.*, Zarsky, *supra* note 239, at 1547–48 ("To assure dignity and lack of targeting, the individual should receive assurances as to the precision, effectiveness, and lack of discrimination in the process.").

[247] Kaminski & Urban, *supra* note 36; *see also* Rebecca Crootof, *The Internet of Torts: Expanding Civil Liability Standards to Address Corporate Remote Interference*, 69 DUKE L.J. 583,

Of course, as is the case throughout this typology, humans could in practice merely rubber-stamp bureaucratized decisions without fulfilling the lofty goals their presence is intended to achieve. Moreover, while dignity is often characterized as paramount by those who subscribe to it, it is also often not the only value at play. Sometimes there is a need to balance legitimate protections for dignity against protections from other significant harms, including other dignitary harms.[248]

E. Accountability Roles

Some fear that humans will delegate difficult decisions to algorithms out of a desire to duck responsibility for undesirable outcomes.[249] As a result, sometimes humans will be included in the loop to ensure that someone is legally liable, morally responsible, or otherwise accountable for the system's decisions.[250] A more negative view of legal accountability suggests a human might be there so that she can be influenced by powerful actors, including (perhaps captured) regulators. Even more cynically, sometimes the human is there to be the fall guy for an organization or for the algorithm's developer. (Recall how Tesla hands off control to human drivers mere milliseconds before a crash, allowing Elon Musk to tout that *technically* Tesla's autopilot has never been on during an accident.)[251]

If the human in the loop has the power, information, judgment, and time to make the final decision in the human-algorithmic system,

655–56 (2019) (discussing a Connecticut law that requires corporations to publish the name and number of a human who can process complaints about electronic self-help measures).

[248] *See, e.g.*, Rebecca Crootof, *A Meaningful Floor for "Meaningful Human Control,"* 30 TEMP. INT'L & COMP. L.J. 53 (2016).

[249] *E.g.*, 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 [https://perma.cc/LKJ7-7X32] ("[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 OF HEALTH (Feb. 10, 2017), https://blog.petrieflom.law.harvard.edu/2017/02/10/artificial-intelligence-and-medical-liability-part-ii/ [https://perma.cc/4PES-TGUU] ("[B]y decreasing the degree of discretion physicians exercise in diagnosis and treatment, medical algorithms could reduce the viability of negligence claims against health care providers."); W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419, 457 n.188 (2015) ("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?").

[250] *E.g.,* Bettina Berendt & Sören 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).

[251] *See supra* notes 29–33 and accompanying text.

then the human might legitimately be held responsible. Consider clinical decision support software that makes recommendations but specifies that it exists simply to present information that should be taken into account by the physician. As the system is envisaged, the physician maintains the authority—and consequently the moral and legal responsibility—for the final decision.

But responsibility can also be assigned to a human 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.[252]

All humans in the accountability role—both legitimate ones and "liability sponges"—may simultaneously serve as a "moral crumple zone."[253] Both soak up liability, but the moral crumple zone explicitly does so to protect another entity: "While 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 human operator."[254] Not only does the human in the loop protect the system itself from censure, but 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.[255] U.S. judges, for example, regularly attribute tort liability for accidents involving robots to a human in the loop, rather than to a robotic system or relevant remote decisionmakers.[256]

[252] M.C. Elish & Tim Hwang, *Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation* 15 (Compar. Stud. in Intelligent Sys., Working Paper #1 V2, 2015), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2720477 [https://perma.cc/JMS8-6GRM].

[253] Elish, *Moral Crumple Zones, supra* note 91, 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.

[254] *Id.*

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

[256] Calo, *supra* note 24, 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).

F. Stand-In Roles

Sometimes a human may be in the loop just to "stand in" for regulators in an abstract sense. This stand-in role consists of demonstrating that, just in case there is something wrong with automation, something has been done. In this sense, a human in the loop can be a sort of proof of work—whether that work (be it corrective, dignitary, or other) has actually happened or not. We suspect that many humans in the loop currently play stand-in roles.

A human in the loop in a stand-in role might nonetheless be effective at one or more of the other roles discussed. Or they might be a mere symbol, a box-ticking practice. We don't know quite why they're there but, like performative royalty, they must mean something, and they get trotted out on special occasions.[257]

G. Friction Roles

Humans' relative slowness may be viewed as a benefit, as many harms arise due to algorithmic speed. Requiring a human in the loop may slow the operation of a system or slow its adoption in useful ways.[258]

The reader will be familiar with clicking an "I Am Not A Robot" checkbox (which measures mouse-cursor irregularities) or struggling to identify which parts of an image match some arbitrary criterion ("click on all images of a pipe").[259] These commonplace hurdles require us to become a human in a loop to block algorithmic systems from completing transactions without human involvement—and thus stop fraud. Humans can similarly be imposed into loops to slow system performance; as we noted above, regulators could combat the risks of high-frequency trading or decrease its comparative advantage by requiring human approval of individual trades.[260]

At a broader level, humans could be included to slow the adoption of automated systems, for instance, by making them more idiosyncratic and less interoperable. While interoperability is often a goal of system designers, there are legitimate reasons (for example, security concerns) and market incentives (industries' interest in creating impenetrable device ecosystems) to not design for

[257] Thanks to Paul Ohm for this analogy.

[258] Paul Ohm & Jonathan Frankle, *Desirable Inefficiency*, 70 FLA. L. REV. 777 (2018).

[259] *E.g.*, Senior Oops Engineer (@ReinH), TWITTER (June 11, 2019, 12:52 PM), https://twitter.com/reinh/status/1138504313469194240 [https://perma.cc/4D6M-CHXA].

