NUS Law Working Paper 2019/016

Artificial Intelligence and The Problem of Autonomy

Simon Chesterman

chesterman@nus.edu.sg

[September 2019]

Forthcoming in the *Notre Dame Journal on Emerging Technologies*

ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF AUTONOMY

*Simon Chesterman**

Artificial intelligence (AI) systems are routinely said to operate autonomously, exposing gaps in regulatory regimes that assume the centrality of human actors. Yet surprisingly little attention is given to precisely what is meant by "autonomy" and its relationship to those gaps. Driverless vehicles and autonomous weapon systems are the most widely studied examples, but related issues arise in algorithms that allocate resources or determine eligibility for programs in the private or public sector. This article develops a novel typology of autonomy that distinguishes three discrete regulatory challenges posed by AI systems: the practical difficulties of managing risk associated with new technologies, the morality of certain functions being undertaken by machines at all, and the legitimacy gap when public authorities delegate their powers to algorithms.

* Dean and Provost's Chair Professor, National University of Singapore Faculty of Law. I am deeply grateful to Damian Chalmers, Miriam Goldby, Hu Ying, Arif Jamal, Jeong Woo Kim, Koh Kheng Lian, Lau Kwan Ho, Emma Leong, Lin Lin, Daniel Seng, Sharon Seah, David Tan, Tan Zhong Xing, Umakanth Varottil, Nico Zachert, and others for their comments on earlier versions of this text. Invaluable research assistance was provided by Violet Huang, Eugene Lau, Ong Kye Jing, and Yap Jia Qing. Errors and omissions are due to the author alone.

INTRODUCTION

On a moonless Sunday night in March 2018, Elaine Herzberg stepped off an ornamental median strip to cross Mill Avenue in Tempe, Arizona. It was just before 10 pm and the 49-year-old homeless woman was pushing a bicycle laden with shopping bags. She had nearly made it to the other side of the four-lane road when an Uber test vehicle travelling at 40 mph collided with her from the right. Ms. Herzberg, known to locals as "Ms. Elle," was taken to hospital but died of her injuries, unwittingly finding a place in history as the first pedestrian death caused by a self-driving car.[1]

The Volvo XC90 that hit her was equipped with forward and side-facing cameras, radar and lidar (light detection and ranging), as well as navigation sensors and an integrated computing and data storage unit. A report by the U.S. National Transportation Safety Board (NTSB) concluded that the vehicle detected Ms. Herzberg, but that the software classified her as an unknown object, as a vehicle, and then as a bicycle with an uncertain future travel path. At 1.3 seconds before impact, the system determined that emergency braking was needed—but this had been disabled to reduce the potential for "erratic vehicle behavior."[2]

It is still not entirely clear what went wrong on Mill Avenue that night. Uber removed its test vehicles from the four U.S. cities in which they had been operating, but eight months later they were back on the road—though now limited to 25 mph and no longer allowed to drive at night or in wet weather.[3]

A key feature of modern artificial intelligence (AI) is the ability to operate without human intervention.[4] It is commonly said that such systems operate

---

[1] Greg Bensinger and Tim Higgins, *Uber Suspends Driverless-Car Program After Pedestrian Is Killed*, WALL ST. J., Mar. 20, 2018; Troy Griggs and Daisuke Wakabayashi, *How a Self-Driving Uber Killed a Pedestrian in Arizona*, N.Y. TIMES, Mar. 20, 2018; Faiz Siddiqui and Michael Laris, *Self-Driving Uber Vehicle Strikes and Kills Pedestrian*, WASH. POST, Mar. 19, 2018.

[2] Preliminary Report Highway HWY18MH010 (National Transport Safety Board, Washington, DC, May 24, 2018).

[3] Wakabayashi Daisuke and Kate Conger, *Uber's Self-Driving Cars Are Set to Return in a Downsized Test*, N.Y. TIMES, Dec. 5, 2018.

[4] For a discussion of attempts to define AI, *see* ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 1-5 (Stuart J. Russell and Peter Norvig eds., 3rd ed. 2010). Four broad approaches can be identified: acting humanly (the famous Turing test), thinking humanly (modelling

"autonomously." As a preliminary matter, it is helpful to distinguish between *automated* and *autonomous* activities. Many vehicles have automated functions, for example, such as cruise control, which regulates speed. These functions are supervised by the driver, who remains in active control of the vehicle. Autonomous in this context means that the vehicle itself is capable of taking decisions without input from the driver—indeed, there may be no "driver" at all.

