Fumble! Anti-Human Bias in the Wake of Socio-Technical System Failures

Joseph Avery

In 2018, an autonomous Uber vehicle hit and killed a pedestrian. Although autonomous, the vehicle was not driverless: the onboard artificial intelligence (AI) had handed off control to a safety driver at the last second. "Handed off"—a just-launched National Highway Traffic Safety Administration investigation into Tesla is focusing on this very issue. After all, like many socio-technical systems, semi- and fully autonomous vehicles are designed so that decision-making can be handed off from a machine to a human operator. This is worrisome because decades of human factors research have shown that people perform poorly under such conditions. When situated in the handoff recipient role, humans suffer from a combination of four linked issues: complacency, inattention, skill atrophy, and automation bias (i.e., over-trust in the machine).

There is a second issue of importance here, and it concerns proximate cause. A significant body of experimental jurisprudence research has revealed a nexus between moral conclusions and causal ones. If we perceive someone as morally responsible for an outcome, then we are more likely to perceive that person as factually responsible for the outcome. By extension, and in dialogue with longstanding debates between legal realists and formalists, proximate cause conclusions are dependent upon perceptions of entities' moral culpability in addition to assessments of those entities' causal responsibility. Socio-technical systems are unique in that their failures result from the collective action of some actors with moral capacity (i.e., humans) and some that lack it (i.e., machines), an imbalance that stands to complicate and impact proximate cause conclusions.

In this article, I garner these two issues to address questions that, in spite of their importance to current and emerging issues in the law of artificial intelligence, have gone heretofore unstudied. How do jurors attribute fault when a socio-technical system leads to harm, when a human-machine collaboration ends in a bad outcome? What is the psychology undergirding these fault attributions, and might there be systematic biases in them?

To fill this lacuna, I conducted two sets of original experimental studies on lay imputations of fault in post-handoff harmful events. In one experiment, the harm was the result of a semi-autonomous car accident. In

the other experiment, the harm was the result of a racially biased algorithm-assisted bail decision. I hypothesized that, in the wake of socio-technical system failures, human operators would be systematically over-faulted, and it would be to the extent that they were perceived as moral actors. That was precisely what happened.

The findings revealed systematic bias, which I am calling "fumble bias": when socio-technical systems feature machine-to-human handoffs, the human operator (i.e., the handoff-recipient) receives the bulk of the fault, even in scenarios in which the operator's performance is so disadvantaged that it is likely incapable of meeting any relevant standard-of-care. In short, such systems set the human operator up for failure—and jury-eligible individuals are more likely to place fault on the human than on the machine, its developers, or the company that is responsible for it. The article's studies showed that fumble bias appears regardless of which term in the fault lexicon is used: fault, proximate cause, factual cause, blame, legal liability, norm violation, moral failure, and so on. Moreover, the studies showed that this effect is not owing to handoffs in general: when a human made a handoff to another human, fault was equally distributed across the two entities. Finally, and most importantly, there was a significant correlation between the perceived moral capacity of the entity and the extent to which that entity was considered at fault.

This article fills in a large knowledge gap by revealing how fault is imputed when socio-technical systems lead to harm. By identifying the effect and a potential mechanism precipitating it, the article provides much-needed information for litigators, legislators, and scholars, and it shows that technology companies might have the upper-hand in the initial wave of these cases that makes it to the courts. In addition, the results provide key data in the rapidly developing field of experimental jurisprudence, especially for a line of research that focuses on proximate cause. A nexus between moral and causal conclusions was relatively well-established, but no one had heretofore explored what happens to this nexus when one entity, such as a machine actor, lacks moral capacity. With the rise of semi-autonomous vehicles and algorithm-assisted professional decision-making, and with the waxing phenomenon of "ubiquitous computing," the reach and impact of these findings is significant.

Having empirically identified potential doctrinal and cognitive entry points for understanding liability imputations following socio-technical system failures, I end by suggesting some steps—such as jury instructions and rules of thumb for regulators—that the legal system could consider taking in light of the identified bias. Indeed, through a final study presented in this article, I show that education about handoffs might partially correct for

fumble bias, although more research along this line is needed. What is clear is that this baseline is necessary for development of guides that will move jurors back onto the path of proper understanding and application of the law. Relatedly, legal scholars have presented countless plans for how best to regulate and legislate semi- and fully autonomous vehicles, and more plans regarding other socio-technical systems are emerging as well. These scholars will be successful in promoting their solutions, and legislators in getting those solutions adopted, only if they are attuned to biases and intuitions that may cut against (or in favor of) their desired solutions.

INTRODUCTION

In 2018, an autonomous Uber vehicle hit and killed a pedestrian. Although autonomous, the vehicle was not driverless: a safety driver had been tasked with oversight and was to monitor the artificial intelligence's ("AI") decision-making and even take over should an emergency arise.[1] In the months following the accident, video of the final, fatal seconds was released. In the video, the safety driver is looking down before her gaze suddenly shoots up; she gasps and is wide-eyed and clearly in shock at what is happening beyond the frame.[2] Was she to blame? Countless commentators across the internet weighed in.[3] An official investigation was conducted and a subsequent report published. The safety driver was indicted by a grand jury on a count of negligent homicide, and the matter is still pending as of this writing.[4]

This machine-to-human ("m2h") handoff failure is not without precedent. In 2009, Air France Flight 447 was en route to Paris from Rio de Janeiro when a faulty mechanical device, which had been documented and was being replaced in other Air France aircraft, caused the autopilot to disconnect.[5] A

[1] Kate Conger, *Driver Charged in Uber's Fatal 2018 Autonomous Car Crash*, N.Y. TIMES (Dec. 7, 2020), https://www.nytimes.com/2020/09/15/technology/uber-autonomous-crash-driver-charged.html [https://perma.cc/TN28-U6P3].

[2] Ray Stern, *Uber Backup Driver Indicted in 2018 Self-Driving Crash That Killed Woman*, PHX. NEW TIMES (Sept. 15, 2020), https://www.phoenixnewtimes.com/news/uber-backup-driver-in-phoenix-indicted-over-fatal-self-driving-car-crash-in-18-11494111 [https://perma.cc/LD62-Q4EH].

[3] *See, e.g.*, Sam Levin, *Video Released of Uber Self-Driving Crash That Killed Woman in Arizona*, GUARDIAN (Mar. 21, 2018), https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona [https://perma.cc/TU76-E9VY].

[4] Stern, *supra* note 2.

[5] BUREAU D'ENQUÊTES ET D'ANALYSES, FINAL REPORT ON THE ACCIDENT ON 1ST JUNE 2009 TO THE AIRBUS A330-203 REGISTERED F-GZCP OPERATED BY AIR FRANCE FLIGHT AF 447 RIO DE JANEIRO – PARIS, at 123–24 (2012), https://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf [https://perma.cc/G46K-7ADH].

series of fatal pilot errors followed.[6] The plane stalled and did not recover: it plummeted into the Atlantic Ocean, killing all 228 people aboard.[7] As Madeleine Clare Elish notes in a 2019 article, blame was heaped upon the pilots, and the mechanical error was largely treated as an afterthought.[8] This was true both in lay opinions during the period following the accident—it took nearly two years to recover the flight recorders from the ocean floor—and in the official accident report.[9]

It may surprise readers to learn that I leaned heavily on machine assistance as I drafted this article. I used word processing software. This was not quite a collaborative enterprise, as I actively monitored the word processing software to make sure that it recorded the precise letters, symbols, and spaces that I instructed it to record. This was a less collaborative undertaking than, say, using Gmail to draft an email, as Gmail incorporates word and sentence completion, which takes steps in the direction of a human-machine collaborative dyad. As one works up the human-machine collaborative ladder, moving further and further into intertwined labor, one encounters many emerging technologies. Most legal associates, pressed to conduct tiresome and tedious document review, do such labor in coordination with e-discovery tools.[10] In this distributed system, the machine performs the initial review, and the associate is called upon to consider only those documents that have already been flagged by the software.[11] As a second example, one might think of surgery robots, such as the da Vinci Surgical Robot, which conducts most of actions necessary for a specified surgical procedure, the surgeon taking over at only a few junctures.[12] If one continues up this ladder, the furthest reaches of such systems manifest as brain-machine interfaces, where computational devices are embedded within a human's brain, and the collaborative dyad becomes something resembling a unity.[13]

[6] *Final Air France Crash Report Says Pilots Failed To React Swiftly*, CNN (July 5, 2012), http://www.cnn.com/2012/07/05/world/europe/france-air-crash-report/index.html [https://perma.cc/2FK6-M4JR].

[7] BUREAU D'ENQUÊTES ET D'ANALYSES, *supra* note 5, at 21–24.

[8] Madeleine Clare Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAGING SCI., TECH., & SOC'Y 40, 48 (2019).

[9] *Id.* at 48–50.

[10] John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, N.Y. TIMES (Mar. 4, 2011), https://www.nytimes.com/2011/03/05/science/05legal.html [https://perma.cc/6PJU-WXL6].

[11] *See id.*

[12] Brian Flood, *Surgical Robot Manufacturer Escapes Personal Injury Lawsuit*, BLOOMBERG L. (Apr. 6, 2021), https://news.bloomberglaw.com/us-law-week/surgical-robot-manufacturer-escapes-personal-injury-lawsuit [https://perma.cc/DQK9-LJ85].

[13] THE ROYAL SOC'Y, IHUMAN: BLURRING LINES BETWEEN MIND AND MACHINE 28–34 (Sept. 2019), https://royalsociety.org/-/media/policy/projects/ihuman/report-neural-interfaces.pdf [https://perma.cc/8WC6-JHNU].

In nearly all collaborative endeavors, be they joint human enterprises or the topic of this article, joint human-machine enterprises, handoffs are a key feature. For example, in healthcare, a patient might present to the emergency department where she is treated by a nurse and an on-call physician. The treatment might necessitate admission and an overnight stay. The patient sleeps. When she awakes, a different nurse and a different doctor stand over her, and what they know about her case has been communicated to them via the previous nurse and doctor. Much litigation in medical malpractice and related lawsuits stems from such handoffs. Did the initial treating doctor incorrectly dictate an otherwise correct diagnosis and prescription? Did the pharmacy fumble the handoff and produce the wrong medication—or the correct medication at an incorrect dosage? Such types of questions regarding handoff-related responsibility and liability certainly apply to the above-described motor vehicle and airplane accidents. Nearly every socio-technical system is designed so that decision-making can be shifted—handed off—from the machine to a human operator.

Despite the increasing ubiquity of socio-technical system failures, key questions, ones foundational to current and emerging issues in the law, have gone empirically unaddressed in the legal academy to date. How do we attribute fault when a human-machine collaboration ends in a bad outcome? What is the psychology undergirding these fault attributions and might there be systematic biases in them? More specifically, the fundamental question of how lay decisionmakers without legal training—more specifically, jury-eligible individuals—are likely to attribute fault in a m2h post-handoff harmful event has never been empirically tested. To fill this lacuna, I conducted two sets of original experimental studies on lay imputations of fault in post-handoff harmful events. In one experiment, the harm was the result of a semi-autonomous motor vehicle accident. In the other experiment, the harm was the result of a racially-biased algorithm-assisted bail decision.

At the root of this research is an old bugbear of legal scholarship: proximate cause. Proximate cause, fault, legal responsibility, and the related constellation of concepts have received renewed interest in legal scholarship.[14] Some of this focus can be attributed to developments in the

[14] *See, e.g.*, Peter Bach-y-Rita, *The Causal Mechanism Theory of Legal Causation*, 34 RATIO JURIS. 57 (2021) (exploring proximate cause in terms of probability of harm theories); Mark A. Geistfeld, *Proximate Cause Untangled*, 80 MD. L. REV. 420 (2021) (discussing proximate cause in the context of the directness and foreseeability tests); Eric A. Johnson, *Dividing Risks: Toward a Determinate Test of Proximate Cause*, 2021 U. ILL. L. REV. 925 (2021) (exploring the application of a framework of increasing risks); Sandra F. Sperino, *The Emerging Statutory Proximate Cause Doctrine*, 99 NEB. L. REV. 285 (2020) (making the case for a statutory proximate cause doctrine).

behavioral sciences.[15] Some of it is the result of advances in neuroscience, which are slowly shifting our understanding of human agency and will.[16] Some is the result of a durative interest in the peculiar cases of drug use and drug addiction and what they mean for responsibility and liability.[17] But the bulk of the scholarly handwringing has been precipitated by technology.[18] More specifically, it has been precipitated by an increasing number of actionable incidents that arise from the conduct of not one person, not two people, not even a group of people, but rather from agency that is distributed across both humans and machines: the complex, semi-automated socio-technical systems such as those described above,[19] especially ones that involve intelligent assistance, artificial intelligence, or robotics.

This interest in liability pertaining to socio-technical systems is warranted, as such systems are increasingly in use, especially in the legal domain. Police departments have been using predictive technologies[20] and facial recognition tools,[21] while prosecutors and the courts have been using risk assessment

[15] Joshua Knobe & Scott J. Shapiro, *Proximate Cause Explained: An Essay in Experimental Jurisprudence*, 88 U. CHI. L. REV. 165, 197–99 (2021).

[16] Joshua Greene & Jonathan D. Cohen, *For the Law, Neuroscience Changes Nothing and Everything*, 359 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON B 1775 (2004).

[17] *See, e.g.*, Kristen S. Jones, *The Opioid Epidemic: Product Liability or One Hell of a Nuisance?*, 39 MISS. COLL. L. REV. 32 (2021) (discussing proximate cause in the context of the opioid epidemic).

[18] Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 VA. L. REV. 127 (2019) (discussing the issue in the context of semi- and fully autonomous vehicles); Frank Griffin, *Artificial Intelligence and Liability in Health Care*, 31 HEALTH MATRIX 65 (2021) (exploring the liability implications of AI in healthcare); Amy L. Landers, *Proximate Cause and Patent Law*, 25 B.U. J. SCI. & TECH. L. 329 (2019) (arguing for an expansion of proximate cause's role in patent litigation).

[19] Gordon Baxter & Ian Sommerville, *Socio-Technical Systems: From Design Methods to Systems Engineering*, 23 INTERACTING WITH COMPUTS. 4 (2011).

[20] *See, e.g.*, Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 113 (2017) (discussing how predictive policing has been increasingly used by police departments); Letter from Jonathan J. Wroblewski, Dir., Off. of Pol'y & Legis., to Hon. Patti B. Saris, Chair, U.S. Sent'g Comm'n (July 29, 2014), https://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf [https://perma.cc/RZ4S-WGTS] (discussing how predictive policing was becoming ubiquitous in police departments).

[21] Bernard Keenan, *Automatic Facial Recognition and the Intensification of Police Surveillance*, 84 MOD. L. REV. 886 (2021).

algorithms pre-trial and various tools at sentencing.[22] Insurance companies,[23] plaintiffs' attorneys,[24] and many others involved in litigation[25] have been leaning heavily on technology.

But what about those instances in which machines make mistakes and people are harmed? The scholarship that explores this has largely focused on generating prescriptive arguments for how liability ought to be attributed and distributed.[26] This scholarship is excellent and necessary, but I believe an important piece is missing. We lack the descriptive: we do not have insight into how fault is likely to be imposed in litigation stemming from socio-technical system failures. Will there be systematic bias that benefits technology companies, or will technology companies be unfairly burdened? Without a baseline understanding of how and to what extent jurors deviate from proper understanding and application of the law, it is nearly impossible to develop instructions that will guide jurors back onto the proper path. Moreover, legal scholars have presented countless plans for how best to regulate and legislate semi- and fully autonomous vehicles, but scholars will be successful in promoting their solutions, and legislators in getting those solutions adopted, only if they are attuned to biases and intuitions that may cut against (or in favor of) their desired solutions. Finally, the descriptive is needed because it furthers exploration of deep issues regarding proximate cause, fault, legal liability, and related concepts. This is an exploration that very much falls within the experimental jurisprudence branch, and its results will help us to better understand how these concepts are used in the U.S. legal system.

In this article, I focus on the space between exclusive human error and exclusive mechanical failure. Indeed, most socio-technical system failures

[22] *See, e.g.*, Andrew Guthrie Ferguson, *Predictive Prosecution*, 51 WAKE FOREST L. REV. 705, 705–08 (2016) (providing an overview of predictive prosecution); Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT'G REP. 237, 237 (2015) (discussing the nexus of race and risk assessment).

[23] MATTHEW J. DEGAETANO & JAMES MATHIS, COLOSSUS AND OTHER INSURANCE ADJUSTMENT SOFTWARE: MAXIMIZING AUTO ACCIDENT CLAIM SETTLEMENT OFFERS (2017), http://media.straffordpub.com/products/colossus-and-other-insurance-adjustment-software-maximizing-auto-accident-claim-settlement-offers-2017-10-17/reference-materials.pdf [https://perma.cc/3AP8-FJKQ].

[24] See, for example, CLAUDIUS LEGAL INTELLIGENCE, https://www.claudius.ai/ [https://perma.cc/JH2H-V6TT], for a description of one company's artificial legal intelligence software.

[25] See, for example, BLUE J LEGAL, https://www.bluej.com/ [https://perma.cc/6GMZ-FL9T], for a description of another company's predictive tax software for use in lawsuits and governmental investigations.

[26] *See, e.g.*, Steven Shavell, *On the Redesign of Accident Liability for the World of Autonomous Vehicles*, 49 J. LEGAL STUDS. 243 (2020) (arguing for strict liability for drivers of autonomous vehicles with payment made to the state).

are the result of a confluence of factors: just as it takes multiple entities for such systems to work, breakdowns in such systems are multiply determined. In the above-described Air France crash there was a cascade of machine and human errors.[27] In her 1996 article, which was decades ahead of the curve, Helen Nissenbaum identified a "problem of many hands" when attributing fault to a technological system.[28] While this creates a knotty problem, we already have a fair amount of guidance as to how we might apportion responsibility in a situation of "many hands," as it takes many hands to operate an overnight train, and it takes many hands to treat a patient who presents to an emergency department, and so on. We have numerous doctrines on which to draw—some admittedly less apt than others—including those pertaining to *respondeat superior*,[29] drug use,[30] the parent-child relationship,[31] the owner-animal relationship,[32] and, as mentioned, products liability.[33]

But there is novelty that emerges in the special case of socio-technical system failures. To understand this novelty, we first must understand the issues of proximate cause mentioned fleetingly above. A central debate in legal scholarship revolves around whether the legal concept of proximate cause[34] is equivalent to the layfolk concept of causation.[35] There are two opposing camps. In the formalist camp, it is argued that proximate cause is an objective matter that can be determined by descriptive inquiry.[36] In the realist camp, it is argued that proximate cause can be reduced to "responsible

[27] BUREAU D'ENQUÊTES ET D'ANALYSES, *supra* note 5, at 200.

[28] Helen Nissenbaum, *Accountability in a Computerized Society*, 2 SCI. & ENG'G ETHICS 25, 28–32 (1996).

[29] Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043, 1084–1100 (2020).

[30] Dawinder S. Sidhu, *Criminal Law x Addiction*, 99 N.C. L. REV. 1083 (2021) (discussing how addiction interacts with control and volition in criminal law).

[31] Charles A. Marvin, *Discerning the Parent's Liability for the Harm Inflicted by a Nondiscerning Child*, 44 LA. L. REV. 1213 (1984).

[32] KATE DARLING, THE NEW BREED: WHAT OUR HISTORY WITH ANIMALS REVEALS ABOUT OUR FUTURE WITH ROBOTS 60–86 (2021).

[33] K.C. Webb, *Products Liability and Autonomous Vehicles: Who's Driving Whom*, 23 RICH. J.L. & TECH. 1 (2017).

[34] The definition of proximate cause has been debated for well over a century. *See* Henry T. Terry, *Proximate Consequences in the Law of Torts*, 28 HARV. L. REV. 10 (1914); Charles E. Carpenter, *Workable Rules for Determining Proximate Cause*, 20 CALIF. L. REV. 229 (1932); Joseph H. Beale, *The Proximate Consequences of an Act*, 33 HARV. L. REV. 633 (1920).

[35] *See, e.g.*, Clarence Morris, *On the Teaching of Legal Cause*, 39 COLUM. L. REV. 1087, 1088 (1939) (discussing the debate and the deep confusion that attached to it); Richard W. Wright, *Causation in Tort Law*, 73 CALIF. L. REV. 1735, 1737 (1985) ("In all of tort law, there is no concept which has been as pervasive and yet elusive as the causation requirement . . . .").

[36] Beale, *supra* note 34, at 643–44.

cause."[37] As Professor Green wrote, "[T]he inquiry while stated in what seems to be terms of *cause* is in fact whether the defendant should be held responsible."[38] In their 2021 article, Professors Knobe and Shapiro take a middle approach.[39] They argue that an initial moral judgment (was the defendant's action morally improper?) influences judgment about proximate cause (did the defendant cause the harm?), which in turn influences judgment about responsibility (should the state hold the defendant liable for the harm caused?).[40]

Along this line, there has been much research showing that our perceptions of causation are influenced by our beliefs as to whether someone has done something wrong.[41] In essence, to the extent that a causal factor is viewed as morally wrong, that factor also will be viewed as having caused the bad outcome.[42] In the case of m2h handoffs then, we have a strange element. Machines are generally *not* considered morally capable; thus, no matter how terrible the result of a machine decision, the behavior likely will not be deemed morally wrong. Given that such judgments impact causal judgments, machines (and their creators, which I discuss in greater detail below) might be let off the hook when m2h handoffs lead to harm. After all, in the wake of an accident, the machine did not do anything morally wrong; it was just being a machine, even if it was being a suboptimal or even a broken machine. But the behavior of the human operator might be considered morally wrong. The natural hypothesis, then, is that in the wake of socio-technical system failures, human operators—unified moral actors who are close the scene of the harm—will be systematically over-faulted to the extent that they are perceived as moral actors.

But is it so wrong to shift blame onto the human recipient of a m2h handoff? Saving policy discussions for the main text of this article, it is worth noting here that human factors research overwhelmingly concludes that

[37] W. PAGE KEETON ET AL., PROSSER AND KEETON ON THE LAW OF TORTS § 42, at 273 (Dan B. Dobbs et al. eds., 5th ed. 1984).

[38] Leon Green, *Are There Dependable Rules of Causation?*, 77 U. PA. L. REV. 601, 605 (1929).

[39] Knobe & Shapiro, *supra* note 15, at 171.

[40] *Id.*

[41] *See, e.g.*, Mark D. Alicke, *Culpable Causation*, 63 J. PERSONALITY & SOC. PSYCH. 368, 376 (1992) (covering the effect of moral conclusions on causal ones); *see also* Judith Jarvis Thomson, *Causation: Omissions*, 66 PHIL. & PHENOMENOLOGICAL RSCH. 81, 99 (2003) (exploring the topic through a philosophical approach to omissions); Sarah McGrath, *Causation by Omission: A Dilemma*, 123 PHIL. STUD. 125, 132–48 (2005) (also exploring the topic through omissions).

[42] Alicke, *supra* note 41, at 376; Mark D. Alicke, *Culpable Control and the Psychology of Blame*, 126 PSYCH. BULL. 556, 558 (2000).

handoff recipients face nearly insurmountable disadvantage.[43] This disadvantage stems from a combination of four linked issues: automation complacency, inattention, skill atrophy, and automation bias (i.e., over-trust).[44] According to Professor John Leonard at the Massachusetts Institute of Technology, the problem "is unsolvable. The notion that a human can be a reliable backup is a fallacy."[45] This behavior is psychologically reasonable. People are generally efficient actors, taking the route that requires the least cognitive effort.[46] When paired with a machine, the impulse is to offload more and more of one's cognitive load onto the machine.[47]

It is axiomatic that laws should regulate behavior, not entities.[48] In the handoff situations discussed in this article, it is behavior that is the problem. Something has gone wrong. Someone has been hurt. Someone's property has been damaged. Someone's liberty has been taken. The question is, how do we attribute liability in order to correct the behavior? In m2h handoff failures, it makes little sense to place fault on the disadvantaged and meaningfully handicapped human operator.

But the developers of technology are vigorously pushing in the opposite direction, and legislators have been siding with the developers.[49] When it comes to aircraft, the Federal Aviation Administration has specifically addressed the problem in regulation: "The pilot in command of an aircraft is directly responsible for, and is the final authority as to, the operation of that aircraft."[50] For decades courts have consistently ruled in accordance with this regulation, such that the pilot, regardless of how advanced autopilot becomes, is the ultimate bearer of liability.[51] Those in the rapidly emerging autonomous vehicle domain have diligently argued that drivers of semi-autonomous vehicles should be treated likewise: as Elon Musk said of Tesla vehicles,

[43] *See infra* Parts I, E.