[260] Ohm & Frankle, *supra* note 258, at 781–83.

interconnectivity.[261] Privacy scholars have been arguing for years that a number of surveillance harms stem from increased efficiencies brought on by surveillance technology.[262] Humans can slow it all down.

H. "Warm Body" Roles

Concerns about technology displacing humans have long existed.[263] In response, humans are sometimes included in a loop to protect their jobs. Humans can fulfill this role merely by being present; whether, how, or how well they contribute to the ultimate result of decisionmaking is largely irrelevant.

Amidst ongoing debates about whether AI is going to replace certain types of physicians,[264] 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.[265] "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."[266] Doctors value their jobs;[267] ensuring a role for

[261] *See, e.g.*, Brett Frischmann & Madelyn Rose Sanfilippo, Privacy Law Scholars Conference 2022: A Principled Decision-Making Approach to Smart Tech Governance in Cities (June 2, 2022) (discussing interoperability in the context of smart cities).

[262] *See, e.g.*, Harry Surden, *Structural Rights in Privacy*, 60 SMU L. REV. 1605 (2007); Orin S. Kerr, *An Equilibrium-Adjustment Theory of the Fourth Amendment*, 125 HARV. L. REV. 476 (2011).

[263] Rather than being anti-technology, the much-maligned original "Luddites" were opposed to the ill-treatment of underskilled 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, at 26.

[264] *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 [https://perma.cc/486L-CGTU] (cataloging different viewpoints on the likelihood of AI replacing many radiologists).

[265] *Augmented Intelligence in Medicine*, AMA, https://www.ama-assn.org/amaone/augmented-intelligence-ai (last visited Oct. 2, 2022) [https://perma.cc/2FVJ-Z4VL].

[266] *Id.*

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

themselves within algorithmic systems is one way to protect those jobs. Lawyers (and legal academics!) similarly emphasize the importance of keeping human lawyers involved in legal processes rather than relying fully on AI.[268] And fighter pilots push back hard against the idea that they can be replaced by drones.[269]

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."[270] 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.[271] 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."[272] Further, Pasquale emphasizes 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.[273]

One notable feature of the warm body role is that it prioritizes the worth of the individual human in the loop, rather than the humans on whom 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 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 nonincumbents and other stakeholders.

[268] *See, e.g.*, Jerry Levine, *Lawyers Can Be More 'Human' with the Help of AI. Here's How.*, ABOVE THE L. (Sept. 23, 2021, 9:58 AM), https://abovethelaw.com/legal-innovation-center/2021/09/23/lawyers-can-be-more-human-with-the-help-of-ai-heres-how/ [https://perma.cc/7JNW-Z943].

[269] 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/ [https://perma.cc/K587-ZQTH].

[270] PASQUALE, NEW LAWS OF ROBOTICS, *supra* note 8, at 3.

[271] *Id.* at 4.

[272] *Id.* at 12–13; *see also supra* notes 168–172 and accompanying text.

[273] PASQUALE, NEW LAWS OF ROBOTICS, *supra* note 8, at 170–99.

I. Interface Roles

Finally, humans can also serve an interface role, helping users interact with an algorithmic system. Sometimes, it's just easier, cheaper, or faster to retain or insert a human link than to create a user-friendly interface.[274] 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.[275] Conversely, 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.[276] 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.[277] 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 solution, as there is no one-size-fits-all regulation to apply. Instead, we offer three recommendations for policymakers who are thinking about how the law might be employed to improve human-in-the-loop systems.

First, policymakers should be intentional and clear about what role they want a human in the loop to serve. 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. Some examples of what we mean by context include: the existing law and institutions in a particular legal field; the kinds of harms at issue (for example, physical harms versus

[274] *See* Brennan-Marquez et al., *supra* note 15, at 754–55.

[275] *Id.* at 754 (discussing the DMV clerk example).

[276] *See id.* at 755.

[277] *Id.* If so, the human may also be serving a justificatory role. *See supra* Part IV.C.

dignitary harms); the legal construction of those harms (for example, how standing doctrine allows some kinds of harms to be brought into court but not others); how much urgency or speed matters; broader societal dynamics, including distributions of and access to power; and what the specific humans at issue are capable of bringing to the table (for example, what a doctor can or cannot do compared to the driver of a car). 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.[278]

Third, governance should regulate the system as a whole, as focusing too narrowly on just the human in the loop will frequently lead to failure. Various case studies of past regulatory strategies model how this might be accomplished. We close out our recommendations by highlighting that there are other, often-overlooked systemic regulatory approaches that can be complementary to or even more effective than focusing on just the hybrid system.

We then consider two case studies. One focuses on the draft E.U. AI Act, to show how extant law attempting to regulate human-in-the-loop systems comes up short; the second applies our suggested process to regulating law enforcement use of facial recognition systems to demonstrate how regulation could be done better.

A. Clarify 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 become substantially more difficult.[279] Conversely, identifying the intended role(s) fosters systems and organizational design that ensure the human in the loop has the needed authorities and capabilities.

[278] Yes, we see the irony inherent in this sentence.

[279] *See, e.g.*, NAT'L RSCH. COUNCIL, HUMAN ENGINEERING FOR AN EFFECTIVE AIR NAVIGATION AND TRAFFIC-CONTROL SYSTEM 11 (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.").

1. Why Clarify?

Explicitness of purpose is necessary to determine what ability and agency a human in the loop must have.[280] 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.

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 rights."[281] 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 addressing harms to fundamental rights, like dignitary harms.[282] Thus, the Act's human oversight requirement appears to be motivated by a mixture of corrective and dignitary goals.