The vehicle that killed Elaine Herzberg was operating autonomously, but it was not empty. Sitting in the driver's seat was Rafaela Vasquez, hired by Uber as a safety driver. The safety driver was expected to intervene and take action if necessary, though the system was not designed to alert her. Police later determined that Ms. Vasquez had most likely been watching a streaming video—an episode of the televised singing competition "The Voice," it seems—for the twenty minutes prior to the crash. System data showed that, just before impact, she did reach for the steering wheel and applied the brakes about a second later—after hitting the pedestrian. Once the car had stopped, it was Ms. Vasquez who called 911 for assistance.[5]

Who should be held responsible for such an incident: Uber? The "driver"? The company that made the AI system controlling the vehicle? The car itself? No one?[6] The idea that no one should be held to account for the death of a pedestrian strikes most observers as wrong, yet hesitation as to the relative

---

cognitive behavior), thinking rationally (building on the logicist tradition), and acting rationally (a rational-agent approach favored by Russell and Norvig as it is not dependent on a specific understanding of human cognition or an exhaustive model of what constitutes rational thought). Though much of the literature focuses on "general" or "strong" AI (meaning the creation of a system that is capable of performing any intellectual task that a human could) the focus in this article is on the more immediate challenges raised by "narrow" AI—meaning systems that can apply cognitive functions to specific tasks typically undertaken by a human. A related term is "machine learning," a subset of AI that denotes the ability of a computer to improve on its performance without being specifically programmed to do so. The program AlphaGo Zero, for example, was merely taught the rules of the notoriously complex board game *Go*; using that basic information, it developed novel strategies that have established its superiority over any human player. *See* David Silver et al., *Mastering the Game of Go Without Human Knowledge*, 550 NATURE 354 (10/18/online 2017). This process may be supervised or unsupervised, or through a process of reinforcement. *See* KEVIN P. MURPHY, MACHINE LEARNING: A PROBABILISTIC PERSPECTIVE 2 (2012). *See* the discussion of human-in-the-loop and other models in part III.

[5] Samuel Gibbs, *Uber's Self-driving Car Saw the Pedestrian but Didn't Swerve*, THE GUARDIAN, May 8, 2018; *Why Uber's Self-driving Car Killed a Pedestrian*, ECONOMIST, May 29, 2018; Heather Somerville and David Shepardson, *Uber Car's "Safety" Driver Streamed TV Show Before Fatal Crash: Police*, REUTERS, June 22, 2018; HANNAH YEEFEN LIM, AUTONOMOUS VEHICLES AND THE LAW: TECHNOLOGY, ALGORITHMS AND ETHICS 69-80 (2018).

[6] There is also an argument that the late Ms. Herzberg might also have been at least partly at fault.

fault of the other parties suggests the need for greater clarity as to how that responsibility should be determined. As systems operating with varying degrees of autonomy become more sophisticated and more prevalent, that need will become more acute.

Though the problem of autonomy is commonly treated as a single quality of AI systems, this article develops a typology of autonomy that highlights three discrete sets of regulatory challenges, epitomized by three spheres of activity in which those systems display degrees of autonomous behavior.[7]

The first and most prominent is autonomous vehicles, the subject of Part I.[8] Certain forms of transportation have long operated without active human control in limited circumstances—autopilot on planes while cruising, for example, or driverless light rail. As the level of autonomy has increased, however, and as vehicles such as driverless cars and buses interact with other road users, it is necessary to consider how existing rules on liability for damage may need to be adapted, and whether criminal laws that presume the presence of a driver need to be reviewed. Various jurisdictions in the United

[7] The term "regulation" is chosen cautiously. Depending on context, its meaning can range from any form of behavioral control, whatever the origin, to the specific rules adopted by government that are subsidiary to legislation. BARRY M. MITNICK, THE POLITICAL ECONOMY OF REGULATION: CREATING, DESIGNING, AND REMOVING REGULATORY FORMS (1980); ANTHONY OGUS, REGULATION: LEGAL FORM AND ECONOMIC THEORY (2004); THE OXFORD HANDBOOK OF REGULATION (Robert Baldwin, Martin Cave, and Martin Lodge eds., 2010); TONY PROSSER, THE REGULATORY ENTERPRISE: GOVERNMENT, REGULATION, AND LEGITIMACY 1-6 (2010). For present purposes, the focus is on public control of a set of activities. *Cf.* Philip Selznick, *Focusing Organizational Research on Regulation*, *in* REGULATORY POLICY AND THE SOCIAL SCIENCES 363, 363 (Roger Noll ed., 1985) (defining regulation as "sustained and focused control exercised by a public agency over activities that are valued by the community"). This embraces two important aspects. One is the exercise of control, which may be through rules, standards, or other means including supervised self-regulation. The second is that such control is exercised by one or more public bodies. These may be the executive, legislature, courts, or other governmental or intergovernmental entities, but the legitimacy of this form of regulation lies in its connection—however loose—to institutions of the state. The emphasis on public control is intended to highlight avoidance of its opposite: a set of activities that would normally be regulated falling outside the effective jurisdiction of any public entity because those activities are being undertaken by AI systems. Regulation need not, however, be undertaken purely through law in the narrow sense of the command of the sovereign backed up by sanctions. JOHN AUSTIN, THE PROVINCE OF JURISPRUDENCE DETERMINED 18-37 ([1832] 1995). It may also include economic incentives such as taxes or subsidies, recognition or accreditation of professional bodies, and other market-based mechanisms. ROBERT BALDWIN, MARTIN CAVE, AND MARTIN LODGE, UNDERSTANDING REGULATION: THEORY, STRATEGY, AND PRACTICE 3 (2nd ed. [1999] 2012). A crucial question in this context is the extent to which AI systems themselves might have a role to play in regulation. LAWRENCE LESSIG, CODE: VERSION 2.0 ([1999] 2006).