[44] MARÍA ALONSO RAPOSO ET AL., JOINT RSCH CTR., EUR. COMM'N, THE R-EVOLUTION OF DRIVING: FROM CONNECTED VEHICLES TO COORDINATED AUTOMATED ROAD TRANSPORT 45–56 (2017).

[45] John Markoff, *Robot Cars Can't Count on Us in an Emergency*, N.Y. TIMES (June 7, 2017), https://www.nytimes.com/2017/06/07/technology/google-self-driving-cars-handoff-problem.html [https://perma.cc/G8LW-7HVB].

[46] Enrico Coiera, *Technology, Cognition and Error*, 24 BMJ QUALITY & SAFETY 417, 417–22 (2015).

[47] RAPOSO ET AL., *supra* note 44.

[48] Bryan Casey & Mark A. Lemley, *You Might Be a Robot*, 105 CORNELL L. REV. 287, 288 (2020).

[49] *See* 14 C.F.R. § 91.3(a) (2021).

[50] *Id.*

[51] James E. Cooling & Paul V. Herbers, *Considerations in Autopilot Litigation*, 48 J. AIR L. & COM. 693, 708–11 (1982).

> It's almost to the point where you can take your hands off [. . .] but we're very clearly saying this is not a case of abdicating responsibility . . . . The hardware and software are not yet at the point where a driver can abdicate responsibility . . . . [The system] requires drivers to remain engaged and aware when Autosteer is enabled.[52]

Which brings us to the need for the present article. Beneath the prescriptive and regulatory posturing, there are currents of psychological intuition regarding fault, proximate cause,[53] and legal liability. Via empirical research, this article begins the task of observing, documenting, and understanding these currents. Exemplars of the two types of m2h handoffs are presented, and the contours of fault attributions in them are explored. This is done in the light of perceived moral capacity, and potential solutions to the documented bias are unpacked. As a result, this article yields the following contributions.

First, it shows that, in m2h handoff failures, there is a systematic bias towards placing fault for the bad outcome on the nearest human operator, even when the machine that made the handoff is shown to have significantly, if not entirely, caused the outcome. Importantly, this effect is not related to the nature of handoffs in general, as equivalent human-to-human handoffs do not lead to equal attribution of fault across the handoff-maker and handoff-recipient. In other words, there is something about m2h handoffs that precipitates the systematic bias. This leads to the second contribution, such that the article underscores the importance of moral inferences for proximate cause. Because moral blame and proximate cause operate in a feedback loop, when there are two actors and one lacks moral capacity, blame will be shifted onto the more morally capable actor. Third, I show that these results are important both for applied and academic reasons. On the applied end, the article should precipitate a rethinking of instinctual responses to breakdowns in socio-technical systems, with ramifications for jury and judge decision-making. The empirical work conducted for this article shows that education regarding handoffs can lead to more equitable attributions of fault. On the academic end, by identifying the effect and a potential mechanism precipitating it, the article stands to advance scholarship in a host of areas, including torts and criminal law, the nascent field of empirical jurisprudence, and public policy. Moreover, these results and this article should form a

[52] Bob Sorokanich, *Tesla Autopilot First Ride*, RD. & TRACK (Oct. 14, 2015), http://www.roadandtrack.com/new-cars/car-technology/news/a27044/tesla-autopilot-first-ride-almost-as-good-as-a-new-york-driver/ [https://perma.cc/4RPT-CVT4] (first omission in original).

[53] *See* Alicke, *supra* note 41; Thomson, *supra* note 41; McGrath, *supra* note 41.

foundation for much future research on the psychology of apportioning fault, guilt, responsibility, moral blame, and related concepts in the context of collaborative human-machine decision-making.

This article proceeds in four Parts. Part I presents the legal and psychological frameworks within which I investigate lay imputations of fault in socio-technical systems failures. Part II, the empirical pith of this article, presents the methodology and quantitative results of two sets of original experimental studies that demonstrate systematic bias and the precipitating role of moral capacity in lay perception of m2h handoffs. Part III proposes potential mechanisms to explain the experimental findings, drawing from legal theory, empirical jurisprudence, and social and cognitive psychology. Part IV suggests ways in which scholars, legislators, and the legal system and its primary players can build upon these insights and revise existing practices to address the novel aspects of human-machine collaborations that this research brings to light. It points the way to more precisely identifying and remedying misalignments between legal assumptions and the psychological realities of lay adjudication, as the psychology of how lay decision makers impute legal liability in the context of socio-technical systems will feature prominently in legislation, litigation, and legal scholarship over the next decade. I suggest that those crafting policy might need to rethink both their proposals and their intended audiences. The final part also highlights further sociopsychological variables and legal doctrines that merit experimental investigation for a fuller understanding of how lay decisionmakers determine legal liability in the wake of socio-technical system failures. Future studies could shed light on the variables, moderators, and mediators at play in such failures, mainly through study of additional scenarios, types of handoffs, and other aspects of socio-technical systems. They also could explore the factors that influence perception of moral capacity, in machines for sure, but in humans with different demographic characteristics as well.

I. LEGAL AND PSYCHOLOGICAL FRAMEWORK

In legal education, distributed liability is typically first considered across the plaintiff-defendant divide. Did the plaintiff's own negligence contribute to the accident and/or the injuries? Depending on the jurisdiction, different standards of apportioning liability and subsequent damages judgment are imposed.[54] These standards are the well-known contributory negligence and

[54] JOHN C. P. GOLDBERG ET AL., TORT LAW: RESPONSIBILITIES AND REDRESS ch. 7 (4th ed. 2021).

comparative negligence, as well as variants of the two.[55] But distributed liability is also considered within a set of defendants.[56] If the actions of two or more defendants may have jointly caused harm, how should liability be apportioned across these defendants? This ground is well-trodden. For instance, many, if not most, medical malpractice lawsuits implicate multiple individuals and entities, from the doctor to the nurse to the medical facility itself.[57]

Technology is increasingly providing a variant on this second distributed liability analysis. Since the 1980s, we have witnessed the rise of "ubiquitous computing," where cars, homes, even jewelry and nearly everything else with which we interact is embedded with computer chips and some form of machine intelligence.[58] When there is a single human operator, say, a driver at the wheel of a semi-autonomous vehicle, ostensibly there is just one defendant: the human operator. But this conclusion is complicated when one considers the actual driving environment. In the time leading up to the accident, the semi-autonomous vehicle may have been making most, if not all, of the driving decisions. It was doing this through the real-time functioning of an artificial intelligence, which in turn was designed and trained by a team of developers and engineers, who in turn were employed by one or more companies.[59] In essence, when a semi-autonomous vehicle gets into an accident, that accident is really a collision between the plaintiff, the defendant, and a host of other individuals and entities. That some of these actors are human and others are machines might not matter from a theoretical stance (there are humans, after all, behind the curtain of all machines), but it certainly could matter in the theory and application of legal principles. First, regarding theory, there has been a flurry of scholarship on how liability ought to be distributed when such accidents occur. Second, regarding application, there has been a conspicuous lack of research into how individuals will apply principles of fault to accidents that involve both human and machine actors.

This part of the article provides an overview of socio-technical systems with an emphasis on the sociology of handoffs within such systems. Then, it covers the legal posture pertaining to socio-technical system liability. It then provides a synthesis of the various prescriptive analyses that have been

[55] *Id.*

[56] *Id.*

[57] *Id.*

[58] Mark Weiser, *Some Computer Science Issues in Ubiquitous Computing*, 36 COMMC'NS ACM 75 (1993); James Gleick, *Watch This Space*, N.Y. TIMES MAG., July 9, 1995, at 14.

[59] Jack Karsten & Darrel M. West, *Semi-Autonomous Vehicles Must Watch the Road and the Driver*, BROOKINGS INST. (Jan. 30, 2017), https://www.brookings.edu/blog/techtank/2017/01/30/semi-autonomous-vehicles-must-watch-the-road-and-the-driver/ [https://perma.cc/9FKD-4P5M].

proffered for such fault attributions. These discussions foreground the importance of buttressing the prescriptive with a descriptive account of liability for socio-technical system failures. The part then concludes with an overview of general psychological understandings of distributed fault and proximate cause. It is on this foundation that the article's hypotheses are based.

A. Handoffs in Socio-Technical Systems

Socio-technical systems[60] have been empirically studied at least since the mid-twentieth century, when Trist and Bamforth published foundational work at the Tavistock Institute of Human Relations in the United Kingdom.[61] While this early work was focused more on mechanization than automation, the latter became a more prominent focus in the 1970s and more recently, as seen in the work of DeGreene.[62] With the rise of ubiquitous computing, the emphasis of such study shifted more and more to analysis of the human-machine collaborative dyad and, more specifically, potential breakdowns in performance related to this collaboration.[63] One particularly treacherous moment is the "handoff," which I define below but which can be introduced through a brief example.

At its nascence, manned airflight was largely a manual operation.[64] One or more human pilots controlled the machine and made all necessary flight decisions.[65] But a mere nine years after the Wright brothers' successful flight at Kitty Hawk, the Sperry Corporation developed a rudimentary form of autopilot.[66] As reported in *Popular Science Monthly*, the "remarkable gyro-electric mechanism holds the stick and guides an airplane on its course for three hours without human aid."[67] In the decades since, and especially in the modern era of flight, Boeing and other companies have revolutionized flight by increasing the sophistication and reach of these automated systems, creating "automated cockpits."[68] Of course, there were and still are human

[60] Baxter & Sommerville, *supra* note 19.

[61] Eric Lansdown Trist & Kenneth W. Bamforth, *Some Social and Psychological Consequences of the Longwall Method of Coal-Getting*, 4 HUM. RELS. 3 (1951).

[62] KENYON B. DE GREENE, SOCIOTECHNICAL SYSTEMS (1st ed. 1973).

[63] Baxter & Sommerville, *supra* note 19.

[64] Walter J. Boyne, *History of Flight*, ENCYC. BRITANNICA, https://www.britannica.com/technology/history-of-flight [https://perma.cc/A9R3-LSUD].

[65] *Id.*

[66] *Now—The Automatic Pilot*, POPULAR SCI. MONTHLY, Feb. 1930, at 22.

[67] *Id.*

[68] Dan Manningham, *The Cockpit: A Brief History*, 80 BUS. & COM. AVIATION 56 (1997).

pilots, and their remit includes the gaps beyond the automation.[69] For instance, when there is a mechanical issue of some sort, autopilot typically will pass control to the pilots: there is a "handoff."[70]

Nearly every socio-technical system is designed so that decision-making can be shifted—handed off—from the machine to a human operator.[71] This occurs in one of two primary instances. First, it may occur when there is an atypical or overly complex circumstance that overwhelms the machine's ability to function.[72] One common example is when a semi-autonomous vehicle encounters heavy rain that obstructs its camera vision, and control is handed off to the human driver.[73] While this is the most salient version, handoffs also appear in a less well-identified version, one that is commonly discussed in the healthcare domain, where a handoff is "described as the transfer of patient information and knowledge, along with authority and responsibility, from one clinician or team of clinicians to another clinician or team of clinicians."[74] We can clarify this definition as follows: an entity provides information that another entity must rely upon to make a decision, and the latter entity has no real way of checking whether the information provided is accurate.[75] Handoffs are fraught and intellectually complex because they (1) demand that the handoff recipient suddenly assume the decision-making role; (2) place the recipient at a distinct disadvantage because of the handoff; and (3) are "conjunctive cases," such that the outcome would not occur if either the handoff-maker or the handoff-recipient were absent, and both must be present for it to occur.[76]

Point (2) is a key point, and it might not be intuitive. However, human factors research has long known that handoffs are fraught.[77] While the focus of this article is on m2h handoffs, it is worth starting this branch of the

[69] Boyne, *supra* note 64.

[70] CHARLES E. BILLINGS, NAT'L AERONAUTICS & SPACE ADMIN., PUB. NO. 103885, HUMAN-CENTERED AIRCRAFT AUTOMATION: A CONCEPT AND GUIDELINES 1 (1991); *see also* Nadine B. Sarter & David D. Woods, *Pilot Interaction with Cockpit Automation: Operational Experiences with the Flight Management System*, 2 INT'L J. AVIATION PSYCH. 303 (1992).

[71] *See, e.g.*, BILLINGS, *supra* note 70; *see also* Sarter & Woods, *supra* note 70.

[72] BILLINGS, *supra* note 70, at 1–2*; see also* Sarter & Woods, *supra* note 70.

[73] Thierry Bellet, Jean-Michel Hoc, Serge Boverie & Guy Andre Boy, *From Human-Machine Interaction to Cooperation: Towards the Integrated Copilot*, *in* HUMAN-COMPUTER INTERACTION IN TRANSPORTATION 129 (C. Kolski ed., 2011).

[74] COMM. ON PATIENT SAFETY & QUALITY IMPROVEMENT, AM. COLL. OBSTETRICIANS & GYNECOLOGISTS, PUB. NO. 517, COMMUNICATION STRATEGIES FOR PATIENT HANDOFFS 1 (2012).

[75] *Id.*

[76] Thomas F. Icard, Jonathan F. Kominsky & Joshua Knobe, *Normality and Actual Causal Strength*, 161 COGNITION 80, 82 (2017).

[77] Thierry Bellet et al., *From Semi to Fully Autonomous Vehicles: New Emerging Risks and Ethico-Legal Challenges for Human-Machine Interactions*, 63 TRANSP. RSCH. PART F: TRAFFIC PSYCH. & BEHAV. 153, 155–60 (2019).

discussion by describing how such handoffs are fraught even when they are human-to-human handoffs. As mentioned above, handoffs are a feature of the medical system.[78] Patient care is seldom managed by a single provider; it is often managed by a team that is differently constituted over time.[79] In studying doctor-to-doctor handoffs, nurse-to-nurse handoffs, and facility-to-facility handoffs, multiple studies have shown that these handoffs represent a major jeopardy to safe patient care.[80] The problem is so significant that entire research teams are focused on improving the issue.[81]

Of course, handoffs are necessary. Humans get fatigued, they lose their vigilance and mental acuity. They need sleep. Indeed, just as human-to-human handoffs are designed on account of these very needs, machines are often introduced into human-machine collaborative dyads because they likewise can help meet such needs. A few decades ago, a report by the Flight Safety Foundation analyzed automated cockpits with the aim of identifying the extent to which such cockpits decreased pilot workload and subsequent fatigue.[82] It concluded that, while workload went down, pilot performance did not necessarily improve.[83] It mentioned a fatal accident in 1974, where the aircraft crashed short of the runway during a rather standard approach.[84] "There is certainly evidence that underarousal and, possibly, complacency" were to blame: at the final key moments, the conversation in the cockpit was "quite casual and completely unrelated to the flight task."[85]

Or consider the first known fatal accident involving an autonomous car. In May 2016, Joshua Brown was in a Tesla Model S with the vehicle operating in self-driving mode.[86] A white tractor trailer drove across the highway, and neither the autopilot nor Mr. Brown detected the obstacle; the

[78] *See* COMM. ON PATIENT SAFETY & QUALITY IMPROVEMENT, *supra* note 74.

[79] *See id.*

[80] Vineet Arora et al., *Communication Failures in Patient Sign-Out and Suggestions for Improvement: A Critical Incident Analysis*, 14 BMJ QUALITY & SAFETY HEALTHCARE 401 (2005); Leora I. Horwitz et al., *Transfers of Patient Care Between House Staff on Internal Medicine Wards: A National Survey*, 166 ARCHIVES INTERNAL MED. 1173 (2006); Leora I. Horwitz et al., *Consequences of Inadequate Sign-Out for Patient Care*, 168 ARCHIVES INTERNAL MED. 1755 (2008).

[81] Lindsay J. Blazin et al., *Improving Patient Handoffs and Transitions Through Adaptation and Implementation of I-PASS Across Multiple Handoff Settings*, 5 PEDIATRIC QUALITY & SAFETY, July/Aug. 2020, at 1.

[82] Alan H. Roscoe, *Workload in the Glass Cockpit*, FLIGHT SAFETY DIG. 1 (1992).

[83] *Id.*

[84] *Id.*

[85] *Id.* at 6 (quoting post-crash report).

[86] Will Oremus, *The Tesla Autopilot Crash Victim Was Apparently Watching a Movie When He Died*, SLATE (July 1, 2016), https://slate.com/business/2016/07/tesla-autopilot-crash-victim-joshua-brown-was-watching-a-movie-when-he-died.html [https://perma.cc/ADV5-5MGK].

brakes were not applied.[87] At the time of the collision, Mr. Brown was watching a Harry Potter movie.[88]

In brief, the handoff problem is primarily a combination of four linked issues: complacency, inattention, bias (over-trust in the entity making the handoff), and skill atrophy.[89] While all four of these apply to both h2h and m2h handoffs, let us explore them in the context of m2h handoffs and the voluminous research on the topic.

The first three are closely related. As automation increases, inattention and complacency increase. Automation complacency refers to the tendency to monitor one's environment less frequently and less astutely when technology is providing information about the same.[90] Evidence of automation complacency has been found in all of the following settings: industrial monitoring,[91] air traffic control,[92] aviation crashes,[93] and the grounding of a passenger ship,[94] among countless others. The NTSB has been examining automation complacency in the operation of limited semi-autonomous vehicles, such as those that can maintain control and handle, say, slowing traffic but also require driver monitoring due to the limited capabilities.[95] The NTSB has documented that even such limited semi-autonomous vehicles lead to complacency and contributed to recent accidents in Florida and California.[96]

In the aftermath of an Uber autonomous vehicle collision, the official report generated by the National Transportation Safety Board cited, among

[87] *A Tragic Loss*, TESLA (June 30, 2016), https://www.teslamotors.com/blog/tragic-loss [https://perma.cc/VHA3-TBN3].

[88] Oremus, *supra* note 86.

[89] RAPOSO ET AL., *supra* note 44.

[90] Raja Parasuraman & Dietrich H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381, 382 (2010).

[91] Raja Parasuraman, Robert Molloy & Indramani L. Singh, *Performance Consequences of Automation-Induced 'Complacency'*, 3 INT'L J. AVIATION PSYCH. 1 (1993).

[92] Ulla Metzger & Raja Parasuraman, *The Role of the Air Traffic Controller in Future Air Traffic Management: An Empirical Study of Active Control Versus Passive Monitoring*, 43 HUM. FACTORS 519 (2001).

[93] Ken Funk et al., *Flight Deck Automation Issues*, 9 INT'L J. AVIATION PSYCH. 109, 109–23 (1999).

[94] NAT'L TRANSP. SAFETY BD.., REP. NO. NTSB/MAR-97/01, GROUNDING OF THE PANAMANIAN PASSENGER SHIP ROYAL MAJESTY ON ROSE AND CROWN SHOAL NEAR NANTUCKET, MASSACHUSETTS, JUNE 10, 1995 (1997).

[95] NAT'L TRANSP. SAFETY BD., REP. NO. NTSB/HAR-17/02, COLLISION BETWEEN A CAR OPERATING WITH AUTOMATED VEHICLE CONTROL SYSTEMS AND A TRACTOR-SEMITRAILER TRUCK NEAR WILLISTON, FLORIDA, MAY 7, 2016 (2017) [hereinafter TRACTOR COLLISION]; NAT'L TRANSP. SAFETY BD., REP. NO. NTSB/HAB-19/07, REAR-END COLLISION BETWEEN A CAR OPERATING WITH ADVANCED DRIVER ASSISTANCE SYSTEMS AND A STATIONARY FIRE TRUCK, CULVER CITY, CALIFORNIA, JANUARY 22, 2018 (2019) [hereinafter FIRE TRUCK COLLISION].

[96] TRACTOR COLLISION, *supra* note 95; FIRE TRUCK COLLISION, *supra* note 95.

other factors, the "lack of adequate mechanisms for addressing operators' automation complacency."[97] As we will see, this conclusion is somewhat misguided, since, while it might seem that automation complacency could be mitigated with proper training and/or expertise, a meta-analysis found that it occurs in both naïve and expert individuals and cannot be prevented by training or explicit instructions to verify the recommendations of the AI.[98]

There is a sense in which, if not complacency, then something similar seems *desirable*, as automation has been touted as capable of freeing humans to focus on more important decisions.[99] But complacency occurs in tandem with both inattention and over-trust, and these bring consequences.[100] A study on pilots showed that, as automation increases, awareness decreases.[101] In the medical domain, this presents as marked failure to detect machine mistakes, with studies showing that nearly half of all human users will fail to detect any machine errors over the course of a day of collaborative work.[102] If the human role is primarily a monitoring one, poor performance should be expected.[103] This is especially true when the technical component has a low failure rate.[104]

This behavior is psychologically reasonable. People are generally efficient actors, taking the route that requires the least cognitive effort.[105] When paired with a machine, the impulse is to offload more and more of one's cognitive load onto the machine.[106] But this tendency leads to cognitive bias:[107] automation bias is the tendency to favor technological guidance, even in the face of signs that the guidance is incorrect.[108] Some have referred to this as

[97] NAT'L TRANSP. SAFETY BD., COLLISION BETWEEN VEHICLE CONTROLLED BY DEVELOPMENTAL AUTOMATED DRIVING SYSTEM AND PEDESTRIAN, TEMPE, ARIZONA, MARCH 18, 2018, at 59 (2020).

[98] Parasuraman & Manzey, *supra* note 90, at 397.

[99] Lisanne Bainbridge, *Ironies of Automation*, *in* ANALYSIS, DESIGN AND EVALUATION OF MAN–MACHINE SYSTEMS 129–35 (1983); Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse*, 39 HUM. FACTORS 230, 230–53 (1997).

[100] Bainbridge, *supra* note 99, at 129–35; Parasuraman & Riley, *supra* note 99, at 230–35.

[101] Stephen M. Casner & Jonathan W. Schooler, *Thoughts in Flight: Automation Use and Pilots' Task-Related and Task-Unrelated Thought*, 56 HUM. FACTORS 433 (2014).

[102] *See* Parasuraman & Manzey, *supra* note 90, at 389; Parasuraman et al., *supra* note 91.

[103] Victoria A. Banks et al., *Is Partially Automated Driving a Bad Idea? Observations from an On-Road Study*, 68 APPLIED ERGONOMICS 138 (2018).

[104] DAVID ROY DAVIES & RAJA PARASURAMAN, THE PSYCHOLOGY OF VIGILANCE (1982).

[105] *See* Coiera, *supra* note 46, at 420.

[106] *Id.*

[107] Kate Goddard, Abdul Roudsari & Jeremy C. Wyatt, *Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators*, 19 J. AM. MED. INFORMATICS ASS'N 121 (2012) [hereinafter *Automation Bias 2012*].

[108] *Id.*

"over-trust."[109] Such errors are often observed in the medical arena.[110] Recent analyses have shown that medical clinicians will override their own (correct) judgments and follow (incorrect) guidance from technology.[111]

The automation of driving is sometimes conceived of in levels. The Society of Automotive Engineers (SAE) created a taxonomy that ranges "from no driving automation (level 0) to full driving automation (level 5)".[112] Autonomous vehicles are now in the middle levels, wherein the AI requires handoffs to humans at various points.[113] Complacency, inattention, and over-trust are all products of the monitoring role.[114] One does not have to perform the task; one just watches. A product of watching rather than performing is that skills deteriorate. As automation becomes more reliable, and the human driver is called upon to act less frequently, the human driver's performance becomes worse.[115] Without practice and use, human operators' skills atrophy.[116] This is partly reflected in the extremely poor performance observed in the wake of m2h handoffs. "Human factors research has proven this 'handoff' scenario detracts from, rather than enhances, human performance."[117]

Time is another factor that impacts post-handoff performance. In general, humans are not well-adapted to regain control in a limited time frame.[118]

[109] John D. Lee & Katrina A. See, *Trust in Automation: Designing for Appropriate Reliance*, 46 HUM. FACTORS 50, 55 (2004).

[110] *See* INST. FOR SAFE MEDICATION PRACS., *Understanding Human Over-Reliance on Technology* (Sept. 8, 2016), https://www.ismp.org/resources/understanding-human-over-reliance-technology [https://perma.cc/FAX2-NQ6G].