If 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 role (mandated, presumably, by labor unions), or to serve as a liability sponge (though we suspect system designers will be loath to admit this).[283]

To the extent that designers may resist clarifying roles—or casting them accurately—regulators could offer carrots or threaten consequences. A well-defended characterization could win the benefits of a regulatory safe harbor, where the system was subject to less

[280] There are various questions about the identity and characteristics of the human in the loop—including what abilities they must have, how they should be trained, and who are they (in terms of identity, representativeness, and other characteristics)—which we do not tackle here.

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

[282] Veale & Zuiderveen Borgesius, *supra* note 65, 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.").

[283] 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.

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 better argue about role priority and success.

2. Role Complexities

Identifying relevant roles will often be 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). Alternatively, humans can be included in the loop in such a way as to have essentially no performance impact—but to fulfill a dignitary, justificatory, "stand-in," or other rationale. 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.[284] Consider the constrained-but-blamable clerk at the Department of Motor Vehicles, who can only respond to complaints of inflexibility with a rote "that's all the computer will let me do."

Another set of problematic interactions arises 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, 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. It is beneficial to articulate these "warm body," job-preserving aims when they exist, to foster a transparent debate on the benefits and trade-offs 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.

[284] Brennan-Marquez et al., *supra* note 15.

B. Consider Context

Context matters. Regulators should also consider the greater context within which the hybrid system is operating. This will include the contextual law in a particular legal field (including all of the subtle influences associated with the already-existing law-of-the-loop),[285] as well as specifics about the human who will be placed in the loop, the kinds of benefits and harm at issue, broader societal dynamics, and more.

Although the features of humans in the loop have some commonalities across fields, the importance of different roles varies by context. For example, dignitary and justificatory rationales are less important (though nonnegligible) when the relevant decisions do not impinge the dignity of human beings. Incoming missiles shot down by an automated defense system have no particular dignitary claim to a human making the targeting decision.[286] Using AI to optimize telecommunications networks, predict needed sewer maintenance, and reduce energy use arguably also has minimal direct impact on human dignity.[287] 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 nonetheless counsel in favor of retaining a central role for humans in sentencing processes.[288] This is arguably true of the judicial system more broadly speaking. In France, for example, automated decisionmaking is banned in the judicial context and limited elsewhere in the administrative state.[289]

Context also matters in terms of determining what exactly human intervention brings to the table. The value of a human in overcoming bias depends on the relative bias of human decisionmakers and algorithms, which will change depending on the field, human predilections, and the data available. Some humans are experts. Some have no training. Some are managers, and some are peons. Some work

[285] *See supra* Part II.

[286] *See, e.g.*, Jen Kirby, *Israel's Iron Dome, Explained by an Expert*, VOX (May 14, 2021, 3:40 PM), https://www.vox.com/22435973/israel-iron-dome-explained [https://perma.cc/YPR3-SUXC] (describing Israel's automated air defense system).

[287] *See, e.g.*, *AI in Networks*, ERICSSON, https://www.ericsson.com/en/ai (last visited Dec. 28, 2022) [https://perma.cc/DSC7-9ANP] (explaining how Ericsson, a telecommunications company, incorporates AI in its networks).

[288] *E.g.*, Kiel Brennan-Marquez & Stephen Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137, 147 (2019); Crootof, *supra* note 15, at 238–42.

[289] Malgieri, *supra* note 235.

in systems that provide authority, reporting infrastructure, and even whistleblower protection. Others are deeply embedded in the culture and rationales of an organization using automated decisionmaking, such that inserting a human in the loop is very unlikely to alter (possibly company-preferred) outcomes.[290]

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 reflects the problem. Context includes not just the capacities and capabilities of particular humans but entire organizational and legal infrastructures within which they are embedded—and which are themselves shaped by humans. As we discuss in the next Section, there is much that we can learn from experience about regulating systems as complex systems.

C. Regulate Hybrid Systems Using Lessons from Engineering

With a clear understanding of the human role and larger relevant context, policymakers can take steps to regulate the hybrid system as a whole. Three examples of successful hybrid regulation provide some generalizable lessons on how to go about that task.

1. Examples of Successful Hybrid Regulation

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, 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

[290] *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).

with complex human-machine systems that have long presented analogous concerns about human-machine interfaces, inadequate operator training, and failure cascades.[291]

a. Railroads

Federal Railroad Agency ("FRA") regulations provide an example of how a regulator can consider systems-level human factors and create tremendously detailed rules for system designers and operators. The agency 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."[292] 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.[293]

For example, the FRA requires product designers of Positive Train Control Systems, intended to automatically stop a train to prevent an accident, to address the "[h]uman factor engineering principle."[294] Specifically, product designers must consider an operator's limited ability to process large amounts of information,[295] imperfect long term memory,[296] and expectation that there will be "consistent relationships between actions and results."[297] Designers must create a system such that the human operator is "in the loop" for a minimum of thirty minutes at a time; does not lose situational awareness and is not

[291] Jones, *Ironies of Automation*, *supra* note 15, at 92–93 (discussing, among others, railroad law as an example of the law of automation).

[292] Safety Assurance Criteria and Processes, 49 C.F.R. pt. 236, app. C(a) (2022).

[293] Recommended Practices for Design and Safety Analysis, 49 C.F.R. pt. 229, app. F (2022); Human-Machine Interface (HMI) Design, 49 C.F.R. pt. 236, app. E (2022); *see also* Railroad Safety Program Plan (RSPP), 49 C.F.R. § 236.905(b)(3) (2022) ("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."); Product Safety Plan (PSP), 49 C.F.R. § 236.907(a)(11) (2022) ("The PSP must include the following: . . . (11) A human factors analysis, including a complete description of all human-machine interfaces . . . .").

[294] 49 C.F.R. pt. 236, app. C(b)(5). 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.*

[295] 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)).

[296] *Id.* at (c)(4)(ii).