[111] *Automation Bias 2012*, *supra* note 107; *see also* Kate Goddard, Abdul Roudsari & Jeremy C. Wyatt, *Automation Bias: Empirical Results Assessing Influencing Factors*, 83 INT'L J. MED. INFORMATICS 368 (2014) [hereinafter *Automation Bias 2014*].

[112] SAE INT'L, TAXONOMY AND DEFINITIONS FOR TERMS RELATED TO DRIVING AUTOMATION SYSTEMS FOR ON-ROAD MOTOR VEHICLES (2018), https://www.sae.org/standards/content/j3016_201806/ [https://perma.cc/F9DT-RAX7].

[113] *See id.*

[114] *Id.*

[115] Paul C. Schutte, *How To Make the Most of Your Human: Design Considerations for Human–Machine Interactions*, 19 COGNITION, TECH. & WORK 233 (2017).

[116] William Langewiesche, *The Human Factor*, VANITY FAIR (Sept. 17, 2014), http://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash [https://perma.cc/KM8D-8FYG]; Nadine B. Sarter, David D. Woods & C. E. Billings, *Automation Surprises*, *in* HANDBOOK HUM. FACTORS & ERGONOMICS 1926 (Gavriel Salvendy ed., 1997).

[117] Elish, *supra* note 8, at 50.

[118] Zhenji Lu, Xander Coster & Joost De Winter, *How Much Time Do Drivers Need To Obtain Situation Awareness? A Laboratory-Based Study of Automated Driving*, 60 APPLIED ERGONOMICS 293 (2017).

Humans don't do well when they have to take a handoff at the last minute.[119] In the driving context, studies have shown that there is a period of high risk that follows m2h handoffs.[120] At first, there is an attentional lag. Distracted drivers require more than five seconds to regain proper control of their vehicle.[121] Other research has shown that reorientating oneself back to a driving task might take even longer, with results showing a minimum of eight seconds[122] and up to forty seconds.[123] When drivers are given limited time to take the handoff, subsequent decision-making is predictably suboptimal.[124]

Changing conditions also make reorientation difficult. Researchers have shown that gaps in control, such as being in control at low speeds and not being in control again until at high speeds, create performance issues.[125] Moreover, these issues are not readily amendable to correction.[126] Even when auditory, visual, and haptic warnings are used to alert drivers to an impending handoff, drivers still struggle to regain attention.[127] Google's self-driving car program concluded that it could not solve the handoff problem,[128] and it has

[119] Roscoe, *supra* note 82, at 6; *cf.* EARL L. WEINER, NAT'L AERONAUTICS & SPACE ADMIN., PUB. NO. 177528, HUMAN FACTORS OF ADVANCED TECHNOLOGY ("GLASS COCKPIT") TRANSPORT AIRCRAFT 169–181 (1989); *see also* Victoria A. Banks, Katherine L. Plant & Neville A. Stanton, *Driver Error or Designer Error: Using the Perceptual Cycle Model To Explore the Circumstances Surrounding the Fatal Tesla Crash on 7th May 2016*, 108 SAFETY SCI. 278 (2018).

[120] Alexander Eriksson & Neville A. Stanton, *Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and from Manual Control*, 59 HUM. FACTORS 689 (2017); Christian Gold, Riender Happee & Klaus Bengler, *Modeling Take-Over Performance in Level 3 Conditionally Automated Vehicles*, 116 ACCIDENT ANALYSIS & PREVENTION 3, 3–4 (2018).

[121] *See* Brian Mok et al., *Tunneled In: Drivers with Active Secondary Tasks Need More Time To Transition from Automation*, PROC. OF THE 2017 CHI CONF. ON HUM. FACTORS IN COMPUTING SYS. 2840 (2017) ("[T]he majority of drivers in the 5 or 8 second conditions were able to navigate the hazard situation safely.").

[122] Ravi Agrawal et al., *Effects of a Change in Environment on the Minimum Time to Situation Awareness in Transfer of Control Scenarios*, 2663 TRANSP. RSCH. REC. 126 (2017).

[123] Natasha Merat et al., *Transition to Manual: Driver Behaviour When Resuming Control from a Highly Automated Vehicle*, 27 TRANSP. RSCH. PART F: TRAFFIC PSYCH. & BEHAV. 274 (2014).

[124] Natasha Merat et al., *Highly Automated Driving, Secondary Task Performance, and Driver State*, 54 HUM. FACTORS 762 (2012).

[125] Holly E. B. Russell et al., *Motor Learning Affects Car-to-Driver Handover in Automated Vehicles*, SCI. ROBOTICS, Dec. 6, 2016, at 1–5, https://www.science.org/doi/10.1126/scirobotics.aah5682 [https://perma.cc/MA4F-J8ZR].

[126] *See id.*

[127] *Ford's Dozing Engineers Side with Google in Full Autonomy Push*, INDUSTRYWEEK (Feb. 17, 2017), https://www.industryweek.com/innovation/product-development/article/22007061/fords-dozing-engineers-side-with-google-in-full-autonomy-push [https://perma.cc/YD94-RFVB].

[128] John Markoff, *Google Car Exposes Regulatory Divide on Computers as Drivers*, N.Y. TIMES (Feb. 10, 2016), http://www.nytimes.com/2016/02/11/technology/nhtsa-blurs-the-line-

decided to try to skip the semi-autonomous stages and progress directly to fully autonomous vehicles.[129]

B. The Law of Socio-Technical System Liability

How do socio-technical systems lead to harm? To answer this, let us consider the primary types of socio-technical system failures.[130] First, there is the case of clear human operator error. This would result in a negligence-based tort, possibly professional negligence depending upon the specifics. Second, there is the case of clear mechanical failure. Especially when this can be established by a history of such failures in the same (or the same class of) product, the matter will be moved into the realm of products liability. For example, the company that produced the previously mentioned da Vinci Surgical Robot has faced a slew of product liability-type lawsuits.[131] In this article, I am primarily concerned with the in-between space, where there is neither clear human nor clear machine error.

The first legal (or quasi-legal) rule for assigning responsibility when an autonomous agent causes harm involved a highly complex but relatively low-tech agent: an ox.[132] Both the Code of Eshnunna and the Code of Hammurabi, which date from about the eighteenth-century BCE, hold that if the owner of an ox knew that the ox was a "habitual gorer", then the owner would be at fault.[133] If the ox was not previously known to gore, then the owner would be likely to escape liability.[134] Who would be at fault? Technically, the ox. What does that look like? Probably somewhat like Arizona in recent years, where people have made news for castigating Waymo autonomous vehicles through

between-human-and-computer-drivers.html?smid=tw-share&_r=0 [https://perma.cc/LUQ4-EUSK].

[129] JONES DAY PUBL'NS, LEGAL ISSUES RELATED TO THE DEVELOPMENT OF AUTOMATED, AUTONOMOUS, AND CONNECTED CARS 1, 2–3 (2017), https://www.jonesday.com/files/Publication/f5cf8577-3267-4f78-bbf8-ec32333cc49b/Preview/PublicationAttachment/4a78a73f-67e6-4d18-9845-ed6b0eb7561e/Legal%20Issues%20Related%20to%20Autonomous%20Cars.pdf [https://perma.cc/DU2E-A8AZ].

[130] If the matter involves a fully autonomous machine, such that the machine has a human creator but there is no ongoing human involvement, then this is not a socio-technical system. There is no "socio." If a security robot injures a bystander, one must seek recourse by suing the owner of the robot. Roger Michalski, *How To Sue a Robot*, 5 UTAH L. REV. 1021, 1025 (2018).

[131] Flood, *supra* note 12.

[132] DARLING, *supra* note 32, at 67–69.

[133] *Id.* at 67–68.

[134] *Id.* at 68.

physical attacks,[135] just as in the Middle Ages animals were routinely tried for crimes and duly punished with imprisonment, hanging, or torture.[136]

More recently, the issue of liability when alleged harmful conduct is distributed across human and machine actors has been debated on multiple fronts.[137] Some claim that such torts and potentially criminal acts pose no great hurdle for current regulatory and legal frameworks; after all, the law knows how to assemble all the potential tortfeasors and evaluate joint and several liability,[138] and any difficulties would concern the complex but mundane task of how best to apportion liability among the various creators of the technology.[139] Others argue that existing frameworks will be sorely tested and stretched. For instance, a recent article by a multidisciplinary team of scholars noted, "[T]he legal community and ethicists are struggling to come to terms with the implications of Automated Vehicles (AV) in terms of liability regimes and questions of responsibility and culpability. Key to understanding the inherent complexities is the notion of *handover/takeover transitions* . . . ."[140] In this section, I explore the contours of the law of socio-technical system liability, beginning with civil suits before addressing criminal matters.

The standard legal response to incurred harm[141] is to trace causation until landing upon a negligent (or *the* negligent) party.[142] This negligent (or reckless or intentionally harmful or omissive) party may then be sued with the aim of—well, the aim depends upon the type of behavior and harm. For

[135] *See* Simon Romero, *Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars*, N.Y. TIMES (Dec. 31, 2018), https://www.nytimes.com/2018/12/31/us/waymo-self-driving-cars-arizona-attacks.html [https://perma.cc/67JE-CZ6V].

[136] DARLING, *supra* note 32, at 76–77.

[137] *See, e.g.*, F. Patrick Hubbard, *"Sophisticated Robots": Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 1803 (2014); Daniel A. Crane, Kyle D. Logue & Bryce C. Pilz, *A Survey of Legal Issues Arising from the Deployment of Autonomous and Connected Vehicles*, 23 MICH. TELECOMM. & TECH. L. REV. 191 (2017).

[138] *See, e.g.*, RESTATEMENT (SECOND) OF TORTS § 875 (AM. L. INST. 1979) (stating that tortfeasors causing indivisible harm are jointly liable for such harm); *id.* § 881 (stating that tortfeasors causing divisible harms are only severally liable for such harms); RESTATEMENT (THIRD) OF TORTS: APPORTIONMENT OF LIAB. §§ 10–21 (AM. L. INST. 2000) (discussing liability of multiple tortfeasors for indivisible harm).

[139] Crane et al., *supra* note 137, at 259–61.

[140] Bellet et al., *supra* note 77, at 153.

[141] In tort law, people are protected only when behavior causes harm. A plaintiff may recover damages only if the defendant breached the duty of care and harm was caused by the breach. *See, e.g.*, RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 39 cmt. D (AM. L. INST. 2010) ("[Liability is limited to when] there is a close connection between the breach of duty and the ensuing harm."); *id.* § 26 cmt. B.

[142] Strict liability, of course, eliminates the need to identify negligence, recklessness, and so on.

brevity's sake, we can say the aim is to make the injured party whole. While wholeness is almost always reduced to an economic question (what amount of money will restore the injured party?), the question pertaining to liability is multifaceted. "Proximate cause" (sometimes called "legal cause") is the star of the show, but discussion of proximate cause must be preceded by discussion of "factual cause" (also called "cause-in-fact"): is X a but-for cause of Y? If yes, if it is the case that Y would not have occurred without X, then X is a factual cause of Y.[143] When there are multiple defendants with distributed liability, factual cause still is necessary. In *Sindell v. Abbott Laboratories*,[144] for example, the Supreme Court of California used a "market share" approach, holding that the various defendants, each of which manufactured a drug that had harmed the plaintiff (but none of which could be conclusively identified as the manufacturer of the specific drug that had rendered the harm), would be held liable in proportion to their market share, in other words, in proportion to the odds that they had manufactured the specific harmful drug.[145]

If there is one primary weakness in the concept of factual cause, it is overbreadth. After all, if a person commits murder, that person's mother is a but-for cause of the murder. If she had not given birth to the murderer, the murder would not have occurred. In the interest of saving mothers from the gallows, then, proximate cause is introduced. From the set of all factual causes, which ones ought to be held responsible?

The answer might rest on policy considerations.[146] It might rest in something else.[147] It often at least partially rests on a consideration of whether the injury was a "reasonably foreseeable" outcome of the behavior,[148] just as it often rests on notions of different classes of causes, since surely there were

[143] RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 26 cmt. B (AM. L. INST. 2010).

[144] 607 P.2d 924 (Cal. 1980).

[145] *Id.* at 937.

[146] RESTATEMENT (SECOND) OF TORTS § 431 cmt. A (AM. L. INST. 1965); *see* In re M.S., 896 P.2d 1365, 1386–87 (Cal. 1995) (Kennard, J., concurring); FOWLER V. HARPER & FLEMING JAMES JR., THE LAW OF TORTS § 20.4 (1986) ("[P]olicy considerations underlie the doctrine of proximate cause.").

[147] James Angell McLaughlin, *Proximate Cause*, 39 HARV. L. REV. 149, 155–60 (1925) (discussing various purposes for proximate cause, including fairness and justice); Mitchell v. Gonzales, 819 P.2d 872, 882–85 (Cal. 1991) (Kennard, J., dissenting) (describing a "social evaluative process" inherent to deciding which causes will be held legally liable).

[148] *See, e.g.*, HARPER & JAMES, *supra* note 146, § 20.5 (discussing foreseeability within the causation doctrines).

several intervening and superseding causes between the mother giving birth to the future murderer and the murder itself.[149]

Moreover, proximate cause might not even be a coherent concept.[150] Courts often mix tests for factual and proximate causes.[151] They sometimes make a conclusion about liability and appear to slap the proximate cause label on after the conclusion is made.[152] There are deep issues here, interesting ones concerning the psychology of what people are concluding when they make proximate cause conclusions. These issues are discussed in the final section of this part.

Applying these principles to socio-technical systems is a struggle. Writing in 1996, Judge Curtis E. A. Karnow put it thus:

> [M]ultiple agent systems imply at least the causal input of multiple independent programmers of the basic scripting or authoring software, a vast number of users creating distinct intelligent agents, and an unpredictable number of agent to agent interactions on an unpredictable number of interwoven platforms, operating systems, distributed data and communications programs, each of which in turn incorporates at least some further limited programming. This inevitable causal complexity poses problems for traditional tort law, in which a determination of proximate cause is essential, as it evaluates the liability of an intelligent machine system.[153]

That said, there are some relatively clear points. If there is a manufacturing defect, such that the machine does not function in accordance with its design, then this would move the matter out of the negligence realm and into the relatively more straightforward realm of strict liability, as outlined in Section 2(a) of the Products Liability Restatement ("PLR").[154] We might say that there are four theories of strict liability.[155] One relates to ultrahazardous activities, which is less relevant here, and the rest relate to products liability, where there might be a failure to warn, a design defect, or a manufacturing defect.[156] For the latter, and in the context of autonomous vehicles, the

[149] O'Brien v. B.L.C. Ins. Co., 768 S.W.2d 64, 68 (Mo. 1989); *see also* Doe v. Manheimer, 563 A.2d 699 (Conn. 1989) (finding the conduct of a rapist not reasonably foreseeable); Erikson v. Curtis Inv. Co., 447 N.W.2d 165 (Minn. 1989) (finding that a parking operator owes customers a duty of care to protect against foreseeable criminal behavior).

[150] Mark F. Grady, *Proximate Cause Decoded*, 50 UCLA L. REV. 293, 294 (2002).

[151] Green, *supra* note 38, at 623.

[152] *Id.*

[153] Curtis E.A. Karnow, *Liability for Distributed Artificial Intelligences*, 11 BERKELEY TECH. L.J. 147, 182 (1996).

[154] *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(a) (AM. L. INST. 1998).

[155] Anderson v. Owens-Corning Fiberglas Corp., 810 P.2d 549, 553 (Cal. 1991).

[156] *See id.*; Barker v. Lull Eng'g Co., 573 P.2d 443, 446–47 (Cal. 1978).

complicating issue would be that of proof, such that the plaintiff would have the burden to persuade that the defect actually occurred.

Indeed, for fully autonomous vehicles, there is a "shared conclusion" that manufacturers will be responsible as a simple matter of products liability.[157] The claims will primarily involve defects in design or warnings,[158] with such claims falling under PLR Section 2(b), which covers risk-utility design defects, or PLR Section 2(c), which covers failures to provide reasonable instructions or warnings.[159] Of course, there is an array of sub-issues that would complicate any such claims, including ones pertaining to determination of design defects and others pertaining to what standards for warning ought to be adopted.[160]

One subset of the warnings issue—indeed, a subset that probably moves it from a warnings issue to one of design defect—involves those that pertain to handoffs, most obviously in the context of semi-autonomous vehicles. How much and what type of warning should be given in the moments preceding a handoff? Here, one relevant tort doctrine is the Restatement (Third) of Torts, with its mandate for the adoption of fault-tolerant product designs: "[I]nstructions and warnings may be ineffective because users of the product may not be adequately reached, may be likely to be inattentive, or may be insufficiently motivated to follow the instructions or heed the warnings."[161] This echoes the handoff scenarios that lead to inattention, such as in the case of prolonged inactivity when autopilot or similar systems are engaged. The Restatement mandates for safer design instead of warning alone: "[W]hen a safer design can reasonably be implemented and risks can reasonably be designed out of a product, adoption of the safer design is required over a warning that leaves a significant residuum of such risks."[162] So, in the wake of a handoff-related vehicle collision, a plaintiff could claim that the manufacturer of the semi-autonomous vehicle did not adopt a reasonably safe, fault-tolerant design. The pith of the tort inquiry would involve the risk-utility test: would incorporation of additional safety features cost less than the associated safety benefit?[163]

While this seems reasonable enough, what many commentators fail to realize is that such additional safety features would defeat one of the primary

[157] DOROTHY J. GLANCY, ROBERT W. PETERSON & KYLE F. GRAHAM, A LOOK AT THE LEGAL ENVIRONMENT FOR DRIVERLESS VEHICLES 35 (2016).

[158] *Id.*

[159] *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(b)–(c) (AM. L. INST. 1998).

[160] *Id.*

[161] *Id.* at § 2, cmt. 1.

[162] *Id.*

[163] Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 CALIF. L. REV. 1611, 1627–28 (2017).

purposes of automation: lessening cognitive load and attentional demands. Semi-autonomous vehicles are often considered valuable *because* they free the driver from the rather exhausting task of driving. If a manufacturer were to incorporate constant reminders and other ways of stimulating the driver's attention, then why not just have the driver drive? If the driver is already paying full attention, there is little reason to have the autonomous feature. As those working in human factors research have shown, balancing such factors is an incredibly nuanced and involved undertaking.[164]

In addition, there are complications that inhere to the design defect claim. If something unpredictable happens, such as a child darting into the street, or something unusual occurs, such as rare light conditions that interfere with machine vision, what is the point at which the interceding hazards transcend that which can be expected of design engineers? As Professors Abraham and Rabin argue, there might not be anything conceptually distinctive about the needed analysis, but the issues will certainly be technically complex, not least because socio-technical systems will be continually developing, with state-of-the-art an ever-moving goalpost.[165]

Socio-technical systems also have significant applications to criminal law, although many of these are applications and do not entail criminal sanction for the socio-technical system failure. For example, one such application, which will be explored in the experiments presented below, is algorithm-aided judicial decision-making.[166] A judge might work in tandem with an algorithm to determine optimal bail amounts or to determine appropriate criminal sentencing.[167] Moreover, scholars are proposing much more extensive automation of the criminal justice system.[168] If harm results from a breakdown in such socio-technical systems, the entities likely will not be subject to criminal sanction. Granted, the breakdown will impact the functioning of the criminal justice system, but that is a different matter.

When it comes to criminal sanction for joint human-machine behavior, most recent scholarship has focused on the problem of robot behavior in isolation.[169] This is a good place to start, as it gets at the core worry regarding machine criminal liability: *mens rea*. While there is no doubt that the factual

[164] *See* Bellet et al., *supra* note 77.

[165] Abraham & Rabin, *supra* note 18, at 141–42.

[166] *See infra* Part II.C.

[167] *See infra* Part II.C.

[168] BENJAMIN H. BARTON & STEPHANOS BIBAS, REBOOTING JUSTICE: MORE TECHNOLOGY, FEWER LAWYERS, AND THE FUTURE OF LAW (2017).

[169] *See, e.g.*, Gabriel Hallevy, *I, Robot – I, Criminal: When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses*, 22 SYRACUSE SCI. & TECH. L. REP. 1, 18–25 (2010).

element requirement (*actus reus*) can be met when the defendant is a robot,[170] the mental element is trickier. The *mens rea* requirement encapsulates the offender's internal and subjective relation (mentation) to the external and objective commission/omission (behavior).[171] Criminal law recognizes a few forms of mentation: (1) intent and specific intent, (2) indifference, and something along the lines of (3) rashness.[172] The latter two fall under a general sense of recklessness.[173] In a homicide matter, it is hard to imagine that a jury would find that a machine *wanted* a victim to die, absent evidence of a malicious programmer. For an advanced AI, it is still hard to imagine such intent being imputed, as awareness, will, and intent raise thorny issues pertaining to the psychology and philosophy of consciousness. But we certainly can imagine a jury finding that a machine was indifferent to causing a victim's death (although some of the same consciousness concerns might be raised here) or assumed unreasonable risk relating to the victim's death—and with even more certainty we can imagine a jury finding that a programmer held such intent through the machine.

Regardless, these issues are not quite as thorny as those in civil matters, as all parties can be held criminally liable. Just because a robot is found guilty of a crime, this does not mean that the manufacturer, programmer, and user are absolved of their individual criminal liability.[174] But it does raise serious questions about the purpose of holding the robot liable.[175] The purpose of criminal sanction is disputed, implicating notions of retribution, restitution, norm validation, and much more.[176] In the Middle Ages, a rooster was put on trial in Basel, Switzerland; despite deft legal representation, the rooster was convicted of a crime and burned at the stake.[177] We might hesitate, perhaps, even beyond our distaste for capital punishment, to treat Siri in such a manner.

[170] GABRIEL HALLEVY, WHEN ROBOTS KILL: ARTIFICIAL INTELLIGENCE UNDER CRIMINAL LAW 41–45 (2013).

[171] *Id.* at 47.

[172] *Id.* at 48.

[173] *Id.*

[174] Ying Hu, *Robot Criminals*, 52 U. MICH. J.L. REFORM 487, 487 (2019).

[175] *Id.* at 504–10.

[176] Albert W. Alschuler, *The Changing Purposes of Criminal Punishment: A Retrospective on the Past Century and Some Thoughts About the Next*, 70 U. CHI. L. REV. 1 (2003); Henry Weinhofen, *The Purpose of Punishment*, 7 TENN. L. REV. 145 (1929).

[177] DARLING, *supra* note 32, at 76–77.

C. The Prescriptive Landscape

Most legal scholarship on socio-technical systems failures is more focused on the prescriptive rather than the descriptive. We might bend our current framework to accommodate human-machine pairings or even robots in isolation, but how precisely should we bend the framework? In order to grasp the implications of the experiments I present in this article, it is worth briefly surveying this work.

Judge Karnow was an early entrant to the debate, proposing in 1996 that a "Turing Registry" be put in place.[178] Believing that the risks associated with the use of an intelligent agent can be predicted, Judge Karnow suggested that AI developers submit their creations to a certification procedure that can analyze the risk along a spectrum of automation ("the higher the intelligence, the higher the risk") and generate an insurance premium.[179] If we were to classify the prescriptive arguments, this would be one that focuses on the manufacturers, pushing liability and fault onto them and sparing operators/users.

Another example of such an approach can be found in the work of Professor Omri Rachum-Twaig, who has argued for imposing a predetermined level of care on different stakeholders, thus creating a presumption of negligence.[180] This would include a host of duties, including ones of monitoring and emergency shut-down.[181] Liability would arise only when a stakeholder fails to meet one of these duties, although meeting them might still result in a common negligence action.[182]

Professor Geistfeld has focused on fully autonomous vehicles, arguing that they will transform driving into a collective system where there is but one driver: the operating system.[183] Subsequent liability analyses should thus focus on the performance of the entire fleet of vehicles, where the tort obligation would stem from the reasonably safe programming or design of the operating system.[184] Professor Geistfeld draws the products liability line such that, in premarket testing, autonomous vehicles must be at least twice as safe as conventional vehicles.[185] Insurers could then establish risk-adjusted annual premiums based on premarket testing, and the National Highway

[178] Karnow, *supra* note 153, at 193.

[179] *Id.*

[180] Omri Rachum-Twaig, *Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots*, 2020 U. ILL. L. REV. 1141, 1174–75 (2020).

[181] *Id.*

[182] *Id.*

[183] Geistfeld, *supra* note 163, at 1621.

[184] *Id.* at 1627.

[185] *Id.* at 1679.