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

overly reliant on the machine;[298] has adequate warning before she must take action; and receives "timely feedback . . . regarding the system's automated actions, the reasons for such actions, and the effects of the operator's manual actions on the system."[299] Additionally, the regulations note that the interface design shouldn't itself distract the operator from safety-related duties.[300]

Rather than assuming human perfection or blaming humans for foreseeable problems, the FRA requires product designers to create complex technological systems around the persons operating them to maximize system success. The regulations are detailed, addressing the design of displays and controls as well as the design of communications to the human operator, all with an awareness of the operator's physical characteristics, education, and cognitive capacity.[301] Designers must "locate displays as close as possible to the controls that affect them," "arrange controls according to their expected order of use," "group similar controls together," and "design controls to allow easy recovery from error."[302] 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.[303]

In addition to design, the FRA addresses training[304] and organizational policies.[305] It also establishes risk-management processes, suggesting that companies assess complex human-machine systems—as systems—for risks of failure.[306] For example, companies are required to keep hazard logs and document assumptions about human performance so they can later be evaluated for accuracy.[307]

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

[299] *Id.* at (c)(1)(ii).

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

[301] *See id.* at (e)(9) ("Design 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 . . . .").

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

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

[304] *See, e.g.*, Training, 49 C.F.R. § 228.411(b) (2022) (discussing training to reduce fatigue).

[305] *See* Guidance on Fatigue Management Plans, 49 C.F.R. pt. 228, app. D (2022).

[306] *See* Recommended Practices for Design and Safety Analysis, 49 C.F.R. pt. 229, app. F(f)(3) (2022) ("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 . . . . ").

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

b. Nuclear Reactors

The Nuclear Regulatory Commission ("NRC") also incorporates human factors in regulation of nuclear reactors, but in far more general terms than the FRA's regulations. The NRC refers to human factors engineering in regulation[308] but leaves the details to external documents, including related Guidelines,[309] the Three Mile Island ("TMI") Action Plan Report,[310] and specific company documents incorporated into certification standards by reference.[311]

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."[312] 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."[313]

Rather than spell out relevant considerations, the regulations reference the NRC's TMI Action Plan Report and subsequent related NRC Guidelines.[314] The NRC's 563-page Human-System Interface Design Review Guidelines provide detailed instruction for everything from interface displays and user-interface interaction to alarm systems and the design of workstations.[315] The Guidelines additionally reference

[308] *See, e.g.*, 10 C.F.R. § 50.34(f)(2)(ii) (2022).

[309] *See* U.S. NUCLEAR REGUL. COMM'N, HUMAN-SYSTEM INTERFACE DESIGN REVIEW GUIDELINES (2020), https://www.nrc.gov/docs/ML2016/ML20162A214.pdf [https://perma.cc/36BN-37YV] [hereinafter DESIGN REVIEW GUIDELINES]; *see also* U.S. NUCLEAR REGUL. COMM'N, CLARIFICATION OF TMI ACTION PLAN REQUIREMENTS 3–51 (1980), https://www.nrc.gov/docs/ML0514/ML051400209.pdf [https://perma.cc/5DKN-HWV2].

[310] This report was made in response to the Three Mile Island disaster. U.S. NUCLEAR REGUL. COMM'N, NRC ACTION PLAN DEVELOPED AS A RESULT OF THE TMI-2 ACCIDENT (1980), https://www.nrc.gov/docs/ML0724/ML072470524.pdf [https://perma.cc/5A5Y-AY7Z] [hereinafter NRC ACTION PLAN DEVELOPED]; U.S. NUCLEAR REGUL. COMM'N, 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 [https://perma.cc/BQZ4-NJHW].

[311] *See, e.g.*, Design Certification Rule for the ESBWR Design, 10 C.F.R. pt. 52, app. E(II)(A) (2022) (incorporating documents provided in its GE-Hitachi Nuclear Energy application for certification of the Economic Simplified Boiling-Water Reactor ("ESBWR") design).

[312] 10 C.F.R. § 50.34(f)(2)(iii) (2022).

[313] *Id.* § 50.34(f)(2)(ii) (emphasis added).

[314] *Id.* § 50.34(f)(1) n.10, (f)(2)(ii) (referencing DESIGN REVIEW GUIDELINES, *supra* note 309; NRC ACTION PLAN DEVELOPED, *supra* note 310, at 3–51).

[315] DESIGN REVIEW GUIDELINES, *supra* note 309. Section 9 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.*

operator training programs and processes for verification and validation.[316]

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 intended user."[317]

Like the NRC, FDA has not promulgated regulations 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.[318] 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's user interface.[319] It discusses processes such as validation testing and actual use testing.[320] Like other regulators, FDA directs companies to consider user training as an aspect of device operation and risk mitigation.[321]

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."[322] 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.[323] 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.[324]

[316] *Id.* at 9-8; B-16 to B-17.

[317] *E.g.*, External cardiac compressor, 21 C.F.R. § 870.5200(b)(4) (2022). Similar requirements apply to Cardiopulmonary resuscitation (CPR) aid, 21 C.F.R. § 870.5210 (2022); Coronary vascular physiologic simulation software device, 21 C.F.R. § 870.1415 (2022); and elsewhere.

[318] FDA, APPLYING HUMAN FACTORS AND USABILITY ENGINEERING TO MEDICAL DEVICES: GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF (2016), https://www.fda.gov/media/80481/download [https://perma.cc/D5HN-M8P6].

[319] *Id.* at 4, 7–11.

[320] *Id.* at 21, 28–29.

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

[322] *Id.* at 11.

[323] *Id.* at 10.

[324] *Id.*:

* * *

These three examples from U.S. law showcase three possible models of how to regulate the design of a complex human-in-the-loop system. Regulators could promulgate formal and detailed regulations dictating the design of user interfaces, precise training requirements, and organizational measures. 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.[325] And these models are only three of many ways to more comprehensively regulate human-in-the-loop systems.[326]

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.

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

There are many lessons to be learned from the regulation of safety-critical systems. Clearly, if you're going to put a human in the loop, you need a *lot* of implementing regulations to ensure the system 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

> 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.

[325] To the extent that each of these approaches envisions 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).

[326] *Cf.* Green *supra* note 8, at 14 (arguing that policymakers should require that vendors or agencies conduct preliminary evaluations of whether people can collaborate with the algorithm in a desirable manner before adopting algorithmic systems).

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 to 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.

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 will handle failure.[327]

Well-designed interfaces are critical. A poorly designed or inadequate interface creates opportunities for critical information to be garbled or lost in translation. As discussed above, during the Three Mile Island meltdown, there were no sensors to detect how much water covered the nuclear reactor, and an instrument erroneously communicated that an open valve was closed.[328] Similarly, Boeing failed to design an effective interface between pilots and the MCAS system, and the U.S. Navy failed to deploy ships with an effective interface between helmsmen and the navigation system.[329] 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.[330]

Machines must also be built to respond to useful human interaction.[331] 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 an individual application; input by the head of a medical practice who determines that there has been a pattern of errors over multiple uses;

[327] *See* Dekker & Woods, *supra* note 200, at 7; *see also* Robin R. Murphy & David D. Woods, *Beyond Asimov: The Three Laws of Responsible Robotics*, IEEE INTELLIGENT SYS., July–Aug. 2009, at 14, 17–18.

[328] *See supra* Part III.C; *see also* Elish, *Moral Crumple Zones*, *supra* note 91, at 40–50 (detailing how human-in-the-loop interface design issues contributed to the Three Mile Island and Air France Flight 447 disasters).

[329] *See supra* Part III.C.

[330] *See* Dekker & Woods, *supra* note 200, at 7.

[331] *See* Murphy & Woods, *supra* note 327, at 17–18.

and input by the hospital administration trying to determine hospital policy over when the algorithm should be used and by whom.[332] Meanwhile, when there is a handoff between machine and human, 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.[333]

Technological elements of hybrid systems should be designed to be resilient and facilitate organizational resilience.[334] Failure, or at least encountering unpredicted events, is inevitable for complex systems.[335] Principles from resilience engineering include employing a safe failure mode (such as "return home" or "stop movement"),[336] designing for smooth transfer of control,[337] and recording crash events using a black box for purposes of accident diagnosis and future debugging. 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.[338]

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

[332] *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: NETWORKS, ALGORITHMS, HUMANITY 137 (Kevin Werbach ed., 2020).

[333] *See* Murphy & Woods, *supra* note 327, at 18 ("[B]umpy transfers of control have been noted as a basic difficulty in human interaction with automation that can contribute to failures.").

[334] *Id.* 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).

[335] Waldemar Karwowski, *A Review of Human Factors Challenges of Complex Adaptive Systems: Discovering and Understanding Chaos in Human Performance*, 54 HUM. FACTORS 983, 985 (2012) ("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.").

[336] Murphy & Woods, *supra* note 327, at 17; *see* Choi, *supra* note 156.

[337] Murphy & Woods, *supra* note 327, at 18.

[338] *Id.* ("[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.").

machine.[339] 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.)[340]

Other known cognitive factors, such as the human tendency to use heuristics or oversimplify, have implications for training materials and practices,[341] especially for training nonexperts 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.[342] Jones observes that "the more advanced and reliable the automation, the more important the human operator must be."[343] Training can enable humans to develop these needed skills.

Both interface design and training can affect automation bias, a common human willingness 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 otherwise be warranted.[344] 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 system's report of launched U.S. missiles was probably a false alarm.[345] In contravention of his orders, he decided against responding with force—then sweated for twenty minutes before finding out he was right.[346]

But undertrust creates new problems. One of the aims of introducing machine intelligence into a decisionmaking system is to

[339] *Id.* at 15.

[340] Natalie Kitroeff & David Gelles, *In Reversal, Boeing Recommends 737 Max Simulator Training for Pilots*, N.Y. TIMES, https://www.nytimes.com/2020/01/07/business/boeing-737-max-simulator-training.html (last updated Jan. 8, 2020) [https://perma.cc/QBY5-BBVS].

[341] WOODS ET AL., *supra* note 129, at 13; *see also id.* at 17 (discussing training for flexibility under pressure).

[342] DAVID D. WOODS, EMILY S. PATTERSON & EMILIE M. ROTH, CAN WE EVER ESCAPE FROM DATA OVERLOAD? A COGNITIVE SYSTEMS DIAGNOSIS 19 (1998) https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.6079&rep=rep1&type=pdf [https://perma.cc/XBX8-R3TF] ("[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.").

[343] Jones, *Ironies of Automation*, *supra* note 15, at 91.

[344] *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/ [https://perma.cc/FP3B-54KY].

[345] Marc Bennetts, *Soviet Officer Who Averted Cold War Nuclear Disaster Dies Aged 77*, GUARDIAN (Sept. 18, 2017, 11:24 AM), https://www.theguardian.com/world/2017/sep/18/soviet-officer-who-averted-cold-war-nuclear-disaster-dies-aged-77 [https://perma.cc/ZL79-3DL5].

[346] *Id.*

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[347] or decreased performance when humans overrule correct but counterintuitive algorithmic decisions.[348] Further, undertrust can also contribute to accidents. Due to its repeated glitches, the USS *John McCain*'s captain quickly learned not to trust the new navigation system, so he often used it in backup manual mode.[349] Unfortunately, this practice created new risks, as the manual mode disabled built-in safeguards, the lack of which made the *McCain* disaster possible.