Transit Safety Administration could adopt federal regulations that support this regulatory approach.[186]

Professors Abraham and Rabin believe that the liability standard that Professor Geistfeld develops is insufficiently exacting.[187] They argue that products liability law has never been premised on relative reasonableness, and thus Professor Geistfeld's two-times safer standard is misguided: an unsafe design is still unsafe even when it's less unsafe than some other design.[188]

Other scholars have taken very different approaches, prescribing liability that is not premised on defectiveness.[189] Professor David Vladeck, for instance, has developed a "common enterprise liability" approach in which there would be strict joint and several liability for all autonomous vehicle-related injuries, with the liability extending to both car manufacturers and manufacturers of component parts.[190]

For their part, Professors Abraham and Rabin propose "Manufacturer Enterprise Responsibility," which would be a manufacturer-financed, strict responsibility bodily injury compensation system that is administered by a fund created through assessments levied on autonomous vehicle manufacturers.[191] Somewhat similar in aim, but focused on individual accidents, Bryan Casey has argued that for an emphasis on "robot *ipsa loquitur*." This pun on *res ipsa loquitor* is meant to highlight that advanced data-logging technologies in machines can be used to provide detailed records of accidents that, in turn, will speak to the fault of the parties involved.[192]

There are many more offerings: Ryan Abbott proposed that liability be based on a negligence standard that treats the vehicle as a person.[193] Dylan LeValley, in contrast, argued that autonomous vehicle manufacturers should

[186] *Id.* at 1674–75.

[187] Abraham & Rabin, *supra* note 18, at 145.

[188] *Id.* at 145–46.

[189] *See, e.g.*, Hubbard, *supra* note 137, at 1866–67 (describing how automobile distributors could create a fund that compensates victims for injuries caused by autonomous automobiles); Jeffrey K. Gurney, Comment, *Sue My Car Not Me: Products Liability and Accidents Involving Autonomous Vehicles*, 2013 U. ILL. J.L. TECH. & POL'Y 247, 271–72 (2013) (arguing that, rather than the manufacturer of the car, the manufacturer of the autonomous driving technology should be liable for accidents).

[190] David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 129 n.39, 146–48 (2014).

[191] Abraham & Rabin, *supra* note 18, at 147.

[192] Bryan Casey, *Robot Ipsa Loquitur*, 108 GEO. L.J. 225, 225 (2019).

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

be treated as "common carriers."[194] Kevin Funkhouser argued for a standardized no-fault compensation system.[195]

Perhaps the most eloquent understanding of the problem, one that is not limited to a subset of automation (i.e., not focused on autonomous vehicles, which is the topic of most of the above authors), is that presented by Jack Balkin, who emphasizes making a connection to the law of nuisance and environmental law.[196]

> The algorithm doesn't have intentions, wants, or desires. . . . Hence it is useless to model the duty or liability of algorithm operators on a *respondeat superior* theory . . . . Instead, we have to focus on the social effects of the use of a particular algorithm, and whether the effects are reasonable and justified from the standpoint of society as a whole.[197]

While most of these arguments do not foreclose imputation of liability to operators, that is not their focus. Others, though, have focused almost exclusively on the operators, arguing that liability should be placed there. Duffy and Hopkins argued for strict liability for autonomous vehicle owners.[198] Julie Goodrich, emphasizing the social benefits of autonomous vehicles, proposed a legislative scheme that immunizes autonomous vehicles from civil liability.[199] Professor Shavell has proposed a rather novel and interesting solution: strict liability for autonomous vehicle accidents, with payment made to the state.[200] He argues that this would more properly incentivize both operators and manufacturers.[201]

Finally, Xuan Di, Xu Chen, and Eric Talley have used game theory to model the various factors that arise in the context of autonomous vehicles and how they impact user and manufacturer behavior, especially in the context of road safety.[202] Their aim was to empirically inform design of socially optimal

[194] Dylan LeValley, Comment, *Autonomous Vehicle Liability—Application of Common Carrier Liability*, 36 SEATTLE U. L. REV. 5, 6 (2013).

[195] Kevin Funkhouser, Note, *Paving the Road Ahead: Autonomous Vehicles, Products Liability, and the Need for a New Approach*, 1 UTAH L. REV. 437, 440 (2013).

[196] Jack M. Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1234 (2017).

[197] *Id.* (emphasis added).

[198] Sophia H. Duffy & Jamie Patrick Hopkins, *Sit, Stay, Drive: The Future of Autonomous Car Liability*, 16 SMU SCI. & TECH. L. REV. 453, 479–80 (2013).

[199] Julie Goodrich, Comment, *Driving Miss Daisy: An Autonomous Chauffeur System*, 51 HOUS. L. REV. 265, 284 (2013).

[200] Shavell, *supra* note 26, at 243.

[201] *Id.* at 247.

[202] Xuan Di, Xu Chen & Eric Talley, *Liability Design for Autonomous Vehicles and Human-Driven Vehicles: A Hierarchical Game-Theoretic Approach*, TRANSP. RSCH. PART C: EMERGING TECHS., Sept. 2020, at 118.

liability rules for autonomous vehicles and human drivers.[203] They used game theory to simulate examples and investigate the emergence of human drivers' moral hazard, manufacturers' role in traffic safety, and lawmakers' role in liability design, proposing that their model help inform policy.[204]

There are many good ideas among those just presented, and they are worth holding in mind for the discussion that follows the presentation of the experiments. There, I will return to these prescriptive arguments, as the experiments bear upon both the feasibility and the likely persuasiveness of them.

D. Why the Descriptive?

The legal psychology of assessing fault in the wake of socio-technical system failures matters for several reasons. First, it matters for litigation, especially litigation that reaches jury or bench trials. At present, if an accident occurs in the wake of a m2h handoff in a semi-autonomous vehicle, and a party is injured beyond property (vehicle) damage, there likely will be a personal injury lawsuit. In such a suit, jurors' beliefs about liability are vitally important. If the average juror were to perceive the machine—the semi-autonomous vehicle, including its AI, its developers, and the company that created it—as predominantly at-fault, this would change defense and insurance adjusters' thinking, and it would change pre-trial settlement offers. If the average juror were to perceive the human driver as predominantly at-fault, this also would change defense and insurance adjusters' thinking, but in the opposite direction. Madeleine Clare Elish's worry about such a bias was expanded on by Kate Darling when she wrote, "The problem here isn't that our legal systems don't have a solution for this sort of liability—they do. The problem is that our *perception* of fault in these situations is often different. . . . We need to be extremely careful that this bias doesn't let companies deflect legal liability."[205] Just as it matters in civil proceedings, it matters in criminal law in similar ways. When there are multiple defendants, all might be held criminally liable, but no one defendant exists in isolation. Jurors, to the extent that evidence regarding multiple actors is allowed in, take such evidence into account.

Second, lay imputations of fault in the wake of socio-technical system failures matter for oversight and auditing. As I will explore in the experiments below, sometimes judges are informed by algorithmic tools.[206] If racial bias

[203] *Id.*

[204] *Id.*

[205] DARLING, *supra* note 32, at 85.

[206] *See infra* Part II.

were to emerge in judicial decision-making, and that bias could be traced to a specific machine handoff, which entity would the auditors or academic researchers blame? The judge or the algorithm or both? If there is systematic bias in such placements of blame, that bias might stand in the way of developing optimal solutions.

Third, only once we know how lay individuals perceive such events can we formulate appropriate jury instructions or other measures that would further policy prescriptions. Without a baseline from which to judge how and to what extent jurors deviate from the proper understanding of the law, it is nearly impossible to develop instructions that will guide jurors back onto the proper path. Relatedly, as we saw above, legal scholars have presented countless plans for how best to regulate and legislate semi- and fully autonomous vehicles, and more theories regarding other socio-technical systems are emerging as well. Scholars will be successful in promoting their solutions, and legislators in getting those solutions adopted, only if they are attuned to biases and intuitions that may cut against (or in favor of) their desired solutions. Research on lay perception of socio-technical system failures is vital for this.

Lastly, this research is needed because it explores deep issues regarding proximate cause, fault, legal liability, and related concepts. It is an exploration that very much falls within the experimental jurisprudence branch,[207] and its results help us to better understand how these concepts are used in the U.S. legal system.

E. The Psychology of Fault

Who gets blamed when m2h handoffs lead to harm? As noted above, there is a dearth of research in this area. That said, there are two lines of work that are worth discussing, as they directly inform the experiments presented in the succeeding section. First, there is Madeleine Clare Elish's theory of "moral crumple zones."[208] Dr. Elish reviewed a few notable socio-technical system failures, including the nuclear meltdown at Three Mile Island and the crash of Air France Flight 447, and argued that, when such failures occur, responsibility may be misattributed to a human actor who had limited control over the behavior.[209] Her primary argument is that this misattribution of responsibility serves the purpose of protecting the integrity of the system; in other words, the human operator serves as the crumple zone, absorbing the

[207] Knobe & Shapiro, *supra* note 15, at 171.

[208] Elish, *supra* note 8, at 40.

[209] *Id.* at 41–50.

figurative force of impact.[210] Dr. Elish argues that this is a problematic bias, as the entities who possess equal if not greater control over the behavior of a purportedly "autonomous" system are the designers, engineers, and manufacturers.[211]

This is a compelling thesis, and it certainly makes sense when one considers the incentives of the different players. The company that created the system undoubtedly wants to shift blame away from the system and towards individual actors whose behavior can be characterized as anomalous. But the evidence marshalled for this theory is anecdotal, and the theory also does little to account for the psychology of uninvested parties, such as jurors or lay commentators. In this article, I provide data that speaks to these points, but it is worth hesitating now on the psychology: what is the psychology of imputing fault?

The studies in this article were designed to gather initial impressions as to whether, when an accident occurs in the wake of a m2h handoff, the machine or the human is perceived as more at fault. But fault, of course, is a slippery concept. It is related to a central debate in legal scholarship, one that revolves around whether the legal concept of proximate cause[212] is equivalent to the lay/folk concept of causation.[213] There are two opposed camps. In the formalist camp, it is argued that proximate cause is an objective matter that can be determined by descriptive inquiry.[214] For formalists, judges make conclusions regarding proximate cause first, through an analysis of the facts of a matter; only once such analysis is completed do they use those conclusions to make a moral judgment regarding the matter.[215]

In the realist camp, it is argued that proximate cause can be reduced to "responsible cause."[216] When a judge concludes that a defendant caused a plaintiff's harm, this is merely the judge concluding that the defendant is morally and legally responsible for the harm.[217] As Professor Leon Green wrote, "[T]he inquiry while stated in what seems to be terms of *cause* is in fact whether the defendant should be held responsible."[218] In short, realists conclude that causal judgments are determined by moral ones.

Accidents involving socio-technical systems are relatively common. A few examples were given in the introduction to this article. A self-driving

[210] *Id.* at 41.

[211] *Id.* at 42, 45.

[212] *See* sources cited *supra* note 34.

[213] *See* sources cited *supra* note 35.

[214] Knobe & Shapiro, *supra* note 15, at 174.

[215] *Id.* at 179.

[216] KEETON ET AL., *supra* note 37, § 42, at 273.

[217] *Id.*

[218] Green, *supra* note 38, at 605.

Uber strikes and kills a pedestrian—this has happened, and it will happen again in the future. Multiple questions emerge in the aftermath of such accidents.

In their 2021 article, Professors Knobe and Shapiro argue that, when something goes amiss like this, there are three separate but linked questions that fall along the proximate cause spectrum: was the behavior morally right or wrong? Based on this moral judgment, who or what was the proximate cause of the harm? Based on this proximate cause judgment, who is responsible?[219] This is an approach that takes up the middle group between the formalists and realists.[220] Drawing on empirical work and vetting their theory against the results of actual cases, they argue that an initial moral judgment (was the defendant's action morally improper?) influences judgment about proximate cause (did the defendant cause the harm?), which in turn influences judgment about responsibility (should the state hold the defendant liable for the harm caused?).[221]

Professors Knobe and Shapiro were writing about proximate cause in general, not in the specific case of machine failure or socio-technical system failure. Thus, it might not be surprising to realize that applying their triadic analysis to a situation in which the malfeasor is a machine raises illuminating issues. For one, while machines might effectively act like moral actors, very few people would attribute moral capacity or moral agency to machines.[222] A human faced with the trolley problem conducts a moral deliberation.[223] A machine faced with the trolley problem arguably does something quite different.[224] We might say that the developers of the machine morally deliberated, but there are problems even here: with many types of artificial intelligence algorithms, outcomes are not explicitly programmed.[225] In addition, most AIs are designed by teams of developers, and it is not

[219] Knobe & Shapiro, *supra* note 15, at 171.

[220] *Id.*

[221] *Id.*

[222] Yochanan E. Bigman & Kurt Gray, *People Are Averse to Machines Making Moral Decisions*, 181 COGNITION 21, 32 (2018); Patrick Gamez et al., *Artificial Virtue: The Machine Question and Perceptions of Moral Character in Artificial Moral Agents*, 35 AI & SOC'Y 795, 805 (2020).

[223] Lance Eliot, *AI Ethicists Clash Over Real-World Aptness of the Controversial Trolley Problem, but for Self-Driving Cars It Is the Real Deal*, FORBES (July 27, 2020, 11:53 PM), https://www.forbes.com/sites/lanceeliot/2020/07/27/ai-ethicists-clash-over-real-world-aptness-of-the-controversial-trolley-problem-but-for-self-driving-cars-it-is-the-real-deal/?sh=6f34fe105095 [https://perma.cc/D5D5-BZFT].

[224] *Id.*

[225] *See id.*

immediately apparent how moral intention can be established in such a collective.[226]

Professor Michael LaBossiere has argued that there are typically two tests for imputing moral capacity: people either draw on a Kantian notion that bases moral status on rationality and/or they draw on Mill and base moral status on an entity's ability to feel pleasure and pain.[227] The latter approach, which was endorsed by Professor Peter Singer in the context of animal status,[228] seems like a bar too high for machines to reach, although the Kantian conception of rationality is attainable. That said, the Kantian notion likely does not appeal to modern thinkers, as rationality is increasingly equated with non-human, non-moral existence: hence the exhortation, "don't be a robot."[229]

To provide more context, consider that legal decision-making is often deemed a moral undertaking, and people tend to prefer that machine decision-making is limited to mechanical tasks.[230] At its most general, we might understand this concern as one relating to a general reluctance to allow machines to make important decisions,[231] but it is more specific than that. When a decision falls within the general category of a moral one, we tend to care how that decision was reached.[232] Everett, Pizarro, and Crockett showed that how people arrive at their moral decisions influences whether those decisions are perceived as morally permissible.[233] Decision-makers were perceived as less trustworthy when their decisions were the result of calculating costs and benefits rather than a deliberative struggle or a demonstrated sensitivity to others' welfare.[234] One reason people do not trust

[226] *See id.*

[227] Michael LaBossiere, *Testing the Moral Status of Artificial Beings; Or "I'm Going To Ask You Some Questions...", in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 293, 294 (Patrick Lin et al. eds., 2017).

[228] PETER SINGER, ANIMAL LIBERATION (1990).

[229] Bruno Jacobsen, *Can AI Make Us More Rational?*, FUTURES PLATFORM: FUTURE PROOF (July 21, 2019), https://www.futuresplatform.com/blog/can-ai-make-us-more-rational [https://perma.cc/89LE-SC87].

[230] Min Kyung Lee, *Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management*, 5 BIG DATA & SOC'Y 1 (2018).

[231] Aaron Smith & Monica Anderson, *Automation in Everyday Life*, PEW RSCH. CTR. (Oct. 4, 2017), https://www.pewresearch.org/internet/2017/10/04/automation-in-everyday-life/ [https://perma.cc/73WM-HF6U].

[232] Jim Everett, David A. Pizarro & Molly J. Crockett, *Inference of Trustworthiness from Intuitive Moral Judgments*, 145 J. EXPERIMENTAL PSYCH. 772 (2016).

[233] *Id.*

[234] *Id.*

autonomous vehicles is because they perceive the vehicles as lacking moral capacity.[235]

Whereas humans possess highly complex minds capable of nuanced feelings, machines, including artificial intelligence, are perceived as entities that lack nuance and depth of feeling.[236] Weisman, Dweck, and Markman found that inferences of mindedness are more readily granted to decision makers who have the ability to value others' feelings and understand the moral ramifications of their behavior.[237] Thus, if a machine decision maker is perceived as lacking a mind, people may view its decisions as morally deficient. Professor Eugene Volokh has argued that this is one of the concerns that makes people hesitant to adopt robot judges.[238] Professors Kerr and Mathen also have expressed concerns along these lines.[239]

While we can safely assume that machine decision-makers are, on average, perceived as less morally capable than human decision-makers, the truth of this assertion is not overly important here, as I will measure perceived moral capacity of both the human and machine actors presented in the second set of experiments.[240]

This moral capacity issue is important because there has been much research documenting the impact of moral judgments on causal judgments.[241] In essence, our causal judgments are influenced by our judgments as to whether someone has done something wrong. Consider this vignette from an experiment conducted by Professors Knobe and Fraser:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has

[235] April D. Young & Andrew E. Monroe, *Autonomous Morals: Inferences of Mind Predict Acceptance of AI Behavior in Sacrificial Moral Dilemmas*, 85 J. EXPERIMENTAL SOC. PSYCH. 103870 (2019).

[236] Heather M. Gray, Kurt Gray & Daniel M. Wegner, *Dimensions of Mind Perception*, 315 SCIENCE 619 (2007).

[237] Kara Weisman, Carol S. Dweck & Ellen M. Markman, *Rethinking People's Conceptions of Mental Life*, 114 PROC. NAT'L ACAD. SCI. 11374 (2017).

[238] Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1189–90 (2019).

[239] Ian Kerr & Carissima Mathen, Chief Justice John Roberts is a Robot, 38–39 (2014) (unpublished manuscript), http://robots.law.miami.edu/2014/wp-content/uploads/2013/06/Chief-Justice-John-Roberts-is-a-Robot-March-13-.pdf [https://perma.cc/CQT4-CBLV] (The authors express uncertainty about an AI judge's "imagination, and capacity, to perceive the moral underpinnings of its community.").

[240] *See infra* Part II.C.

[241] *See* sources cited *supra* note 41.

> repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message . . . but she has a problem. There are no pens left on her desk.[242]

After reading these vignettes, participants on average concluded that Professor Smith had caused the pen shortage, and the administrative assistant had not.[243] Even though both had taken pens, and it was obvious that both literally caused there to be a lack of pens, because Professor Smith did something morally blameworthy, he was considered more of a cause of the problem.[244]

But what happens to imputations of causation if an entity *cannot* be morally blameworthy? In the case of human-machine collaboration, if a machine has no moral capacity and cannot be thought of as morally blameworthy, then we might assume that this will lead to lowered causal inferences. After all, in the wake of an accident, the machine did not do anything wrong; it was just being a machine, even if it was being a suboptimal or even broken machine. But the behavior of the human collaborative partner surely can be considered morally wrong. Thus, we might have a systematic bias in fault attributions when socio-technical systems lead to harm. That is, human operators will be over-faulted. In the studies that follow, I test this hypothesis.

II. THE EXPERIMENTS

Since this is the first-ever attempt to empirically test this important topic, the scope should be relatively narrow. I have narrowed it to handoffs within socio-technical systems and ones that seem to fall within the gray area between clear human failure and clear machine failure. With this in mind, in this section, I relay the results of two sets of original experimental studies investigating attributions of fault in handoff scenarios, with a focus on m2h handoffs.

The studies were approved by Princeton University's Institutional Review Board. The participants were recruited through Prolific, an online platform

[242] Joshua Knobe & Ben Fraser, *Causal Judgment and Moral Judgment: Two Experiments*, *in* 2 MORAL PSYCHOLOGY: THE COGNITIVE SCIENCE OF MORALITY: INTUITION AND DIVERSITY 441, 443 (Walter Sinnott-Armstrong ed., 2008).

[243] *Id.*

[244] *Id.* at 443–44.

for human intelligence tasks.[245] Like any source of participants for human intelligence tasks, Prolific has its limitations,[246] but such types of sources, such as Amazon's Mechanical Turk, have found widespread acceptance and support in the academic community.[247] Separate samples of participants were used for the different experiments, and participants were blocked from participating in more than one of the experiments. Because I was interested in U.S. lay decision-makers, all of the participants were jury-eligible U.S. citizens and current U.S. residents. Data analysis was performed using the R software/programming language.[248]

A. Study 1: Bias in Machine-to-Human Handoff Fault Determinations

Study 1 was designed to gather initial impressions as to whether, when an accident occurs in the wake of a m2h handoff, the machine or the human is perceived as more at fault. The basic design of Study 1 is that participants were told about an accident involving a semi-autonomous vehicle, and they were asked who was at fault, the human driver or the AI.[249] There were sixteen different accident types, and they systematically varied on dimensions relating to problems inherent to semi-autonomous vehicles, including human driver skill atrophy, timeliness of the handoff, and difficulty of the situation. All of the accident prompts described a m2h handoff: the AI instructed the human to take over the driving prior to the accident occurring. I hypothesized that, regardless of the specifics of the accident, the human driver would be perceived as more at fault.

This experiment is novel, as there are no other experiments or data regarding such handoffs and subsequent fault attributions. Given the novelty, I kept the primary analysis relatively simple, reserving analyses of alternate conditions for the succeeding sections of this article. That is, in Study 1, the primary dependent variable asks about the two primary players: the human driver and the AI that operates within the semi-autonomous vehicle. In the alternate conditions, instead of asking about the AI, I ask about the developers

[245] *See* PROLIFIC.CO, https://www.prolific.co [https://perma.cc/X8DF-DX3H].

[246] Krin Irvine, David A. Hoffman & Tess Wilkinson-Ryan, *Law and Psychology Grows Up, Goes Online, and Replicates*, 15 J. EMPIRICAL LEGAL STUDS. 320, 326 (2018).

[247] *See* Michael Buhrmester, Tracy Kwang & Samuel D. Gosling, *Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?*, 6 PERSP. ON PSYCH. SCI. 3, 5 (2011); Krista Casler, Lydia Bickel & Elizabeth Hackett, *Separate but Equal?: A Comparison of Participants and Data Gathered via Amazon's Mturk, Social Media, and Face-to-Face Behavioral Testing*, 29 COMPUTS. HUM. BEHAV. 2156, 2158 (2013).

[248] R version 3.6.2, R FOUNDATION FOR STATISTICAL COMPUTING (2019), https://www.R-project.org/ [https://perma.cc/T8H3-TR3X].

[249] *See infra* Part II.A.1.

of the AI and the company that created the AI. In addition, in Study 1, the dependent variable asks about fault: who is at fault? In a variant of the study that follows, I ask about the constellation of concepts that come near to fault (such as proximate cause, legal liability, moral culpability, and so on).

1. Methodology: Participants and Design

To determine the sample size, I considered the primary analyses I would run, which were one-sample *t*-tests. At a significance level of 0.05 with power of 0.8, and for a medium effect size, 33 participants were needed. As I anticipated medium to large effects, I aimed for thirty participants. These participants were 47% male and 43% female, with three respondents selecting "Gender Variant/Non-Conforming." They ranged from 19 to 55 years old, with an average age of 34.2 years. All were U.S. citizens who had been born in the U.S., were currently residing in the country, and indicated that they were jury-eligible.

Study 1 introduced participants to autonomous vehicles, which it defined as vehicles in which an artificial intelligence makes most of the driving decisions.[250] It stated that one company in particular had developed an AI that had proven itself in testing and in actual road driving, and was considered ready for the task. Participants then were told about handoffs, that, when there is something beyond the AI's capabilities, the human driver is instructed to take over the driving.

After this introduction, participants were told that there recently were a few accidents involving the vehicles produced by this same company. At this point, participants were presented with reports of accidents involving these vehicles, and they were told to attribute fault however they thought appropriate.

There were sixteen different reports of accidents, and they were presented in random order. Each report reflected a different combination of three attribute categories.[251] The attribute categories were (1) skills atrophy, (2) situational difficulty, and (3) lag from handoff to accident. Skills atrophy refers to how often handoffs are made: is the human driver often or seldom given driving control? There were different types of these. First, there was atrophy on a single trip. The trip had either been a long one, and the human had not been given control until the handoff that directly preceded the accident, or the trip had just started when the handoff was made. Second,

[250] *See infra* App.A.I for the full set of prompts used in Study 1.

[251] *See infra* App. A.I for a visualization as to how the attribute categories were distributed across the prompts.

there was atrophy over a longer period of time. The AI either had a history of making multiple handoffs on every trip, or it would make only one or so handoffs per month.

Situational difficulty refers to how daunting was the problem that necessitated the handoff. It was either a minor problem (slightly congested area) or a more significant problem (a torrential rainstorm broke out). Lag from handoff to accident refers to how much time passed from the time of the handoff until the accident occurred. Was it very quick—just a matter of moments or minutes? Or was it more delayed—did the human driver have the wheel for, say, thirty or more minutes when the accident finally occurred?

After reading each accident report, participants answered a single-item dependent measure: "Who is at fault?" The scale ranged from 0 ("definitely the human driver") to 100 ("definitely the AI"), with a score of 50 also labeled ("both are equally at fault"). After completing all sixteen accident prompts, participants answered demographics questions.

2. Results

As predicted, the participants overwhelmingly found the human driver (versus the AI) at fault. This was true for each of the prompt types, and it also was true for all of the prompts when averaged together. For the latter analysis, a one-sample two-tailed t-test yielded $M = 12.53$, $SE = 2.70$, 95% CI [7.01, 18.05], $t(29) = -13.89$, $p < .001$, Cohen's $d = 2.54$, such that participants on average viewed the human driver as relatively more at fault than the AI.[252] For the individual prompts, all means were < 28 on the scale and were significantly different than the midpoint of the scale at $p < .001$. Even after Bonferroni corrections, all p-values were $< .002$.[253]

[252] Note that a Cohen's d of 0.8 or greater is typically considered a large effect size, so the effect found here was extremely large. Daniël Lakens, *Calculating and Reporting Effect Sizes To Facilitate Cumulative Science: A Practical Primer for* T*-Tests and ANOVAs*, 4 FRONTIERS PSYCH. 3 (2013).

[253] *See infra* App. B.I for full statistical outputs for Study 1.

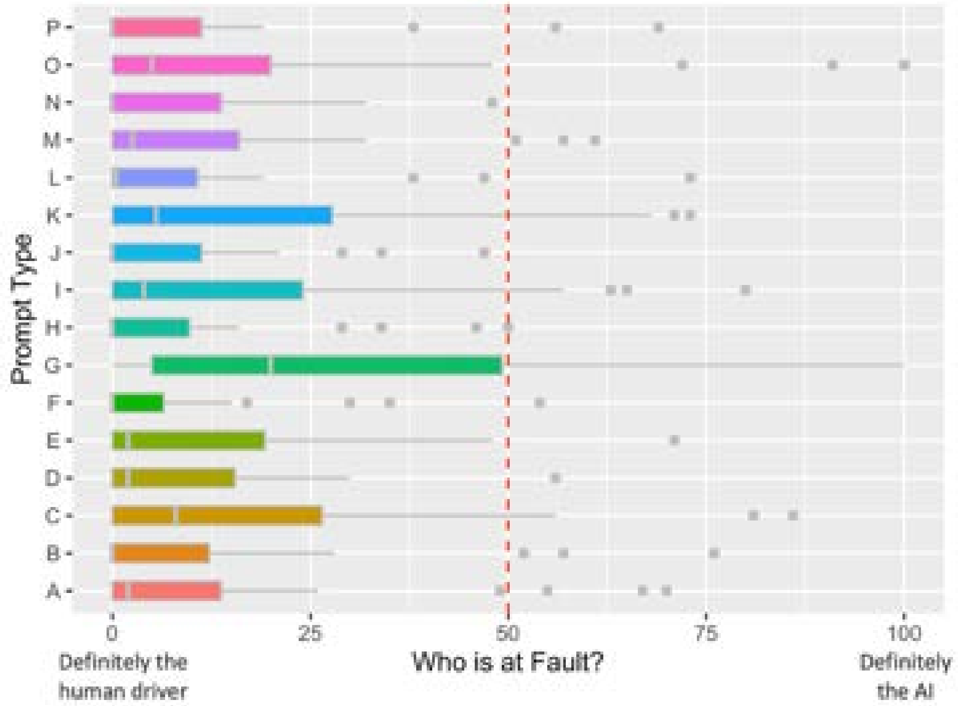


Figure 1. Across all the prompt variants, participants consistently concluded that the human driver was more at fault for the bad outcome following the m2h handoff.

Although the hypothesis proved correct—participants found the human driver more at fault than the AI that had made the handoff—the results still were somewhat surprising. After all, many of the prompts were clear in suggesting that the AI and its process were to blame. Yet, in spite of this, participants still overwhelmingly found the human driver to be more at fault.

While preliminary, this experiment provides strong evidence that, faced with post-m2h handoff harm, individuals are likely to show anti-human operator bias. Granted, there are nuances that are worth exploring further, and these will be broached starting with the next study.

B. Secondary Conditions and Analyses Pertaining to Study 1

While Study 1 was designed to gather initial impressions as to whether, when an accident occurs in the wake of a m2h handoff, the machine or the human is perceived as more at fault, I included a few conditions that answered immediate questions that are important for interpreting the results. In particular, I wanted to expand the referent (to not just the AI but the developers of AI and the company that built the AI) and the dependent variables (to not just fault but proximate cause, legal liability, moral culpability, and other related concepts). These conditions and analysis of the results from them are presented in the succeeding two sections. I hypothesized that, regardless of the condition (AI/developers of the AI/company that built the AI) and regardless of the fault-related dependent variable, responsibility would fall on the human driver.

1. Studies 1A and 1B: The People (or Corporation) Behind the Curtain

In Study 1, the primary dependent variable asked about the human driver and the AI: who was at fault? However, some participants might want to place fault on the developers of the AI, or the company that built the AI, and would perceive the question as being not about these entities but about the AI in isolation. In other words, what happens when we pull back the curtain and let participants find fault in those who created the AI?

a. Methodology: Participants and Design

As these conditions were identical to the primary condition in Study 1, I recruited the same number of participants (thirty in each condition). In Study 1A, the participants were 63% male and 37% female. They ranged from 21 to 67 years old, with an average age of 34.6 years. All were U.S. citizens who had been born in the U.S., were currently residing in the country, and indicated that they were jury-eligible. In Study 1B, the participants were 30% male, 67% female, and 3% gender variant/non-conforming. They ranged from 18 to 60 years old, with an average age of 33.0 years. All were U.S. citizens who had been born in the U.S., were currently residing in the country, and indicated that they were jury-eligible.

The designs were identical to Study 1. Participants read about sixteen different accidents involving a semi-autonomous vehicle, and for each accident they were asked to attribute fault. The only difference from Study 1 was that, in Study 1A, rather than "definitely the AI," the highest point in the dependent variable scale was labeled "definitely the developers of the AI." In Study 1B, the highest point in the dependent variable scale was labeled "definitely the company that created the AI."

b. Results

As predicted, the participants overwhelmingly found the human driver (versus the developers of the AI) at fault. This was true for each of the prompt types, and it also was true for all of the prompts when averaged together. In Study 1A, for the latter analysis, a one-sample two-tailed *t*-test yielded M = 12.45, SE = 2.28, 95% CI [7.79, 17.12], $t(29) = -16.47$, $p < .001$, Cohen's d = 3.01, such that participants on average viewed the human driver as relatively more at fault than the developers of the AI. For the individual prompts, all means were < 37 on the scale and were significantly different than the midpoint of the scale at $p < .001$, with one prompt (prompt G) coming in at $p = .02$. Even after Bonferroni corrections, all *p*-values were < .001, except for G, which rose to $p = .35$.[254]

In Study 1B, for all of the prompts when averaged together, a one-sample two-tailed t-test yielded M = 15.36, SE = 2.85, 95% CI [9.54, 21.18], $t(29)$ = -12.17, $p < .001$, Cohen's $d = 2.22$, such that participants on average viewed the human driver as relatively more at fault than the developers of the AI. For the individual prompts, all means were < 34 on the scale and were significantly different than the midpoint of the scale at $p < .001$, with one prompt (prompt G) coming in at $p = .002$. Even after Bonferroni corrections, all *p*-values were < .001, except for G, which rose to $p = .04$.[255]

In short, the general result was as hypothesized and showed the same pattern as that observed in Study 1: more fault was placed on the human driver than on the developers of the AI and the company that created the AI. I will withhold full discussion of potential reasons for this result until Section IV, but I will venture brief remarks here. There appear to be only two explanations: first, participants might view the AI and its developers and/or parent company as one composite entity, and thus asking about one is equivalent to asking about the others or about all. Second, participants might have similar psychological orientations to a collection of individuals (especially when collected into a corporate unit) as they do to a machine.

2. Study 1C: Fault and Related Sins

Studies 1A and 1B should have increased our confidence in Study 1's results, as the conditions showed nearly identical effects for the parties related to the machine making the handoff, i.e., the developers and the corporate entity. In Study 1C, I prodded the outcome measure—fault—along the lines discussed in Part I of this article. What happens when we ask not about fault but about proximate cause and related concepts? Is it true that, as I hypothesize, lay individuals collapse such concepts into a murky whole

[254] *See infra* App. B.II for full statistical outputs for Study 1A.

[255] *See infra* App. B.III for full statistical outputs for Study 1B.

embodied by "fault"? Or do the terms lead to meaningfully distinct results in this specific application?

While the terms I used are presented below in the Methodology section, I would like to discuss the inclusion of a few of them. In addition to fault, proximate cause, and factual cause, I thought it prudent to test ordinary cause—that is, simply ask, which entity caused the accident? This gets closer to common language usage and intuitive understanding of causation. Second, I asked about morally wrong behavior, as this is important to legal realism, and I also asked about blame, since scholars have indicated a link between moral culpability and blame.[256] I included both legal liability and legal responsibility even though they are near-synonyms. If we can distinguish them, we might say that legal responsibility has more to do with duties.[257] I also included the near-synonyms of norm violation and abnormality, as Professors Knobe and Shapiro used both of these terms in their recent work.[258]

a. Methodology: Participants and Design

As this iteration of the study was identical to the other Study 1 designs, I recruited approximately the same number of participants (thirty-five).[259] These participants were 43% male, 54% female, and 3% were transgender male. They ranged from 18 to 72 years old, with an average age of 33.0 years. All were U.S. citizens who had been born in the U.S., were currently residing in the country, and indicated that they were jury-eligible.

The initial design and the prompts were identical to those presented in Study 1. However, instead of reading about all sixteen different accidents involving a semi-autonomous vehicle, participants read about just one accident each. After reading about the accident, they were asked fourteen questions that related to fault. The questions, which were presented in random order, asked about fault itself, proximate cause, factual cause, ordinary/folk cause ("Who caused the accident?"), blame, legal liability and responsibility, norm violation, moral culpability, and supersedence. Specifically, the questions were as follows:

- Who was at fault?
- "Proximate cause" refers to an action that is legally sufficient to find the defendant liable. [As an extreme example, consider a mother who gave birth to a person who then robbed someone forty-

[256] MICHAEL S. MOORE, CAUSATION AND RESPONSIBILITY: AN ESSAY IN LAW, MORALS, AND METAPHYSICS 33 (2009).

[257] KEETON ET AL., *supra* note 37, § 42, at 273.

[258] Knobe & Shapiro, *supra* note 15.

[259] I recruited slightly more because, under this alternate design, slightly more were needed to achieve multiple participants per prompt-type.

five years later. The mother's giving birth to the robber is not a proximate cause of the robbery.] In the accident you just read about, who was the proximate cause?

- "Factual cause" refers to an action that causes an event. In other words, the event would not have happened had the action not been performed. In the accident you just read about, who was the factual cause?
- Who caused the accident?
- Who deserves the blame?
- Who should be held legally liable?
- Who should be held legally responsible?
- In relation to this accident, would you say that the human driver violated norms for behavior?
- In relation to this accident, would you say that the AI (including its developers and/or parent company) violated norms for behavior?
- In relation to this accident, would you characterize the behavior of the AI (including its developers and/or parent company) as abnormal?
- In relation to this accident, would you characterize the behavior of the human driver as abnormal?
- Was the human driver's behavior morally wrong?
- Was the behavior of the AI (including its developers and/or parent company) morally wrong?
- In your opinion, did the behavior of the human driver supersede any negligence by the AI (including its developers and/or parent company)?

The primary seven items (fault, proximate cause, factual cause, ordinary cause, blame, legal liability, legal responsibility) were graded on a 0 to 100 scale, where 0 = "definitely the human driver," 50 = "both equally," and 100 = "definitely the AI (including its developers and/or parent company)." The remaining seven items were graded on a 0 to 100 scale where 0 = "definitely not," 50 = "unsure," and 100 = "definitely yes." After completing all fourteen questions, participants answered demographics questions.

b. Results

As predicted, the full constellation of fault-like terms came out in the same way as fault itself. The primary seven items were virtually identical: whether it was fault or blame or proximate cause or any of the others, the participants attributed it to the human driver.[260] All means were significantly less than the midpoint of the scale, such that participants attributed the item in question to the human driver, with all p-values significant at < .001; after Bonferroni corrections, all $p < .005$. More importantly, a repeated measures ANOVA failed to reveal a significant difference across the measures, such that they were not meaningfully different from each other: $p = .40$ and ges = .005. Post-hoc comparisons using two-tailed t-tests and Bonferroni's corrections also failed to find a difference across the measures, with all $p > .99$.

For the remaining seven items,[261] six of them were paired (AI versus human on separate scales): behavior violated norms; behavior was immoral; behavior was abnormal. For the normativity, abnormality, and the morality questions, paired t-tests showed a significant difference, with $p = .004$, $p = .04$, and $p = .01$, respectively.[262] Importantly, an ANOVA with post-hoc t-tests corrected with Bonferroni's corrections failed to find a significant difference across the three human results (all $p > .99$) and failed to find a significant difference across the three AI results (all $p > .99$). Finally, participants believed that the human driver's behavior superseded that of the AI, with a mean of 65.69, which was significantly greater than the midpoint of the scale, with $SE = 4.96$, 95% CI [55.60, 75.77], $t(34) = 3.16$, $p = .003$, Cohen's $d = .53$.

In all, Study 1C affirmed that there was little, if any, difference in application of these terms. When participants evaluate behavior that leads to harm, they appear to collapse distinctions in the different terms. I will discuss this in greater detail in Section IV; but, for now, we should feel confident using "fault" as an intuitive folk concept that stands in well for the more nuanced legal concepts, including proximate cause, about which we are also interested.

C. *Study 2: Disambiguating the Effect Across Machine-to-Human and Human-to-Human Handoffs*

While Study 1 was designed to gather initial impressions as to whether, when an accident occurs in the wake of a m2h handoff, the machine or the human is perceived as more at fault, Study 2 extends the design and what conclusions can be drawn. First, it prods the question of whether Study 1's results reflect perceptions of handoffs in general, rather than perceptions of machine handoffs in particular. That is, I added a condition in which the

[260] *See infra* App. B.IV for full statistical output for Study 1C.

[261] *Id.*

[262] After Bonferroni corrections, these p-values rose to .01, .11, and .04, respectively.

handoff is made not by a machine but by a human. Thus, we are able to compare m2h and human-to-human (h2h) handoffs. I hypothesized that there would be systematic bias against the human operator after a machine handoff, but that this bias would not be present when the handoff was made by a human.

Second, in the preceding sections of this article, I discussed the nexus between moral conclusions and causal ones. Socio-technical systems are interesting because their failures result from behavior by some actors with moral capacity (i.e., humans) and some that lack it (i.e., machines). Thus, I hypothesized that, in the wake of socio-technical system failures, human actors, especially human operators—unified moral actors who are close to the scene of the harm—will be systematically over-faulted. In Study 2, I included a moral capacity scale in order to test this hypothesis.

Third, in examining algorithm-aided judicial decision-making that lead to racial bias, Study 2 examines a type of handoff that is important in at least two respects: it involves the vital issue of racial equality, and it involves the application of technology to legal decision-making. Fourth, the design of Study 2 permits investigation of potential solutions to the documented bias; this is explored in Section D.

1. Methodology: Participants and design

To determine the sample size, I considered the primary analyses I would run, which were two-sample *t*-tests. At a significance level of .05 with power of .8, and for a medium effect size, sixty-four participants per cell were needed. I aimed for sixty-five participants per cell, for a total of 130. These participants were 40% male and 58% female, with one respondent selecting "Gender Variant/Non-Conforming" and one respondent indicating that their preferred gender was not listed. The participants ranged from 18 to 70 years old, with an average age of 37.4 years. All were U.S. citizens who had been born in the U.S., were currently residing in the country, and indicated that they were jury-eligible.

Study 2 introduced participants to bail decision-making.[263] Participants were randomly assigned to one of two conditions. In one condition ("m2h condition"), participants were told of a court system that has been using a specially-trained AI to make bail recommendations. The AI had performed well at the task and is considered a true expert.

[263] *See infra* App. A.III for the full set of prompts used in Study 2.

The prompt was tailored to accord with the *State v. Loomis* standard,[264] where it was held that judges can use computer aided risk scores, but the scores cannot comprise judges' sole consideration.[265] In accordance with this, the prompt told participants that there had to be a "handoff." The human judge was ultimately responsible and was required to review each case. The AI's bail recommendation could be only one factor (and not a determining factor) in the judge's review. Participants were also told that, while the average judge in the jurisdiction handles many cases per year, in only a few of these, if any, will a judge not follow the AI's recommendation.

Then participants were presented with a description of a mistake/failure. A twenty-six-year-old Black male had been arrested and charged with assault and battery. At the bail hearing, the AI indicated that the defendant appeared to present a "high risk." Instead of $5,000, which arguably would have been more standard in such a case, the AI recommended bail of $40,000. The judge set bail at $40,000.

The defendant could not make bail and wound up spending a significant amount of time behind bars. However, the charges were ultimately dismissed, and the defendant was released. Three years passed, and the defendant was not arrested for any subsequent offenses. Researchers who were auditing court performance found this case, and they concluded that the seemingly improper bail decision possibly was the result of racial bias.

The other condition ("h2h condition") was identical, except that, instead of an AI, there was a human providing the bail recommendation. The conditions were otherwise identical: for instance, when the AI was described as "extremely good in testing and in actual cases and is considered a true expert at the task," the human was described as "extremely good in testing and in actual cases and is considered a true expert at the task."

After reading the prompt, participants completed the following dependent measures. First, "Who is at fault?" The scale ranged from 0 ("definitely the judge") to 100 ("definitely the AI [definitely the human expert]"), with a score of 50 also labeled ("both are equally at fault"). Second, "To what extent are each of the following at fault?" There were two scales, which ranged from 0 ("Not at all at fault") to 100 ("Completely at fault"), one for the judge and one for the AI/human expert. Then participants were asked to justify the answers they gave. Lastly, they were asked about the extent to which they agreed that the judge [AI/human expert] had the moral capacity necessary for the task. The scale ranged from 0 (complete disagreement with the statement)

[264] 881 N.W.2d 749 (Wis. 2016).

[265] *Id.* at 768–70.

to 100 (complete agreement). To finish, participants answered demographics questions.

2. Results

As predicted, in the m2h condition, the participants overwhelmingly found the human decision maker—i.e., the judge—more at fault than the AI that made the handoff. A one-sample two-tailed *t*-test yielded *M* = 37.58, *SE* = 2.86, 95% CI [31.87, 43.29], *t*(64) = -4.34, *p* < .001, Cohen's *d* = .54, such that participants on average viewed the human decision maker as relatively more at fault than the AI (a score of 50 would have meant that participants viewed the human and the AI as equally at fault). This result replicates Study 1, with the effect now found in a new context—bail decision-making—where before it was found in semi-autonomous driving.

Also as predicted, in the h2h condition, the participants failed to find the judge more at fault than the person making the handoff. A one-sample two-tailed *t*-test yielded *M* = 46.78, *SE* = 2.76, 95% CI [41.28, 52.29], *t*(64) = -1.17, *p* = .25, Cohen's *d* = .14. More importantly, this result was significantly different than the m2h result. A two-sample two-tailed *t*-test yielded *t*(128) = -2.32, *p* = .02, Cohen's *d* = .41, such that the judge was found to be significantly more at fault in the m2h condition than in the h2h condition.

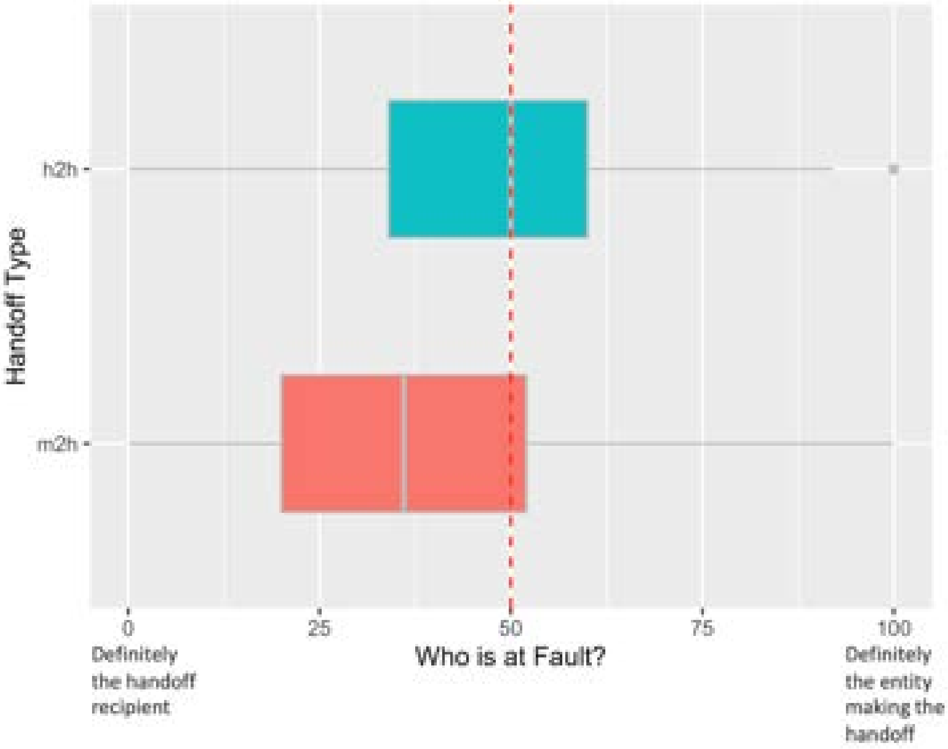


Figure 2. In the m2h handoff condition (labeled "m2h"), participants concluded that the handoff recipient (the judge) was relatively more at fault for the bad outcome than the entity making the handoff (the AI in the m2h condition). In addition, the judge in the m2h condition was considered significantly more at fault than the judge in the human-to-human (labeled "h2h") handoff condition.

As one would expect, the human decision makers were perceived as possessing greater moral capacity than the AI, with means for the judge, the human expert, and the AI of 63.43, 50.49, and 17.62, respectively. Across the different actors, there was a significant correlation between the perceived moral capacity of the entity and the extent to which that entity was considered at fault. A linear regression model yielded the following: β = .11, *t*(258) = 5.48, *p* = .02, such that greater moral capacity correlated with greater fault. In other words, the effect shown in Figures 1 and 2 might be explained by moral capacity, an explanation that aligns with the research cited in the preceding sections of this article. There is compounding circularity that leads causal conclusions and moral ones to move in tandem. When an entity lacks the capacity to be morally responsible, this might in fact lead to lowered imputations of causal responsibility. This is shown quite clearly in Figure 2, where the machine making a handoff is deemed significantly less at fault than the human making a handoff.

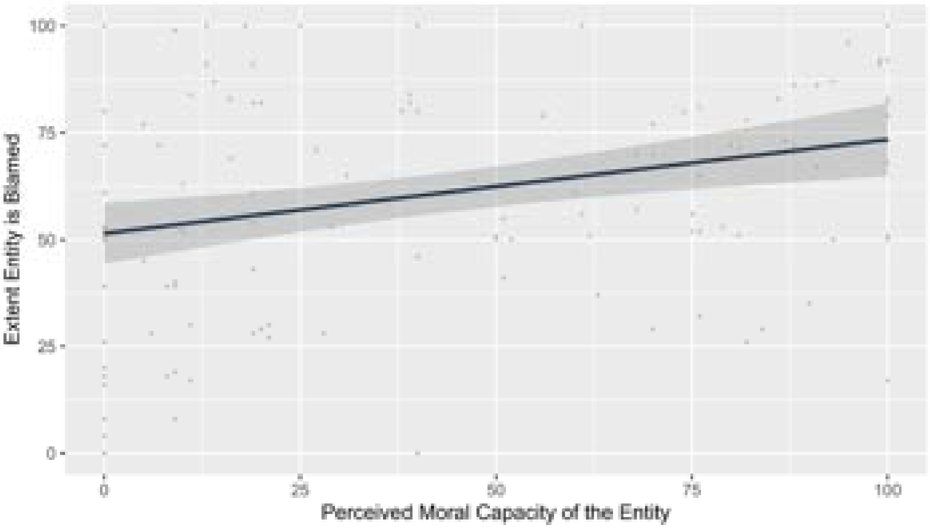


Figure 3. Plot of model of perceived moral capacity and imputation of fault. The shaded region is the 95% confidence interval (p = .02).