Given these known risks, regulators might want to consider mandating certain levels of expertise, experience, or training (which in turn should aim to provide humans with appropriate levels of trust in the system), the ability to identify anomalous results, and the capacity to act accordingly.

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."[350] That is, when individual humans in the loop fail, it is often not just because there are underlying problems with technological design and human preparation but also because there are problems with organizational dynamics.[351]

Organizations shape human-machine interactions in many ways: they select and empower (or not) humans for the loop, assign

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

[348] Price, Gerke & Cohen, *Potential Liability for Physicians Using Medical AI*, *supra* note 118, at 1765.

[349] Miller et al., *supra* note 194. 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*, BMC MED. INFORMATICS & DECISION MAKING, Apr. 10, 2017, art. 36.

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

[351] *Id.*

workloads, and help frame the decisions humans must make.[352] Organizational design and choices regarding individual training and resourcing can make considerable differences in making the human in the loop more effective at her task. It can also make the system more resilient as a whole: resilience engineering suggests that organizations should have a plan for failure, with checklists for contingencies.[353] If technological resilience fails, as it often does, then organizational resilience can provide a backstop.

Perhaps most significantly, organizations can play a large role in affecting whether individual humans in the loop are in a "goal conflict." Failures in complex systems often occur when a human gets trapped in juggling hard-to-reconcile goals (such as safety versus cost).[354] An organization can work to ensure that individual human operators rarely get stuck in these double binds.

For example, in response to the opioid crisis, many states decided 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.[355] The company that created the algorithm, Appriss, was "adamant that a NarxCare score is not meant to supplant a doctor's diagnosis."[356] That is, Appriss insisted that doctors remain in the loop. But doctors are placed in a double bind: overruling the algorithm means risking prosecution for overprescribing or for prescribing to a patient deemed high risk. 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,[357] organizational dynamics will discourage doctors from doing much to correct it.

Human factors engineering and related fields have much to teach regulators about hybrid systems. We hope these approaches will be considered more broadly as regulators continue to address hybrid systems.

[352] *Id.*

[353] *Id.*

[354] *Id.* at 159 (observing that "goal conflicts played a role in the accident evolution, especially when they place practitioners in double binds"); *id.* at 160 ("[C]onstraints imposed by organizational or social context can create or exacerbate competition between goals.").

[355] Maia Szalavitz, *The Pain Was Unbearable. So Why Did Doctors Turn Her Away?*, WIRED (Aug. 11, 2021, 6:00 AM), https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/ [https://perma.cc/26X7-QV73]; *see also* Jennifer D. Oliva, *Dosing Discrimination: Regulating PDMP Risk Scores*, 110 CALIF. L. REV. 47, 51 (2022) (describing PDMP algorithms and physician reliance on them).

[356] Szalavitz, *supra* note 355.

[357] *Id.*

D. Two Case Studies

What might it look like to actually use our process in regulating algorithmic systems? We present two illustrative case studies. The first is something of an anti-example: we point out how the most prominent case of algorithmic regulation, the draft E.U. AI Act, encounters problems across multiple dimensions of effective hybrid regulation. In the second case study, we apply our suggested process to the problem of facial recognition systems: we identify what role or roles a human is there to serve, consider context, and suggest regulatory approaches which incorporate current engineering best practices.

1. The 2021 Draft E.U. AI Act as Anti-Example

At first look, the 2021 draft E.U. AI Act appears to comply with many of the regulatory principles we emphasize. The Act requires the providers of high-risk AI systems to build systems according to a set of standards[358] and conduct a conformity assessment before placing the systems on the market.[359] For example, the Act requires the providers of high-risk AI systems to design systems specifically for human oversight,[360] and the soft law accompanying the Act emphasizes the necessity of educating and training the human providing oversight.[361]

In many ways, the Act emphasizes systemic risk mitigation through ex ante design accompanied by ongoing monitoring—principles embraced by human factors design. The Act tasks providers with performing ex ante risk assessments and risk mitigation.[362] It requires recordkeeping, with automatic logs for high-risk AI systems.[363] It requires providers to design a plan for postmarket monitoring for incidents caused by high-risk AI systems[364] and to report serious incidents.[365]

Viewed through a comparative lens with regard to the regulations discussed above, however, the Act has significant failings. The Act divides regulated entities into providers, who build AI systems, and users, who use them. As a consequence, *nobody is really responsible*

[358] *Draft E.U. AI Act*, *supra* note 1, arts. 41–42.

[359] *Id.* art. 43. Unlike the licensing systems in the regulations above, however, this conformity assessment is self-administered and self-certified. *Id.* at 14 ("A comprehensive ex-ante conformity assessment through internal checks . . . .").

[360] *Id.* art. 14.

[361] *Id.* recital 48.

[362] *Id.* art. 9.

[363] *Id.* art. 12.

[364] *Id.* art. 61.

[365] *Id.* art. 62.

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.[366] But "[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."[367] 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 overreliance, 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 frailties 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."[368] The Act's requirements are broadly functional rather than detailed in nature, which wouldn't be an issue if the Act corrected for problematic incentives. For example, the Act requires that AI tools 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.[369] But 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.[370] By contrast, the safety-critical systems discussed above either involve direct regulation or an ex ante licensing scheme by the relevant authority that can refuse to license a particular product.

Finally, a core challenge for the draft AI Act—and really, for any law that attempts to use human oversight to protect fundamental

[366] *Id.* art. 29.

[367] Veale & Zuiderveen Borgesius, *supra* note 65, at 104.