D. Secondary Condition Pertaining to Study 2: A Potential Solution

In the preceding sections of this article, I discussed an unequivocal finding in behavioral sciences and human factors research: handoff recipients are placed at a distinct disadvantage, and it is hard for them to perform up to the typical standard-of-care. There are many reasons for this, with the primary ones being automation complacency, inattention, skill atrophy, and automation bias (i.e., over-trust).[266] It is hard for individuals to perform well when suddenly given control,[267] and it is hard for them to spot mistakes in the information that machines provide to them prior to a handoff.[268] Related to this second point, it is psychologically difficult for a human to override machine suggestions: even medical doctors will ignore their own instincts to defer to machines' recommendations.[269]

[266] RAPOSO ET AL., *supra* note 44, at 45–56.

[267] Agrawal et al., *supra* note 122, at 132; Eriksson & Stanton, *supra* note 120, at 701–02; Gold et al., *supra* note 120, at 12.

[268] DAVIES & PARASURAMAN, *supra* note 104; NAT'L TRANSP. SAFETY BD., *supra* note 97, at 44.; Parasuraman & Manzey, *supra* note 90, at 397; Parasuraman et al., *supra* note 91.

[269] *Automation Bias 2012*, *supra* note 107, at 121; *see also Automation Bias 2014*, *supra* note 111, at 373.

It likely is the case that the average jury-eligible individual is not aware of this research and does not comprehend the disadvantage the handoff recipients in Study 2 faced. In Study 2A, which I present here, I hypothesized that understanding of the handoff disadvantage would largely eliminate the systematic bias found in Study 2.

1. Methodology: Participants and design

To determine the sample size, I considered the primary analyses I would run, which were two-sample *t*-tests. At a significance level of .05 with power of .8, and for a medium-to-large effect size, about forty participants per cell were needed, so I aimed for a total of eighty across the two conditions. These participants were 38% male and 60% female, with one participant who was transgender female and one respondent who selected "Gender Variant/Non-Conforming." The participants ranged from 18 to 70 years old, with an average age of 39.1 years. All were U.S. citizens who had been born in the U.S., were currently residing in the country, and indicated that they were jury-eligible.

The design was the same as Study 2 except for one change. Instead of one m2h condition and one h2h condition, only the m2h condition was presented, and now there were two variants of it: one was identical to that which appeared in the primary Study 2, and the other also was identical to that which appeared in the primary Study 2 but now included an additional information portion that told participants of the research showing how handoffs disadvantage handoff recipients. As the study was designed to test the extent to which information might lessen the anti-human operator bias, the information presented was clear in stating that handoffs set the human recipient up for failure. The implications of this design for, say, jury instructions are discussed in the succeeding sections of this article.

2. Results

As predicted, and as a replication of Study 2, in the m2h condition with no additional information about handoffs provided, the participants overwhelmingly found the human decision maker—i.e., the judge—more at fault than that AI that made the handoff. A one-sample two-tailed *t*-test yielded $M = 38.28$, $SE = 2.77$, 95% CI [32.67, 43.88], $t(39) = -4.23$, $p < .001$, Cohen's $d = .67$, such that participants on average viewed the human decision maker as relatively more at fault than the AI (a score of 50 would have meant that participants viewed the human and the AI as equally at fault).

Also as predicted, in the m2h condition with the additional information about handoffs, the effect disappeared. The participants failed to find the judge more at fault than the person making the handoff. A one-sample two-tailed *t*-test yielded *M* = 49.43, *SE* = 4.41, 95% CI [40.50, 58.35], *t*(64) = -.13, *p* = .90, Cohen's *d* = .02.[270] More importantly, this result was significantly different than the m2h result. A two-sample two-tailed *t*-test yielded *t*(78) = -2.14, *p* = .04, Cohen's *d* = .48, such that the judge was found to be significantly more at fault in the m2h without handoff information condition.

E. Summary of Key Empirical Findings

This article's studies reveal systematic bias—what I am calling "fumble bias"—in imputations of fault following socio-technical system failures. When such systems feature m2h handoffs, as most do, the human operator (i.e., the handoff recipient) receives the bulk of the fault, even in scenarios in which the operator's performance is disadvantaged and unlikely to meet any relevant standard-of-care. In short, such systems set the human operator up for failure—and lay individuals are more likely to place fault on the human than on the machine, its developers, or the company that is responsible for it.

The studies showed that fumble bias appeared regardless of which term in the fault lexicon was used: fault, proximate cause, factual cause, blame, legal liability, norm violation, moral failure, and so on. Moreover, the studies showed that fumble bias was not owing to handoffs in general: when a human made a handoff to a human, fault was more equally distributed across the two players. The implication, which was supported by data from the studies, is that the perceived moral capacity of the entities correlates with perceived fault. The final study presented above showed that education about handoffs might partially correct for this bias, although more research along this line is needed.

These results are important for multiple reasons. First, they provide evidence as to how liability in the wake of machine failures is likely to be attributed by jury-eligible individuals. There was no previous data or empirical research on this, and these results provide much-needed information for litigators, legislators, and scholars, as discussed below. It also

[270] I noted that the standard error was higher in this condition than in the no information condition, which suggests greater variance in participants' responses. One possible explanation is that the information can lead to extreme opinions: some participants, in learning that handoff recipients typically do not perform well, might blame the recipients even more. At the same time, other participants—and, indeed, a majority of participants, given the overall results—might have the opposite response, blaming the AI and the system designers more than the handoff recipient. At the very least, this difference in standard errors should warrant follow-up research.

shows that tech companies might have the upper-hand in the initial wave of these cases that make it to courts. Second, and equally importantly, the results provide key data in the rapidly developing field of experimental jurisprudence, especially that which focuses on proximate cause.[271] A nexus between moral and causal conclusions has been well-established, but no one has explored what happens to this nexus when one entity, such as a machine actor, lacks moral capacity. The result, as shown in Study 2 above, is that fault is shifted onto the remaining actor who presents a unified morally capable face: the human operator in the standard m2h handoff accident. Third, in addition to advancing legal, psychological, and philosophical scholarship in this area, this finding should give human operators everywhere pause, and it should encourage much activity by both litigators, especially defense attorneys, and legislators who care about shaping a tort system that properly punishes and incentivizes.

III. Explanations and Mechanisms

What led to the systematic bias observed in these studies, and what do the results mean for litigation, for legal scholarship on proximate cause, and for the host of prescriptive proposals discussed above? This part of the article unpacks the results—including the role played by perceived moral capacity—to gain insight into how lay decision-makers construe fault in the context of socio-technical system failures. While the findings suggest that lay understanding might be shaped so as to lessen the systematic bias, they also suggest that there are aspects of personhood that reach beyond legal classification, and the extent to which one perceives personhood in another holds significant meaning for legal outcomes. It bears emphasizing, however, that this article is just a starting point in what likely will be an area of great interest and research over the ensuing decade. While we may glean some understanding from the studies presented above, more research is needed, and I offer directions for future empirical work in the final part of this article.

A. Lay Conflation of Moral Capacity and (Moral) Culpability

Across the series of studies, the primary result was clear: when harm results from a socio-technical system failure—in particular, when harm results in the wake of a m2h handoff—our intuition is to blame the human operator. This was found in the context of semi-autonomous vehicles, and it was found in the context of expert decision-making, such as judicial bail

[271] *See, e.g.*, Knobe & Shapiro, *supra* note 15*,* at 179.

decisions. It was clearly not a product of handoffs in general, as the effect disappeared when the handoff was human-to-human. So what explains it?

As discussed above, Dr. Elish thinks that this systematic bias, which she hypothesized explains responses to large-scale and publicly-visible socio-technical system failures, serves the purpose of protecting the integrity of the system; in other words, the human operators serve as the crumple zone, absorbing the figurative force of impact.[272] This explanation is less compelling in the studies presented above, as there is no reason to expect lay individuals to want to protect the integrity of semi-autonomous vehicles or of algorithm-assisted judicial decision-making. That said, the studies above do empirically affirm, for the first-time ever, the existence of the phenomenon that Dr. Elish hypothesized: the human crumple zone, the bias towards placing fault on the human operator who just happens to be nearest to the socio-technical system failure.

But what might explain this? In Study 1C, the participants consistently showed fumble bias, regardless of whether the outcome measure was fault, proximate cause, legal liability, or a host of other related terms. We might conclude that lay individuals fail to distinguish these terms; alternatively, we might conclude that they distinguish the terms but reach the same conclusion regarding how they should be attributed. Either way, it brings us to the central debate in legal scholarship revolving around proximate cause. What does it mean to be the proximate cause of a harm? Is it, as formalists aver, something objective that can be determined from analysis of the facts of a matter?[273] Is it, as realists aver, something closer to a legal or moral conclusion, perhaps reducible to "responsible cause"?[274] Or is it, as Professors Knobe and Shapiro argue, something in-between, a complex analysis that takes into account moral and other causes?[275]

In Study 2 and as seen in Figure 3, participants' perception of moral capacity correlated with their fault imputations. In other words, as perceived moral capacity increased, so did perceived fault. The machine entities, of course, were perceived as much less morally capable than the human entities, and thus we might begin to understand from where the systematic bias comes. In essence, the participants were undertaking a proximate cause analysis, one in which they sought the legally responsible party, but such analysis includes a moral conclusion—and vice versa, as the two operate in a loop. When it came to the machine, it may have acted inappropriately, it may have been the factual cause of the accident, but given its lack of moral capacity, it was, so

[272] Elish, *supra* note 8, at 41.

[273] Knobe & Shapiro, *supra* note 15*,* at 179.

[274] KEETON ET AL., *supra* note 37, § 42, at 273.

[275] Knobe & Shapiro, *supra* note 15*,* at 171.

to say, saved from the gallows. The participants looked around for the nearest unified moral actor and this so happened to be the human operator. So, fault and responsibility and all the rest were heaped upon the human operator.

This conclusion is made all the more persuasive when we consider the h2h condition in Study 2. There, a human was positioned in the same place as the machine in the m2h condition, such that a human expert made a handoff to a human operator. In this case, no bias against the human operator was observed as the participants could find two morally capable entities, each partly the factual cause, and thus they split fault across these two entities.

But this raises an interesting question. If fault is readily attributed to a human actor, what about a team of human actors, such as developers and engineers, or a corporate entity? Importantly, in the case of socio-technical system failures, both of these groups are behind-the-curtain. The AI performs. The human operator (the user) performs. The developers of the AI perform in the past, long before an accident or failure occurs. Moreover, there is diffusion of responsibility, as there are many individuals who collectively build the AI, but seldom one unified actor who carries off the work. For corporations, the nexus to moral capacity is equally or even more remote. The corporate entity that produces the AI does not resemble a person with moral capacity, even if a corporate entity reflects a number of persons united in one body for a purpose, and even if corporations might receive legal personhood in some contexts.[276] Thus, the results of Studies 1A and 1B are unsurprising, though still illuminating: in the wake of a m2h handoff, the human operator was considered more at fault than both the developers of the AI and the company that created the AI. The systematic bias persisted. As a hypothetical matter, it seems to me that, in order for fault to make its way back to the developers and/or the company, fault would first have to be found in the AI. The AI has the closest nexus to the accident, as does the human operator, and fault at the nexus must first be found for it to then radiate upstream to the other actors. But, as shown, it is unlikely that fault will be found in the AI so long as it is perceived as lacking moral capacity. This should raise concerns about potential injustice when such cases make their way through the courts, a matter I discuss in greater detail in the next section.

In all, these results show that human-machine collaborative endeavors do create a novel problem. They will increasingly lead to matters of distributed fault where one party is perceived as morally capable and the other party as amoral. This, in turn, will lead to systematic bias in fault attributions, and this

[276] Elizabeth Pollman, *Reconceiving Corporate Personhood*, 2011 UTAH L. REV. 1629 (2011).

bias certainly will flummox the legal system as it seeks to optimize punishment and incentivize behavior.

B. Interventions

If we return briefly to the prescriptive arguments relayed above, a few conclusions should present themselves. First, the studies in this article arguably give strength to those prescriptive frameworks mandating regulatory structures that hold manufacturers liable by default.[277] For instance, Professor Abbott's proposal that liability be based on a negligence standard that treats the autonomous vehicle as a person[278] would likely lead to exacerbation of fumble bias and would lead to miscarriages of justice. The bias towards over-faulting human users, even when those users are set-up to fail by the technical system, suggests that there should be some corrective measures in place to mitigate such harm. Strict liability on manufacturers is certainly one such measure. However, most of the prescriptive arguments that suggest this route are focused on fully autonomous systems where the users are more owners than operators. For the time being, whether it is autonomous vehicles or commercial airflight or judicial decision-making, automation is but one component: there are handoffs, as we now know well, and strict liability in handoff scenarios seems less palatable. After all, the human operator does play a role, and some human operators may be negligent while others are not.

In light of this, perhaps Professor Casey's emphasis on "robot *ipsa loquitur*" and the need to carefully examine advanced data-logging technologies in machines is a good start.[279] But we can assume that capable attorneys will conduct good discovery, and such information will be had and examined; whether or not that leads to better resolutions is hard to say. The final study presented above suggests that it might. There, we saw that presenting individuals with information showing that handoffs disadvantage the recipient, making performance that meets the standard of care much less likely, works to mitigate some of the effects of the anti-human operator bias. As a practical matter, this might be useful for trial strategy, including choice of experts and scope of their testimony. As a more standardized matter, embodied in jury instructions, for instance, it may or may not be a feasible route to take. I discuss the possibility in more detail in the succeeding section.

[277] *See, e.g.*, Abraham & Rabin, *supra* note 18, at 145; Geistfeld, *supra* note 163, at 1674–75; Karnow, *supra* note 153, at 193.

[278] Abbott, *supra* note 193.

[279] Casey, *supra* note 192.

Staying on the intervention that I tested above, though, it is worth discussing what it means. In essence, I shifted the factual cause more fully onto the AI. This, in turn, had a genuine effect on fault attributions. Thus, it implies that, though the factual-moral-proximate loop may be a true loop, it has its limits. If one party is overwhelmingly perceived as the factual cause, this may overcome the fact that the party lacks moral capacity. Then again, this is not saying so much, since if one is overwhelmingly perceived as the factual cause, then we have a case of clear error, one potentially approaching design defect.[280]

IV. PATHS FORWARD

This final part proposes some future directions in which these results might take us. As the first shot in what surely is to be a long struggle to understand liability as new forms of actors emerge, this article sets the stage for much future work. In the first section of this part, I outline what some of these lines of research might include. In the second section of this part, I cover potential reforms that might correct for the systematic bias identified in this article. I have hinted at and briefly mentioned such reforms in the preceding sections; now I close in discussing them more fully.

A. Future Directions

This section highlights further sociopsychological variables and legal doctrines that merit experimental investigation for a fuller understanding of how lay decisionmakers determine legal liability in the wake of socio-technical system failures. Future studies could shed light on the variables, moderators, and mediators at play in such failures, mainly through study of additional scenarios, types of handoffs, and other aspects of socio-technical systems. They also could explore the factors that influence perception of moral capacity, in machines for sure, but in humans with different demographic characteristics as well. These studies might, moreover, explore the effects of group deliberation on liability imputations, as jurors function within the collective jury, and group opinions might very well diverge from individuals' opinions. Further work on perception of machine actors might be explored in other areas of the law, especially criminal law, administrative law, and copyright and intellectual property. Lastly, future work might extend these studies in the direction that I have traveled: marshalling cognitive

[280] *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(b) (AM. L. INST. 1998).

science and experimental jurisprudence to buttress more traditional legal scholarship on liability.

1. Disentangling Variables, Moderators, Mediators

To better understand the extent of the systematic bias documented above, it would be useful for follow-up studies to expand the experimental conditions. While the studies above covered multiple semi-autonomous vehicle scenarios and also judicial decision-making, there are countless other socio-technical systems that warrant testing and exploration. In addition, the present work was limited to handoffs. Socio-technical systems may lead to harm in non-handoff scenarios, and these ought to be vetted. These additional studies might experiment with increasing specificity, since legal matters are specific, limited to the facts presented. Greater specificity would only work to increase the nuance of the findings and determine the extent of the doctrinal reach.

Further work is also needed to probe interventions. In Study 2A, I explored one intervention which may or may not be feasibly put into practice. There are a range of potential ones that can be explored, as I outline in the next section, and these should be tested. The results would shed light on prescriptive routes, of course, but they also would shed increasing light on the mechanisms that drive fumble bias.

One of the more important findings presented above concerned the mechanism: perceived moral capacity and how it relates to imputed liability. One line of research that certainly should follow is the interdisciplinary one that explores human impressions of AI.[281] Most of this work to-date has focused on trust: how do machine attributes, such as a masculine or feminine-sounding voice,[282] contribute to human trust in those machines?[283] But the work also extends to perception of other non-human entities, such as corporations.[284] All of this holds legal significance in the ways that I show above, but this dimension of it has not been fully explored. Imagine if Apple

[281] Zahra Ashktorab et al., *Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes*, 4 PROCS. ACM ON HUM.-COMPUT. INTERACTION, no. 96, 2020, at 1; Pranav Khadpe et al., *Conceptual Metaphors Impact Perceptions of Human-AI Collaboration*, 4 PROCS. ACM ON HUM.-COMPUT. INTERACTION, no. 163, 2020, at 1.

[282] Caitlin Chin & Mishaela Robison, *How AI Bots and Voice Assistants Reinforce Gender Bias*, *in* AI IN THE AGE OF CYBER-DISORDER 82, 82–92 (Fabio Rugge ed., 2020).

[283] Kristin E. Schaefer, Jessie Y.C. Chen, James L. Szalma & Peter A. Hancock, *A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems*, 58 HUM. FACTORS 377–400 (2016).

[284] CHRIS MALONE & SUSAN T. FISKE, THE HUMAN BRAND: HOW WE RELATE TO PEOPLE, PRODUCTS, AND COMPANIES (2013).

were to design a warm, competent, trustworthy-seeming AI, and this AI began to work in collaboration with a human assistant district attorney. If the duo produced biased charging decisions, how would fault be placed?

In addition to exploring perceived moral capacity in non-human entities, there should be more work on the same in humans. Some of this will parallel work on competency, as children and minors will be perceived as less competent and, perhaps, less morally capable. As a scholar who has studied racial bias, I would not be surprised if there are differences in perceived moral capacity that track the demographic features of actors. Experimental scenarios could add more explicit information about the actors to hold nonmanipulated characteristics constant. As an alternative, future work could include self-report measures that ask participants how they conceive of the actors' identity with respect to various demographic dimensions.

Finally, these experiments provide important insight into how individuals place fault, but follow-up work is needed to understand whether and how collective processes influence such judgments. While the literature on how group deliberation impacts individual judgments is notoriously inconsistent,[285] researchers have been able to marshal solid evidence for "group polarization,"[286] with the implication that even more extreme positions might be reached via deliberation. If this were the case with the studies presented above, fumble bias would become even more pronounced, and human operators would face even greater disadvantage.

2. Extensions to Other Areas of the Law

Further work on perception of machine actors might be explored in other areas of the law. While this article focused, in part, on criminal law, there is

[285] *See, e.g.*, Shari Seidman Diamond & Jonathan D. Casper, *Blindfolding the Jury to Verdict Consequences: Damages, Experts, and the Civil Jury*, 26 LAW & SOC'Y REV. 513, 559–60 (1992) (identifying that the impact of group deliberation on jury verdicts may be different across criminal and civil trials); Shari Seidman Diamond, Beth Murphy & Mary R. Rose, *The "Kettleful of Law" in Real Jury Deliberations: Successes, Failures, and Next Steps*, 106 NW. U. L. REV. 1537, 1605 (2012) (showing that jury deliberations can "assist in resolving individual misunderstandings[ ]"); Paula Hannaford-Agor, Valerie P. Hans, Nicole L. Mott & G. Thomas Munsterman, *The Timing of Opinion Formation by Jurors in Civil Cases: An Empirical Examination*, 67 TENN. L. REV. 627, 650–51 (2000) ("[A] substantial proportion of jurors in this study reported changing their minds based on discussions with other jurors during the course of the trial or final deliberations."); HARRY KALVEN, JR. & HANS ZEISEL, THE AMERICAN JURY 488–89 (1966) ("[W]ith very few exceptions the first ballot decides the outcome of the verdict . . . the real decision is often made before the deliberation begins.").

[286] *See* David G. Myers & Helmut Lamm, *The Group Polarization Phenomenon*, 83 PSYCH. BULL. 602, 603 (1976); Cass R. Sunstein, *Group Judgments: Statistical Means, Deliberation, and Information Markets*, 80 N.Y.U. L. REV. 962, 982–84 (2005).

significant work that could be done on prosecutorial perception of machine versus human criminal offenders. As discussed above, the prescriptive work on this is voluminous, and it might be greatly enhanced by better understanding of the tenor of criminal sanction that developers of machines might find themselves facing. Moreover, because negligence and other concepts are slightly differently defined in criminal law as compared with torts,[287] explorations that foreground participants in the different lexicons would help to explore the interaction of such understanding and perception of machine actors. Criminal law would also be a good place to explore post-liability determinations, since criminal sanctions include punishment beyond the financial outcomes of most civil matters. Are human operators punished more severely than the bevy of machine-related actors? I suspect so.

We also might think about how administrative judges and regulators will perceive the conduct of joint human-machine actors. At a minimum, the present results suggest that human operators will fare poorly in such proceedings. Similarly, we know that "inventions" created by AI are increasingly making their way to the patent office, and thus perception of human vs. machine creates potential matters for patent determinations, but also for liability arising from unlawful copying, fraudulent distribution, and other matters pertaining to attribution and integrity.

3. Experimental Jurisprudence

Ordinary reasoning about liability is not morality-free. It also is not consistent: the identity of the party under consideration matters. This article showed this through experimental studies, and future work might likewise move in the direction that I have traveled: marshalling cognitive science and experimental jurisprudence to buttress more traditional legal scholarship on liability. Professors Knobe and Shapiro have been traveling this route,[288] as have others,[289] and I think it is a fruitful way of extending legal scholarship and jurisprudence beyond the confines of law school libraries. There is much to be learned with the tools of cognitive and experimental science.

[287] SANFORD H. KADISH, STEPHEN J. SCHULHOFER, CAROL S. STELKER & RACHEL E. BARKOW, CRIMINAL LAW AND ITS PROCESSES: CASES AND MATERIALS 446–68 (9th ed. 2012).

[288] Knobe & Shapiro, *supra* note 15, at 179.

[289] Roseanna Sommers, *Commonsense Consent*, 129 YALE L.J. 2232 (2020) (using empirical methods to identify the folk understanding of consent).

B. Potential Reforms

One way to address systematic bias in lay imputations of fault following socio-technical system failures would be to target points in the legal process at which the key extra-legal factor—perceived moral capacity—is likely to impact decision-making. To this end, I propose new initiatives in jury instructions that might both clarify the law and preempt bias in juror decision-making.

1. Communication

In Connecticut, the state in which I am physically located while I write this article, jury instructions relating to proximate cause run as follows:

> "Were such (injuries/harm) caused by the negligence of the defendant?" This is called "proximate cause." Negligence is a proximate cause of an injury if it was a substantial factor in bringing the (injury/harm) about. In other words, if the defendant's negligence contributed materially and not just in a trivial or inconsequential manner to the production of the (injury/harm), then (his/her/its) negligence was a substantial factor.[290]

The present article's experiments, which yield a novel data-driven understanding of the intersection of moral capacity, proximate cause, and fault in connection to human-machine liability, suggest that it may be worth reconsidering approaches that are as ingenuous as the above jury instruction. Individuals may suppose that they are merely assessing the facts and determining substantial factorship in an objective manner, but the truth is that much is going on unconsciously or semiconsciously, and this has more to do with the identity of the actor than with the actor's behavior. Does the actor have moral capacity? Depending on the answer, individuals will "see" different fact patterns. In light of this, it might make sense to rewrite jury instructions on proximate cause so that the term is compared with the other related terms. This would be similar to what is currently done at trial by attorneys who explain burdens of proof by contrasting criminal and civil standards.[291] By identifying factual cause, as well as moral blame, and also the more conclusive legal liability, we would be able to articulate proximate

[290] CONNECTICUT JUDICIAL BRANCH CIVIL JURY INSTRUCTIONS § 3.1-1 (2015), https://www.jud.ct.gov/ji/civil/Civil.pdf [https://perma.cc/V59C-H2CV].

[291] *See* James H. Seckinger, *Closing Argument*, 19 AM. J. TRIAL ADVOC. 51, 58 (1995) (defense attorneys emphasize to jurors that the standard used in criminal trials is more relaxed than that used in civil trials); John S. Worden, *The Beast of Burden*, *in* FROM THE TRENCHES: STRATEGIES AND TIPS FROM 21 OF THE NATION'S TOP TRIAL LAWYERS 135, 142–43 (2015).

cause without subsuming the other concepts. This is not without precedent, as Professor Avani Mehta Sood has suggested comparative instructions, emphasizing that the approach is, after all, how law students are trained.[292] Why not train jurors in a like manner? Whether or not this would work to lessen the bias is just speculation at this point, and the efficacy of the proposal would need to be tested.

In a similar vein, it makes sense to include rules of thumb for administrators and regulators. These might function similarly to how special verdict instructions function for jurors,[293] as they would serve the purpose of getting administrators and regulators to focus on behavior and not on entities, on specific items rather than general impressions. The hope is that rulings would then better accord with instrumental goals, especially ones mentioned in the next section.

Another way to pump the procedural brakes, so to speak, is to mirror the tactic implemented by District Judge Carl Rubin in the Bendectin cases. Bendectin, of course, is well-known in legal textbooks because of *Daubert v. Merrell Dow Pharmaceuticals*,[294] the case that featured the drug and set a new standard for admitting expert testimony in federal courts. Setting aside that canonical aspect of the Bendectin litigation, there is another issue that is more apposite to the present article. In numerous epidemiological studies, the Food and Drug Administration found that Bendectin was safe, such that there was no evidence that the drug increased the natural rate of birth defects. However, litigants continued to file suit against the developer of Bendectin, Merrell Dow Pharmaceuticals, Inc., as the cases activated the biases of jurors. In short, jurors' hearts went out to the parents of children born with birth defects, and jurors appeared willing to award damages even when causation was in doubt. To rectify this seeming miscarriage of justice, District Judge Rubin separated the issues: first, the jury was to decide the causation piece (i.e., was Bendectin a cause of the birth defects?). Second, if the causation piece was decided in favor the plaintiffs, then the jury would turn to the liability piece (i.e., was the manufacturer liable and to the tune of how much in damages?).

In light of the results of the present study, an argument can be made that, in the wake of socio-technical system failures, because there is an observed

[292] Avani Mehta Sood, *Attempted Justice: Misunderstanding and Bias in Psychological Constructions of Critical Attempt*, 71 STAN. L. REV. 593, 656–57 (2019).

[293] *See* FED. R. CIV. P. 49 (explaining the procedure for special verdicts in civil trials); Franklin Strier, *The Road to Reform: Judges on Juries and Attorneys*, 30 LOY. L.A. L. REV. 1249, 1262 (1997); Elizabeth G. Thornburg, *The Power and the Process: Instructions and the Civil Jury*, 66 FORDHAM L. REV. 1837, 1853–54 (1998).

[294] 509 U.S. 579 (1993).

anti-human bias, determinations of causation and liability should be separated. This would enable jurors to focus on the human factors issues that beset m2h handoffs, without being swayed by the semiconscious or even nonconscious factor of perceived moral capacity.

2. Regulation

The tort system is concerned with protecting people after they are harmed—making them whole again. Administrative regulation is designed to protect people before they are harmed.[295] In relation to the two examples presented in this article's studies, what might we be designing those systems for? With autonomous cars, it is to increase safety and provide efficiency gains. With algorithm-assisted judicial decision-making, surely it is to increase justice. So, we might think about how best to incentivize the socio-technical systems so as to achieve these aims. We would then divvy-up liability in a way that aligns with our incentive aims.

When it comes to driving, systematic bias that places a majority of fault on the human operator might increase safety (we'll have more vigilant human operators), but that is a doubtful proposition given the weight of evidence showing that humans struggle at monitoring tasks. This also would jeopardize the efficiency goal, as the operators would have to be nearly as vigilant as in the old system in which they handled all the driving. Moreover, letting the AI and its developers and parent company off-the-hook likely stymies us in our aim for safety, as there will be less pressure on these manufacturers to create safer programs and safer handoffs and safer driving all-around.

As for judicial decision-making, if fault is placed on the judge, regardless of how biased (and persuasive) the algorithmic decision aid is, we might have more justice in the sense that judges will carefully monitor and overrule the algorithms, but this is doubtful given that the weight of cognitive science evidence suggests that judicial bias emerges from unconscious factors,[296] and thus an overly active judge will be as biased as judges have always been. In other words, judges would benefit from reliable baselines and less discretion.[297] In addition, justice might be served by efficiency, such that innocent criminal defendants would spend less time in jail, hurt plaintiffs would spend less time without just compensation, and so on. How would

[295] CHARLES H. KOCH, JR. & RICHARD MURPHY, ADMINISTRATIVE LAW AND PRACTICE § 1.12, at 16–17 (3d ed. 2010).

[296] Naci Mocan, *Biases in Judicial Decision-Making*, *in* BIAS IN THE LAW 97–113 (Joseph Avery & Joel Cooper eds., 2020).

[297] Joseph J. Avery & Joel Cooper, *Racial Bias in Post-Arrest and Pretrial Decision Making: The Problem and a Solution*, 29 CORNELL J.L. & PUB. POL'Y 257, 269–71 (2019).

putting more pressure on algorithm developers and parent companies further the justice goals of legally oriented socio-technical systems?

Answers to this question and to the implied questions relating to semi-autonomous vehicles cannot be fully developed in this article. But regardless of which routes of regulation and legislation and other intervention are taken, the studies presented herein raise an interesting conundrum: lay individuals are likely to blame human operators. Said again, it does not take much extrapolation to conclude that individuals will be more in favor of measures that penalize operator error, and they may very well shy away from those that place blame on developers and manufacturers. Those who seek change along these latter lines and wish to create proper incentives for socio-technical system operation, may find more opposition than they expect.

CONCLUSION

This article provides an empirical look at lay perception of fault in socio-technical system failures. When, in the wake of a m2h handoff, harm occurs, a majority of fault is placed on the human operator. This is the case even when the human operator is clearly disadvantaged by the handoff and has little to no chance of performing up to the relevant standard-of-care. Moreover, the bias does not attach to handoffs in general: human-to-human handoffs did not lead to bias against the handoff recipient. Rather, the effect was only observed when the handoff was m2h.

This "fumble bias" has significant implications for civil and criminal proceedings in a class of cases—distributed human-machine responsibility—that are increasingly making their way to the courts. It also suggests hurdles that may present when legislators and regulators attempt to craft laws and regulations to properly incentivize and de-incentivize behavior. While the most salient application pertains to semi- and fully autonomous vehicles, the results impinge the vast array of other socio-technical systems that are increasingly in use, including those that accompany judicial decision-making.

These studies demonstrate the role of moral capacity in subsequent causal conclusions and fault impositions. While prior research showed a link between moral and causal conclusions, the present research explores the novel case of collaborative decision-making in which one party lacks moral capacity (the machine) and the other party has moral capacity (the human). The results should worry humans who find themselves laboring in socio-technical systems, as they are likely to bear the brunt of fault for potential errors. Moreover, the results should inform much future research on the moral-causal nexus. Proximate cause is increasingly being studied, including via empirical methods, as developments in technology have made it a pressing concern. Machines are ubiquitous, working alongside us, often making decisions for us. Sometimes they pass decisions on to us, and we must spring to action. What then? As Philip Larkin, the English poet, wrote of an errant toss into a wastebin: it

> Shows less and less of luck, and more and more
> Of failure spreading back up the arm
> Earlier and earlier, the unraised hand calm[.][298]

Indeed, legal scholars and judges have long wrestled with this question, wondering where to stop in the causal chain, whom or what to blame.

While this article uncovered suboptimality in lay decision-making, it also demonstrated that somewhat straightforward interventions, such as education about performance difficulty, can have an ameliorating effect. Whether a form of this intervention can be meaningfully incorporated into the courtroom, and whether its ameliorating effect would persist there, is an open question and one worth exploring further. In fact, the article sets the stage for much future research along this path of empirical jurisprudence and legal psychology. It is a head-start on a problem that will be ever more pressing in the days and years ahead.

[298] Philip Larkin, *As Bad as a Mile*, *in* COLLECTED POEMS 103, 103 (2003).

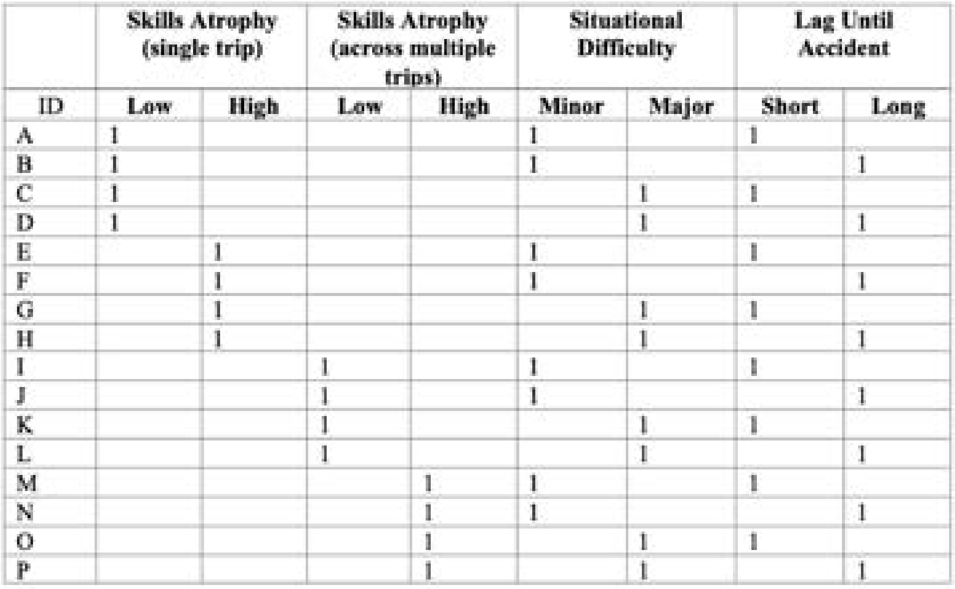
APPENDIX A - STIMULUS MATERIALS

I. STUDY 1

A number of car companies are now developing autonomous vehicles: vehicles in which an artificial intelligence ("AI") makes most of the driving decisions. The AI systems produced by one company, in particular, have proven themselves in testing and in actual road driving and are considered ready for the task. However, when there is something one of these AIs identifies as beyond its capabilities, there is a "handoff," and the human driver is instructed to take over the driving.

There recently were a few accidents involving these vehicles. Please review the following accident descriptions and attribute fault however you deem appropriate. There are no right answers: go with your gut instinct.

[The following—A through P—were presented in random order, and each one was followed by the dependent variable (DV) measure. Also, note that the three attribute categories were distributed across the prompts according to this visualization:]



	Skills Atrophy (single trip)		Skills Atrophy (across multiple trips)		Situational Difficulty		Lag Until Accident	
ID	Low	High	Low	High	Minor	Major	Short	Long
A	1				1		1	
B	1				1			1
C	1					1	1	
D	1					1		1
E		1			1		1	
F		1			1			1
G		1				1	1	
H		1				1		1
I			1		1		1	
J			1		1			1
K			1			1	1	
L			1			1		1
M				1	1		1	
N				1	1			1
O				1		1	1	
P				1		1		1

A. On a local trip to the grocery store, the vehicle had to travel through a slightly congested area of town. The AI had been driving for just five minutes when it indicated that the human should take over. Shortly thereafter, with the human driving, the car got into a fender-bender at a red light.

B. On a quick trip to the hardware store, the vehicle had to travel through a slightly congested area. The AI had been driving for just a minute or two when it indicated that the human should take over. Ten minutes later, with the human driving, the car got into an accident.

C. On a trip to the gym, a very heavy rainstorm broke out, and the AI's sensory system, including its cameras, were obstructed. It told the human driver to take over, and an accident occurred one minute later.

D. On a quick trip to the gym, torrential rain started to fall, and the AI's sensory systems were obstructed. It told the human driver to take over. The human drove for the next 15 minutes before colliding with another vehicle just before arriving at the gym.

E. On a long trip to visit family, the vehicle had to travel through a slightly congested town. The AI had been driving for six hours when it indicated that the human should take over. Shortly thereafter, with the human driving, the car got into a fender-bender at a red light.

F. On a long trip to another state, the vehicle had to travel through a slightly congested area. The AI had been driving for about five hours when it indicated that the human should take over. Thirty minutes later, with the human driving, an accident occurred while trying to exit the highway.

G. While driving from Florida to Georgia, a very heavy rainstorm broke out, and the AI's sensory system, including its cameras, were obstructed. The AI had been driving for hours when it told the human driver to take over, and an accident occurred just moments later.

H. On a trip from Maryland to Virginia, torrential rain started to fall, and the AI's sensory systems were obstructed. The AI had been driving for a long time when it told the human driver to take over. The human drove for the next hour before colliding with another vehicle while changing lanes.

I. One version of the AI–called "Triple J"–makes a lot of handoffs. On a typical trip, it will hand off to the human driver a couple of times. For instance, on a local trip to the grocery store, the vehicle had to travel through a slightly congested area of town. The AI indicated that the human should take over. Shortly thereafter, with the human driving, the car got into a fender-bender at a red light.

J. One version of the AI–called "Triple J"–makes a lot of handoffs. On a typical trip, it will hand off to the human driver a couple of times. For instance, on a quick trip to the hardware store, the vehicle had to travel through a slightly congested area. The AI indicated that the human should take over. Ten minutes later, with the human driving, the car got into an accident.

K. One version of the AI–called "Triple J"–makes a lot of handoffs. On a typical trip, it will hand off to the human driver a couple of times. For example, on a trip to the gym, a very heavy rainstorm broke out, and the AI's sensory system, including its cameras, were obstructed. It told the human driver to take over, and an accident occurred one minute later.

L. One version of the AI–called "Triple J"–makes a lot of handoffs. On a typical trip, it will hand off to the human driver a couple of times. For example, on a quick trip to the gym, torrential rain started to fall, and the AI's sensory systems were obstructed. It told the human driver to take over. The human drove for the next fifteen minutes before colliding with another vehicle just before arriving at the gym.

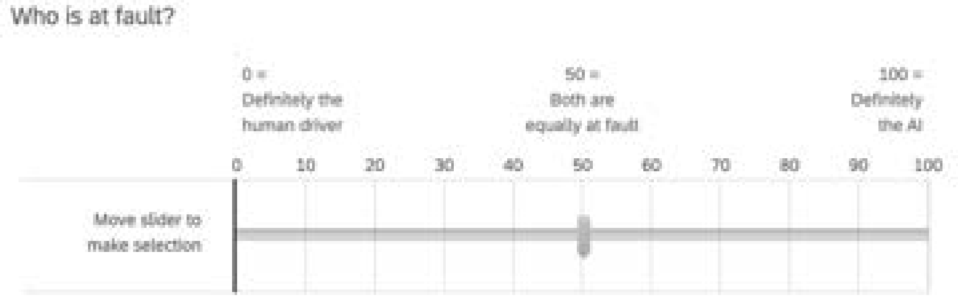
M. One version of the AI–called "Double K"–almost never makes a handoff. Over the past six months, it has averaged less than one handoff per month. However, on a recent local trip to the grocery store, the vehicle had to travel through a slightly congested area of town. The AI indicated that the human should take over. Shortly thereafter, with the human driving, the car got into a fender-bender at a red light.

N. One version of the AI–called "Double K"–almost never makes a handoff. Over the past six months, it has averaged less than one handoff per month. However, on a quick trip to the hardware store, the vehicle had to travel through a slightly congested area. The AI indicated that the human should take over. Ten minutes later, with the human driving, the car got into an accident.

O. One version of the AI–called "Double K"–almost never makes a handoff. Over the past six months, it has averaged less than one handoff per month. However, on a trip to the gym, a very heavy rainstorm broke out, and the AI's sensory system, including its cameras, were obstructed. It told the human driver to take over, and an accident occurred one minute later.

P. One version of the AI–called "Double K"–almost never makes a handoff. Over the past six months, it has averaged less than one handoff per month. However, on a quick trip to the gym, torrential rain started to fall, and the AI's sensory systems were obstructed. It told the human driver to take over. The human drove for the next fifteen minutes before colliding with another vehicle just before arriving at the gym.

- DV Measure

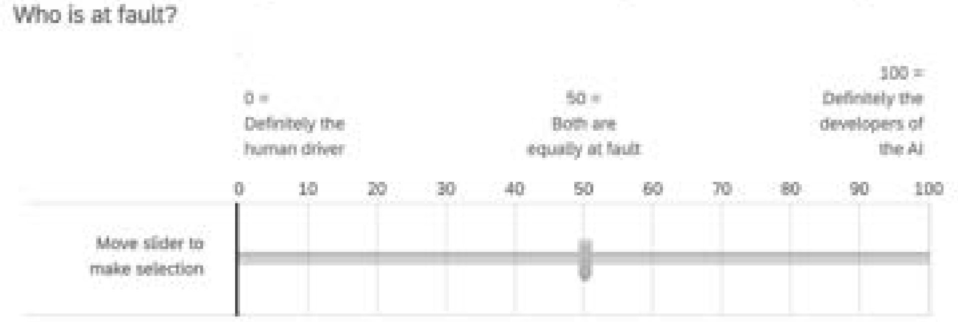


Who is at fault?

0 = Definitely the human driver
50 = Both are equally at fault
100 = Definitely the AI
0
10
20
30
40
50
60
70
80
90
100
Move slider to make selection

II. STUDIES 1A, 1B, AND 1C

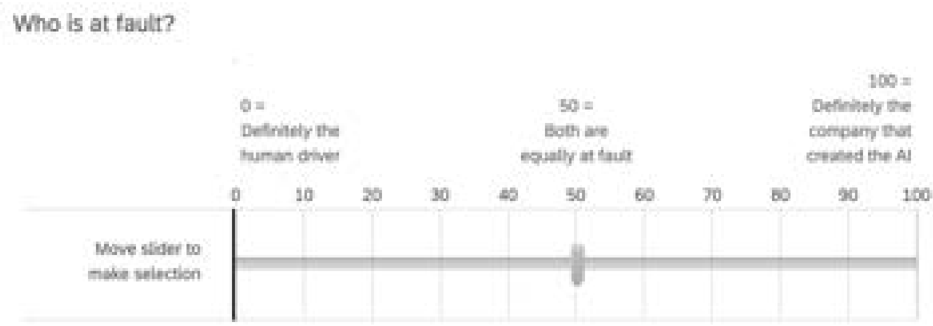
Study 1A. The materials were identical to the primary Study 1, except that the label for the 100-point on the scale was changed to the following:



Who is at fault?

0 = Definitely the human driver
50 = Both are equally at fault
100 = Definitely the developers of the AI
0
10
20
30
40
50
60
70
80
90
100
Move slider to make selection

Study 1B. The materials were identical to the primary Study 1 (and also the secondary Study 1A), except that the label for the 100-point on the scale was changed to the following:



Who is at fault?

Study 1C. The materials were identical to the primary Study 1 (and also the secondary Studies 1A and 1B), except for the following:

1. Participants saw only one of the sixteen prompts;
2. After seeing the prompt, they answered all of the questions presented below (rather than just the fault question); and
3. The first seven items below (fault, proximate cause, factual cause, ordinary cause, blame, legal liability, legal responsibility) were graded on a 0 to 100 scale, where 0 = "definitely the human driver," 50 = "both equally," and 100 = "definitely the AI (including its developers and/or parent company);" the remaining seven items were graded on a 0 to 100 scale where 0 = "definitely not," 50 = "unsure," and 100 = "definitely yes."

The dependent variables were as follows:

- Who was at fault?
- "Proximate cause" refers to an action that is legally sufficient to find the defendant liable. [As an extreme example, consider a mother who gave birth to a person who then robbed someone 45 years later. The mother's giving birth to the robber is not a proximate cause of the robbery.] In the accident you just read about, who was the proximate cause?
- "Factual cause" refers to an action that causes an event. In other words, the event would not have happened had the action not been performed. In the accident you just read about, who was the factual cause?
- Who caused the accident?
- Who deserves the blame?
- Who should be held legally liable?
- Who should be held legally responsible?
- In relation to this accident, would you say that the human driver violated norms for behavior?

- In relation to this accident, would you say that the AI (including its developers and/or parent company) violated norms for behavior?
- In relation to this accident, would you characterize the behavior of the AI (including its developers and/or parent company) as abnormal?
- In relation to this accident, would you characterize the behavior of the human driver as abnormal?
- Was the human driver's behavior morally wrong?
- Was the behavior of the AI (including its developers and/or parent company) morally wrong?
- In your opinion, did the behavior of the human driver supersede any negligence by the AI (including its developers and/or parent company)?

III. STUDY 2

[There were two conditions, one involving a m2h handoff and another involving a human-to-human handoff. Participants saw one of the two conditions. Here, I present the m2h condition first, followed by the h2h condition.]

Making bail decisions is difficult. It's hard to know whether a defendant, if released, will commit additional crimes or fail to show up for subsequent court dates. To help with this problem, one jurisdiction has been using a specially-trained Artificial Intelligence (AI) to make bail recommendations. The AI has proven extremely good in testing and in actual cases and is considered a true expert at the task. But there must be a "handoff." The human judge is ultimately responsible and is required to review each case. The AI's bail recommendation can be only one factor (and not a determining factor) in the judge's review. While the average judge in the jurisdiction handles many cases per year, in only a few of these, if any, will a judge not follow the AI's recommendation.

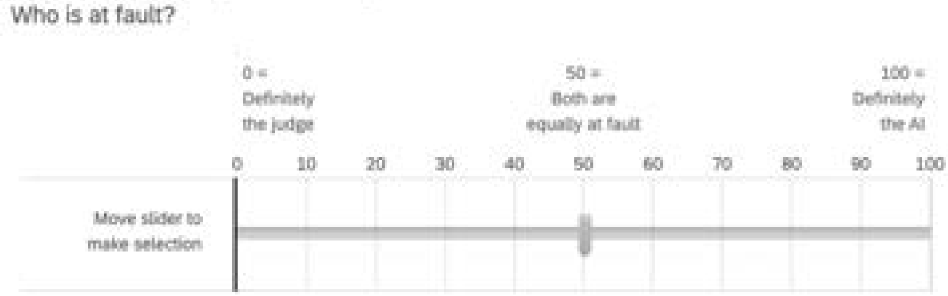
In this same jurisdiction, a twenty-six-year-old Black male was arrested and charged with assault and battery. At the bail hearing, the AI indicated that the defendant appeared to present a "high risk." Instead of recommending bail at $5,000, which arguably would have been more standard in such a case, the AI recommended bail of $40,000. The judge set bail at $40,000.

The defendant could not make bail and wound up spending a significant amount of time behind bars. However, the charges were ultimately dismissed, and the defendant was released.

Three years have passed, and the defendant has not been arrested for any subsequent offenses. Although he struggled to find a new job after being released, he now has a steady, well-paying job, and he and his family are doing well.

Researchers who were auditing court performance found this case, and they concluded that the seemingly improper bail decision possibly was the result of racial bias.

[The following are the dependent variables for the m2h condition.]



Who is at fault?