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

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

[370] Notably, high-risk AI systems that implicate product safety instead would follow an existing system of third-party conformity assessments. *Id.* art. 14, § 5.2.3.

human rights—is how to validate and verify that the human in the loop is accomplishing the desired goals.[371] Given how vague the goals are, that may be nearly impossible. How is the impact on rights measured? What kind of expertise would humans need to have to be effective in their role in the loop? How could deferring to a single person in the loop's take on contested concepts such as "fairness" and "discrimination" make that process legitimate, when additional voices might define "fairness" or "discrimination" differently, with meaningfully different results? 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? Hopefully, these issues will be addressed in future drafts, before the law is finalized.

2. Regulating Better: Facial Recognition Systems

Facial recognition systems ("FRS") are software systems that typically aim to use data analytics to match a recorded instance of a face to a person's identity. Among other industry, administrative, and governmental uses, FRS are used by law enforcement for identity verification and matching. Many people have expressed deep concerns with FRS use, calling for regulation or even a ban.[372] These fears include "unprecedented threats to civil liberties, free expression, privacy, human rights, and democratic accountability."[373] There is also reason to worry about its accuracy. In an era of facial masking, FRS error rates can range from five to fifty percent.[374] And that inaccuracy may be discriminatory with respect to women and people of color, who are disproportionately inaccurately identified by FRS.[375] Thus, concerns

[371] *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.

[372] WOODROW HARTZOG & EVAN SELINGER, JUST. COLLAB. INST., THE CASE FOR BANNING LAW ENFORCEMENT FROM USING FACIAL RECOGNITION TECHNOLOGY (2020), https://theappeal.org/the-lab/report/the-case-for-banning-law-enforcement-from-using-facial-recognition-technology/ [perma.cc/VXS2-V8WN].

[373] *Id.* at 2.

[374] Nat'l Inst. Standards & Tech., U.S. Dep't of Com., *NIST Launches Studies into Masks' Effect on Face Recognition Software*, NEWSNIST, https://www.nist.gov/news-events/news/2020/07/nist-launches-investigation-face-masks-effect-face-recognition-software (last updated Aug. 4, 2020) [perma.cc/A4KC-D64H].

[375] Hartzog & Selinger, *supra* note 372, at 4.

about inaccuracy in facial recognition are also concerns about bias and discrimination.[376]

What if a regulator wanted to address these issues by adding a human in the loop for law enforcement use of facial recognition?[377] (To be clear: we do not argue that a human in the loop is the right approach to regulating the use of facial recognition systems, nor that it should be the only approach. For purposes of our case study, we take as given that a regulator wants to make this particular intervention.)

Specifically, what if a regulator was concerned about the possibility of false arrests made on the basis of incorrect suspect identifications? Our process would guide the regulator to do the following. First, identify the role the human should play. Second, look to context, including what law already governs this space. Third, design regulation to ensure the hybrid system is built according to best practices from engineering.

A regulator might identify several reasons for wanting a human in the loop of law enforcement use of facial recognition systems. The most likely reason is corrective: to prevent errors in general (error correction) and to protect women and people of color from the discriminatory effects of heightened error rates (bias correction). Relatedly, the human might serve a resilience function, able to operate as a failsafe mode when the system malfunctions. A human in the loop might serve a dignitary function, to put a human touch on what is otherwise a decision by a machine. The human might also serve an accountability role, should errors manifest in harms, and might serve (more cynically) to insulate the system from due process challenges. Alternatively, the human may play a friction role, slowing down the use of FRS and thus lessening the speed and capacity of law enforcement and government surveillance.

Regulators will not, however, be writing on a blank slate; they will need to consider the specific requirements of the law enforcement context. For example, certain significant government decisions trigger due process protections. What might a due process analysis require of a human-in-the-loop system?[378] Law enforcement use of surveillance technologies is analyzed under the Fourth Amendment, and recent Supreme Court decisions suggest that law enforcement use of FRS for

[376] *See* Amanda Levendowski, *Resisting Face Surveillance with Copyright Law*, 100 N.C. L. REV. 1015 (2022).

[377] As discussed above, Colorado has recently done this for decisions having significant effects. *See supra* note 71 and accompanying text.

[378] *See* Huq, *Constitutional Rights in the Machine Learning Stat*e, *supra* note 15, at 1880 (discussing due process challenges to government use of algorithms).

location tracking over time would require a warrant.[379] Would the placement of a human in the loop change this analysis? What if law enforcement neglects to get a warrant but still places a human in the loop?

The ongoing surveillance of Black communities also raises significant concerns about law enforcement power and (un)accountability, which should also be considered in a contextual analysis.[380] Policy issues with FRS aren't limited to concerns about accuracy of the system; they go to how much power FRS surveillance systems afford to already-powerful and unaccountable entities.[381] Would placing a human in the decisional loop of FRS in any way address these fears? Probably not. A human in the loop would not meaningfully increase police accountability or address systemic racism. To the contrary: it could plausibly be framed as "ethics washing" by regulators.

This leaves our last step. Should regulators believe that a human in the loop still serves a productive role, they must regulate to ensure that the human succeeds in that role. Let's say the main role is corrective: regulators want a human to catch and correct individual errors and systemic biases. First, policymakers might require the government to procure only FRS that meet some minimum accuracy standard and are designed and tested for successful human interventions.[382] Second, regulators might mandate a training process for humans in the loop or determine who they are and how much independence to give them. For example, regulators might determine whether humans in the loop must be employees or might be external contractors, as well as how much job security they should have, even if they engage in activities like whistleblowing. Third, regulators should set up a means for the human in the loop to review the system's performance, which might include mandating reporting, testing, or auditing, both before and during system use.