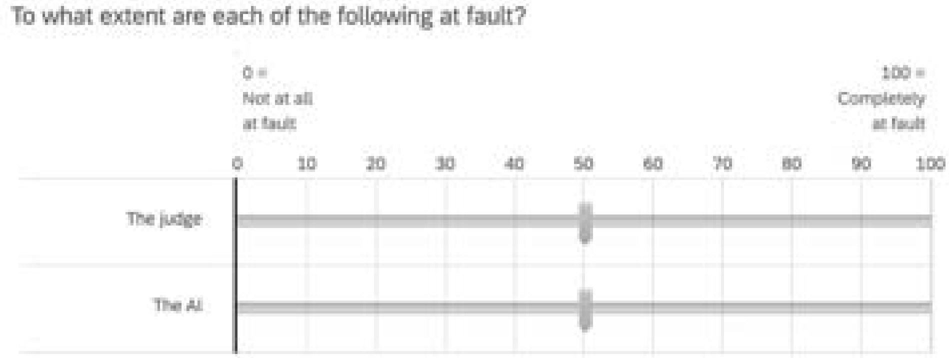
0 = Definitely the judge

50 = Both are equally at fault

100 = Definitely the AI

0 10 20 30 40 50 60 70 80 90 100

Move slider to make selection



To what extent are each of the following at fault?

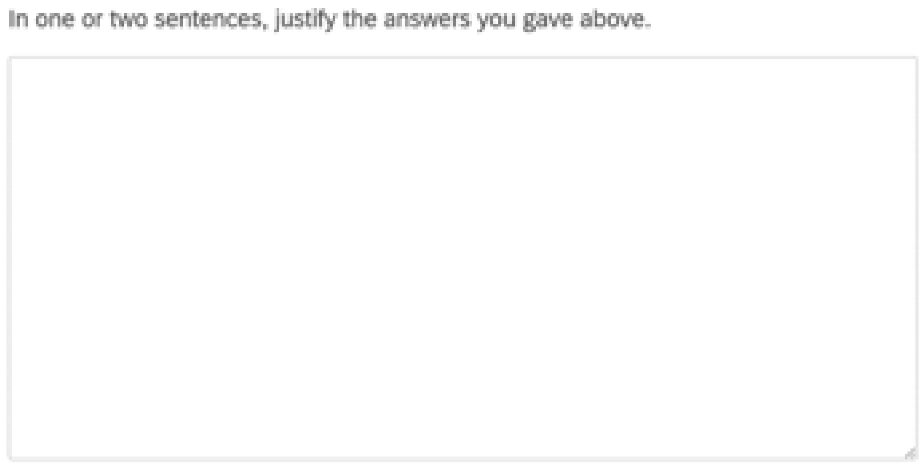
0 = Not at all at fault

100 = Completely at fault

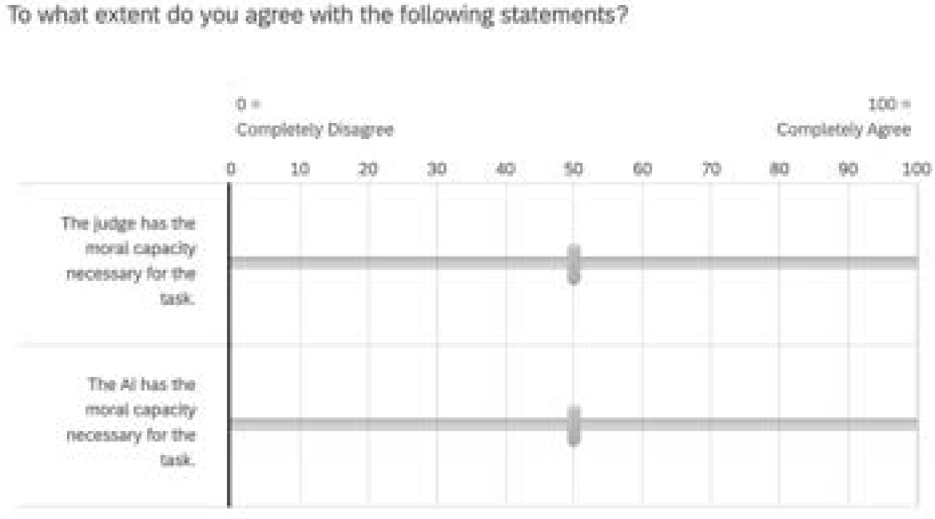
0 10 20 30 40 50 60 70 80 90 100

The judge

The AI



In one or two sentences, justify the answers you gave above.



To what extent do you agree with the following statements?

The judge has the moral capacity necessary for the task.

The AI has the moral capacity necessary for the task.

[The following is the h2h condition.]

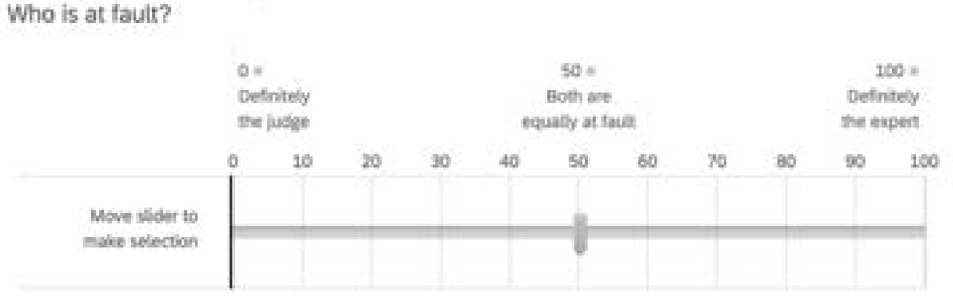
Making bail decisions is difficult. It's hard to know whether a defendant, if released, will commit additional crimes or fail to show up for subsequent court dates. To help with this problem, one jurisdiction has been using a specially-trained individual to make bail recommendations. This individual has proven extremely good in testing and in actual cases and is considered a true expert at the task. But there must be a "handoff." The judge is ultimately responsible and is required to review each case. The expert's bail recommendation can be only one factor (and not a determining factor) in the judge's review. While the average judge in the jurisdiction handles many cases per year, in only a few of these, if any, will a judge not follow the expert's recommendation.

In this same jurisdiction, a twenty-six-year-old Black male was arrested and charged with assault and battery. At the bail hearing, the expert indicated that the defendant appeared to present a "high risk." Instead of recommending bail at $5,000, which arguably would have been more standard in such a case, the expert recommended bail of $40,000. The judge set bail at $40,000.

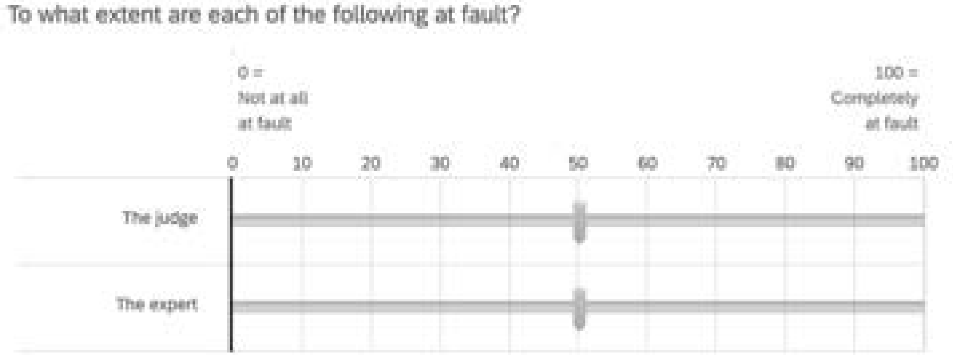
The defendant could not make bail and wound up spending a significant amount of time behind bars. However, the charges were ultimately dismissed, and the defendant was released. Three years have passed, and the defendant has not been arrested for any subsequent offenses. Although he struggled to find a new job after being released, he now has a steady, well-paying job, and he and his family are doing well.

Researchers who were auditing court performance found this case, and they concluded that the seemingly improper bail decision possibly was the result of racial bias.

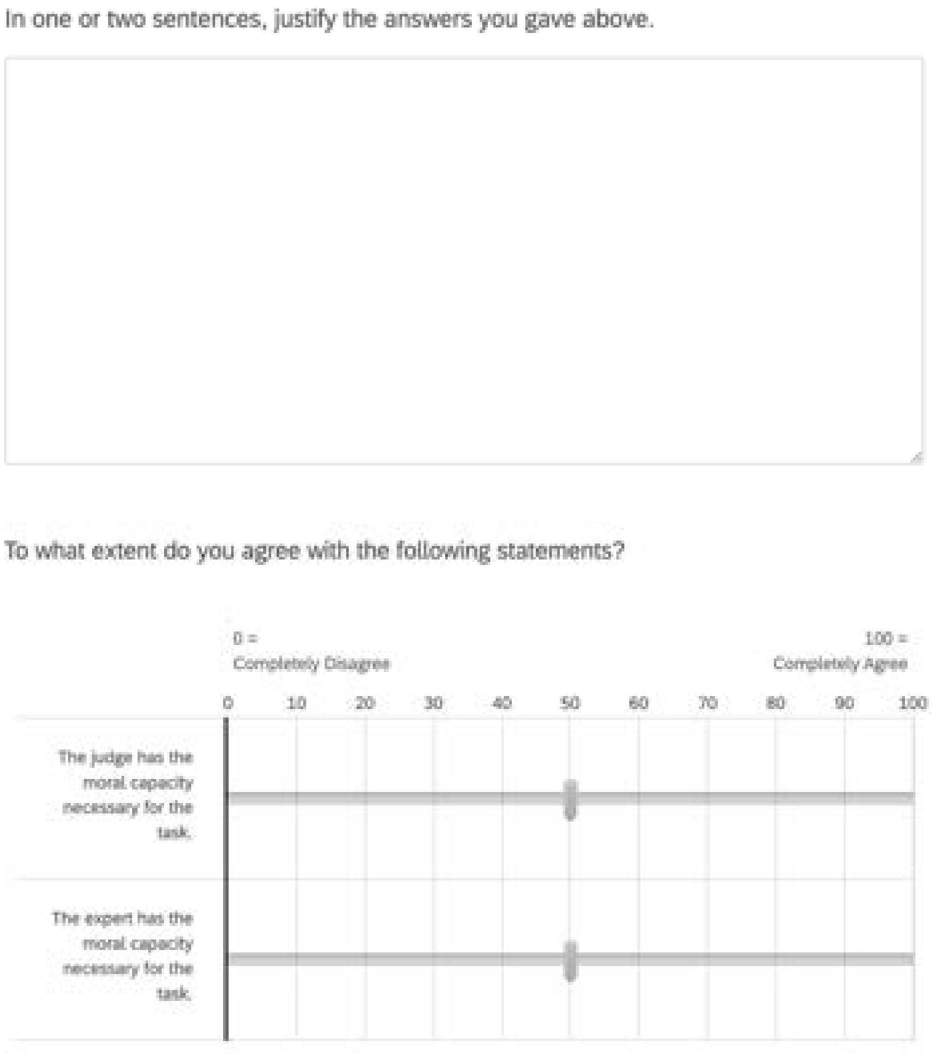
[These are the dependent variables for the h2h condition.]



Who is at fault?



To what extent are each of the following at fault?



In one or two sentences, justify the answers you gave above.

To what extent do you agree with the following statements?

The judge has the moral capacity necessary for the task.

The expert has the moral capacity necessary for the task.

IV. STUDY 2A

The materials were identical to the primary Study 2, except for the following:

1. Instead of one m2h condition and one h2h condition, there was only the m2h condition, except now there were two variants of it: one was identical to that which appeared in the primary Study 2, and the other was also identical to what appeared in the primary Study 2 but now it included an information portion as follows:

> Surprisingly, much research has shown that AI systems can often cause failures that appear to be the result of human mistakes. In other words, the design of the system sets the human up for failure.
>
> As one prominent article concluded: "When it comes to the human capacity to monitor an automated system for its failures, research findings are consistent—humans are very poor at this task." In other words, it was highly unlikely that any judge would have been able to catch a mistake in the AI's bail suggestion.
>
> Moreover, numerous studies have shown that it is difficult psychologically for humans to override machine suggestions. Even medical doctors will ignore their own best instincts and follow incorrect guidance from a machine. This is especially true when there is no meaningful way for the human actor to determine whether the machine is right or wrong. For instance, in the case you read about, the task of calculating defendant risk had been given to the AI, and there was no way, at the time the decision was made, for the judge to discover whether the AI was right or wrong.

2. The DVs were identical to those in the primary Study 2 except that the free response question was not presented.

APPENDIX B - ADDITIONAL ANALYSES FOR STUDIES 1, 1A, 1B, 1C

I. FULL STATISTICAL OUTPUT FOR STUDY 1

Prompt	**Mean**	**95% CI**	***p***	p_{bonf}	**Cohen's *d***
A	12.10	[4.42, 19.78]	< .001	< .001	1.84
B	10.77	[3.59, 17.94]	< .001	< .001	2.04
C	18.33	[9.41, 27.25]	< .001	< .001	1.33
D	8.97	[4.02, 13.91]	< .001	< .001	3.10
E	11.60	[4.96, 18.24]	< .001	< .001	2.16
F	6.30	[1.52, 11.08]	< .001	< .001	3.41
G	27.00	[16.60, 37.40]	< .001	= .002	.83
H	7.70	[2.47, 12.93]	< .001	< .001	3.02
I	16.73	[7.91, 25.55]	< .001	< .001	1.41
J	7.40	[2.85, 11.95]	< .001	< .001	3.50
K	17.27	[8.63 25.91]	< .001	< .001	1.41
L	9.17	[2.98, 15.35]	< .001	< .001	2.47
M	11.40	[4.76 18.04]	< .001	< .001	2.17
N	8.40	[3.80, 13.00]	< .001	< .001	3.38
O	17.87	[7.79, 27.94]	< .001	< .001	1.19
P	9.47	[3.13, 15.80]	< .001	< .001	2.39

II. FULL STATISTICAL OUTPUT FOR STUDY 1A

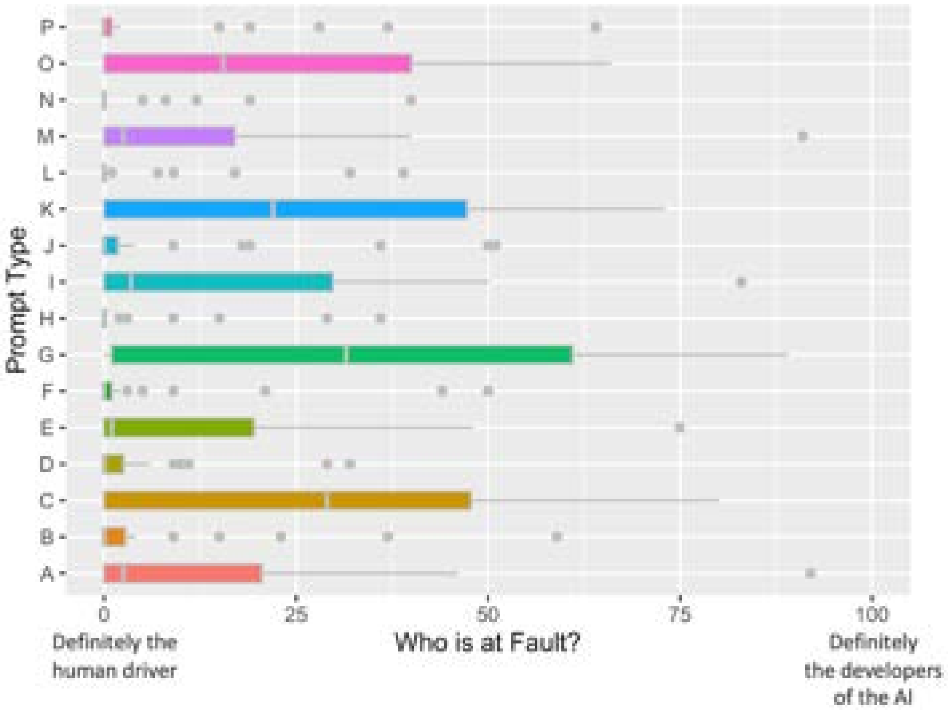


Figure B1. For Study 1A, across all the prompt variants, participants consistently concluded that the human driver was more at fault for the bad outcome following the m2h handoff.

Prompt	Mean	95% CI	*p*	p_{bonf}	Cohen's *d*
A	13.80	[6.02, 21.58]	< .001	< .001	1.74
B	5.43	[.57, 10.29]	< .001	< .001	3.42
C	26.70	[17.54, 35.86]	< .001	< .001	.95
D	3.77	[.75, 6.78]	< .001	< .001	5.73
E	11.53	[4.66, 18.41]	< .001	< .001	2.09
F	4.53	[-.06, 9.12]	< .001	< .001	3.70
G	36.37	[24.85, 47.88]	= .02	= .35	.44
H	3.13	[-.09, 6.36]	< .001	< .001	5.43
I	14.60	[7.08, 22.12]	< .001	< .001	1.76
J	6.37	[1.01, 11.72]	< .001	< .001	3.04
K	24.37	[15.40, 33.33]	< .001	< .001	1.07
L	3.80	[.26, 7.34]	< .001	< .001	4.87

M	12.73	[5.21, 20.26]	< .001	< .001	1.85
N	4.13	[.16, 8.11]	< .001	< .001	4.31
O	22.40	[13.72, 31.08]	< .001	< .001	1.19
P	5.60	[.27, 10.93]	< .001	< .001	3.11

III. FULL STATISTICAL OUTPUT FOR STUDY 1B

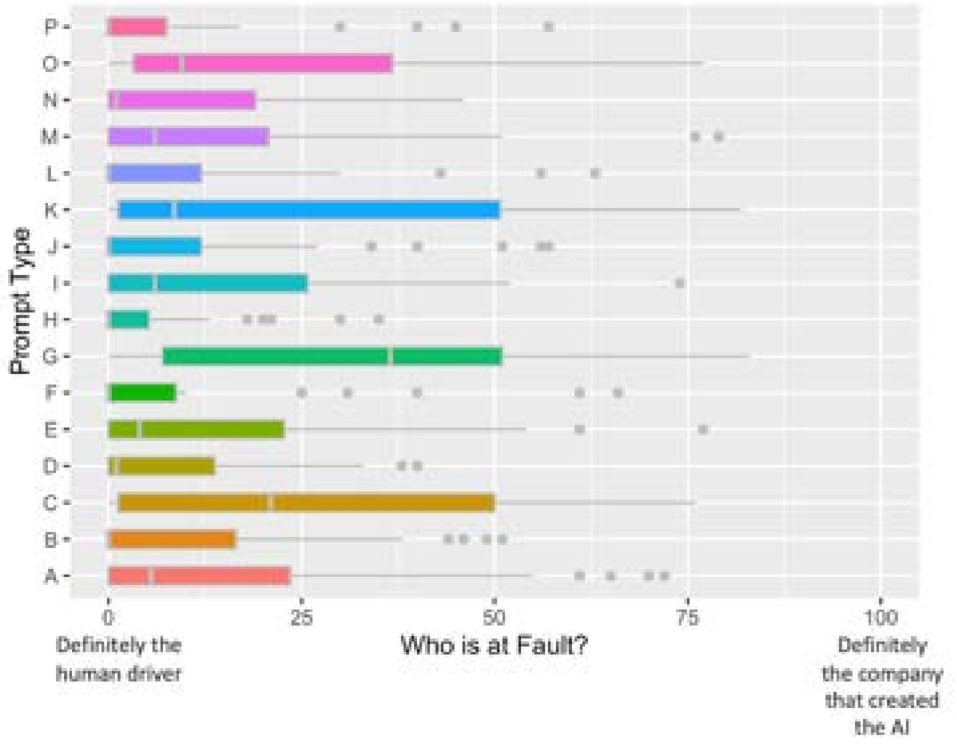


Figure B2. For Study 1B, across all the prompt variants, participants consistently concluded that the human driver was relatively more at fault for the bad outcome following the m2h handoff.

Prompt	Mean	95% CI	*p*	p_{bonf}	Cohen's *d*
A	17.20	[8.30, 26.10]	< .001	< .001	1.38
B	11.27	[4.81 17.72]	< .001	< .001	2.24
C	25.97	[16.18, 35.76]	< .001	< .001	.92
D	8.10	[3.74, 12.46]	< .001	< .001	3.59

E	15.03	[6.94, 23.13]	< .001	< .001	1.61
F	9.03	[2.38, 15.69]	< .001	< .001	2.30
G	33.17	[22.90, 43.44]	= .002	= .04	.61
H	5.27	[1.63, 8.90]	< .001	< .001	4.60
I	14.93	[7.91, 21.96]	< .001	< .001	1.86
J	10.70	[3.89, 17.51]	< .001	< .001	2.16
K	26.83	[15.56, 38.11]	< .001	< .001	.77
L	10.07	[3.57, 16.57]	< .001	< .001	2.29
M	16.33	[7.72, 24.95]	< .001	< .001	1.46
N	10.47	[4.66, 16.28]	< .001	< .001	2.54
O	23.37	[13.80, 32.93]	< .001	< .001	1.04
P	8.07	[2.42, 13.72]	< .001	< .001	2.77

IV. FULL STATISTICAL OUTPUT FOR STUDY 1C

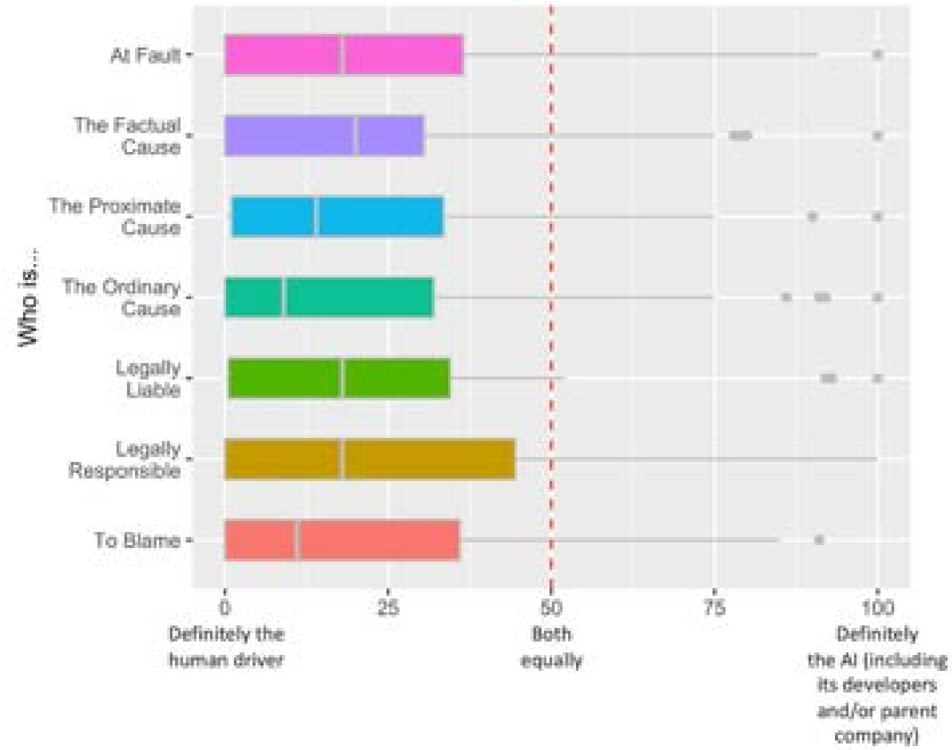


Figure B3. For Study 1C, across all the primary seven fault variants, participants consistently attributed each one more to the human driver.

DV	Mean	95% CI	*p*	p_{bonf}	Cohen's *d*
At Fault	24.46	[14.91, 34.00]	< .001	< .001	.92
Factual Cause	29.11	[17.89, 40.34]	< .001	= .004	.64
Proximate Cause	26.06	[16.71, 35.41]	< .001	< .001	.88
Ordinary Cause	24.11	[14.73, 33.50]	< .001	< .001	.95
Legally Liable	22.77	[12.26, 33.29]	< .001	< .001	.89
Legally Responsible	26.57	[15.76, 37.39]	< .001	< .001	.74
To Blame	23.34	[13.85, 32.84]	< .001	< .001	.96

For the paired items, the results were as follows:

- Behavior violated norms
    - A paired *t*-test showed a significant difference in ratings across the two entities: human (*M* = 51.63) and AI (*M* = 29.43), with *t*(34) = 3.13, *p* = .004, p_{bonf} = .01, Cohen's *d* = .02.
- Behavior was abnormal
    - A paired *t*-test showed a significant difference in ratings across the two entities: human (*M* = 46.11) and AI (*M* = 32.71), with *t*(34) = 2.14, *p* = .04, p_{bonf} = .11, Cohen's *d* = .45.
- Behavior was immoral
    - A paired *t*-test showed a significant difference in ratings across the two entities: human (*M* = 41.94) and AI (*M* = 24.57), with *t*(34) = 2.62, *p* = .01, p_{bonf} = .04, Cohen's *d* = .61.