[379] *See, e.g.*, Margot E. Kaminski, Carpenter v. United States*: Big Data Is Different*, GEO. WASH. L. REV.: ON THE DOCKET (July 2, 2018), https://www.gwlr.org/carpenter-v-united-states-big-data-is-different/ [perma.cc/NL3Z-XZVV].

[380] Chaz Arnett, *Race, Surveillance, Resistance*, 81 OHIO ST. L.J. 1103, 1110–11 (2020); *see also* Isaiah Strong, Note, *Surveillance of Black Lives as Injury-In-Fact*, 122 COLUM. L. REV. 1019 (2022); Sidney Fussel, *How Surveillance Has Always Reinforced Racism*, WIRED (June 19, 2020, 7:00 AM), https://www.wired.com/story/how-surveillance-reinforced-racism/ [perma.cc/B28S-TABE].

[381] *See, e.g.*, Frank Pasquale, *The Second Wave of Algorithmic Accountability*, LPE PROJECT (Nov. 25, 2019), https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/ [perma.cc/CT34-D88T].

[382] Deirdre K. Mulligan & Kenneth A. Bamberger, *Procurement as Policy: Administrative Process for Machine Learning*, 34 BERKELEY TECH. L.J. 773 (2019).

* * *

There are a host of useful lessons to be gleaned from prior regulation of safety-critical systems. Unfortunately, regulators of modern human-in-the-loop systems do not seem to have adopted these principles. As described above, much of the law of the loop is comprised of older, generally applicable rules.[383] Other rules put a human in the loop without considering these systemic concerns.[384] Still others play lip service to hybrid challenges but miss important aspects.[385]

Human involvement 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.

E. Complements and Alternatives to a Human in the Loop

Successfully placing a human in the loop may not be worth the effort necessary to effectively regulate the system.[386] As the examples above demonstrate, including a human may require complex, detailed, even heavy-handed regulation. This kind of regulation—its costs, its interventionism—won't always be appropriate. And many, including us, have argued that algorithmic systems should *not* have such rigid regulation but are often better governed by more flexible, dynamic rules.[387] The goals regulators want to serve by placing a human in the loop may sometimes 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.[388] These approaches can either replace or complement the human-in-the-loop approach we discuss here.[389] For example, corrective goals may be achieved through ex ante interventions on a systemic level, such as conducting risk assessments or setting design goals, and ex post measures, such as articulating performance metrics and conducting

[383] *See supra* Part II.

[384] *See supra* Part III.

[385] *See supra* Section V.D.1.

[386] *Cf.* Green, *supra* note 8, at 13–14 (arguing that policymakers should always first consider whether it is actually appropriate to use an algorithm in a specific context).

[387] *See, e.g.*, Crootof, *Killer Robots Are Here*, *supra* note 16; Kaminski, *Binary Governance*, *supra* note 13; Price, *Regulating Black-Box Medicine, supra* note 17; *see also* Crootof & Ard, *Structuring Techlaw*, *supra* note 23, at 399–413 (discussing the trade-offs between more- and less-flexible legal designs and language).

[388] Crootof, *Cyborg Justice*, *supra* note 15; Crootof, *Killer Robots Are Here*, *supra* note 16; Crootof, *War Torts*, *supra* note 13; Kaminski & Urban, *supra* note 138; Price, *Regulating Black-Box Medicine*, *supra* note 17.

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

audits.[390] Justificatory 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 requiring the disclosure of such reasons to impacted individuals.[391] Both justificatory and accountability goals can be served by incorporating the voices of impacted stakeholders early in the design process[392] and forcing industry actors to internalize the costs of employing harmful technologies.[393] Meanwhile, dignitary goals are furthered by establishing individual rights to contest certain automated decisions.[394]

Humans in the loop[395] are an attractive option for regulators, and we have suggested ways to ameliorate some of the problems hybrid systems raise. But regulators still have other tools in their toolbox, which sometimes will work better.

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.*[396]

Humans in the loop can play important roles in algorithmic decisionmaking systems. But humans aren't 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, but they need to be situated and enabled to succeed in those roles. That requires knowing what those roles are and governing and considering the systems as a whole.

[390] Kaminski, *Binary Governance*, *supra* note 13, at 1595; Margot E. Kaminski & 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 17; Price, *supra* note 13.

[391] Hildebrandt, *supra* note 231, at 114–15.

[392] *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, 64 (2017) (arguing that "public discourse and input" can improve the construction of risk assessment tools); Kaminski, *Binary Governance*, *supra* note 13, at 1533–34; Margot E. Kaminski & Gianclaudio Malgieri, *Participatory AI* (Jan. 11, 2023) (unpublished manuscript) (on file with authors).

[393] Crootof, *War Torts*, *supra* note 13; Rebecca Crootof, *AI and the Actual IHL Accountability Gap*, CIGI (Nov. 28, 2022), https://www.cigionline.org/articles/ai-and-the-actual-ihl-accountability-gap/ [https://perma.cc/2J52-XE9L].

[394] Kaminski & Urban, *supra* note 37.

[395] *See supra* note 34 (illustrating the experience of being a human in a loop).

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

As we each have discussed elsewhere, it may also involve designing governance to make room for changes over time.[397]

The question of how to regulate human-in-the-loop systems will remain a constant for the foreseeable future, but the right answers will vary dramatically depending on the roles humans are meant to play and the context of the particular systems. While recognizing there is much more to be done, we offer guidance for policymakers working to improve the governance of algorithmic systems going forward.

[397] Crootof & Ard, *Structuring TechLaw*, *supra* note 23, at 388, 399; Kaminski, *Binary Governance*, *supra* note 13, at 1560–62; Price, *Regulating Black-Box Medicine*, *supra* note 17, at 459–60.