[8] *See generally* JAMES M. ANDERSON et al., AUTONOMOUS VEHICLE TECHNOLOGY: A GUIDE FOR POLICYMAKERS (2014); AUTONOMOUS DRIVING: TECHNICAL, LEGAL AND SOCIAL ASPECTS (Markus Maurer et al. eds., 2016); LIM, *supra* note 5.

States and elsewhere are already experimenting with regulatory reform intended to reap the anticipated safety and efficiency benefits without exposing road users to unnecessary risk or unallocated losses.

The second example, discussed in Part II, is autonomous weapons.[9] Where driverless cars and buses raise questions of liability and punishment for harm caused, lethal autonomous weapon systems pose discrete moral questions about the delegation of *intentional* life-and-death decisions to non-human processes. Concerns about autonomy in this context focus not only on how to manage risk, but also on whether such delegation should be permissible in any circumstances.

A third set of autonomous practices is less visible but more pervasive: decision-making by algorithm.[10] Many routine decisions benefit from the processing power of computers; in cases where similar facts should lead to similar treatment, an algorithm may yield fair and consistent results. Yet when decisions affect the rights and obligations of individuals, automated decision-making processes risk treating their human subjects purely as means rather than ends. Part III argues that this calls into question the legitimacy of those decisions when made by public authorities in particular.

Each of these topics has been the subject of book length treatments.[11] The aim here is not to attempt a complete study of their technical aspects, but to test the ability of existing regulatory structures to deal with autonomy more generally. Far from a single quality, these examples reveal discrete concerns about autonomous decision-making by AI systems: the practical challenges of managing risk associated with new technologies, the morality of certain decisions being made by machines at all, and the legitimacy gap when public authorities delegate their powers to algorithms.

---

[9] *See generally* STEVEN M. SHAKER AND ALAN R. WISE, WAR WITHOUT MEN: ROBOTS ON THE FUTURE BATTLEFIELD (1988); ARMIN KRISHNAN, KILLER ROBOTS: LEGALITY AND ETHICALITY OF AUTONOMOUS WEAPONS (2009); NEW TECHNOLOGIES AND THE LAW OF ARMED CONFLICT (Hitoshi Nasu and Robert McLaughlin eds., 2014); JAI GALLIOTT, MILITARY ROBOTS: MAPPING THE MORAL LANDSCAPE (2015); AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY (Nehal Bhuta et al. eds., 2016); ALEX LEVERINGHAUS, ETHICS AND AUTONOMOUS WEAPONS (2016); STUART CASEY-MASLEN et al., DRONES AND OTHER UNMANNED WEAPONS SYSTEMS UNDER INTERNATIONAL LAW (2018); DEHUMANIZATION OF WARFARE: LEGAL IMPLICATIONS OF NEW WEAPON TECHNOLOGIES (Wolff Heintschel von Heinegg, Robert Frau, and Tassilo Singer eds., 2018).

[10] *See generally* CHRISTOPHER STEINER, AUTOMATE THIS: HOW ALGORITHMS CAME TO RULE OUR WORLD (2012); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); ARIEL EZRACHI AND MAURICE E. STUCKE, VIRTUAL COMPETITION: THE PROMISE AND PERILS OF THE ALGORITHM-DRIVEN ECONOMY (2016).

[11] *See supra* notes 8-10.

I. DRIVERLESS CARS AND THE MANAGEMENT OF RISK

Modern transportation law generally assumes the presence of a driver, pilot, or captain. In some cases, this is explicit. A "ship," for example, is defined in some jurisdictions as being a "manned" [*sic*] vessel.[12] More often, it is implicit—either because laws were written on the assumption that there would be a person in charge of any vehicle, or because in the absence of such an identifiable individual there is no one to hold to account if a civil wrong occurs or a crime is committed.[13] The 1968 Vienna Convention on Road Traffic, for example, provides that every moving vehicle on the roads "shall have a driver."[14]

Experimentation with varying degrees of automation in cars goes back decades,[15] but truly autonomous vehicles on public roads became a more realistic prospect only in the 2010s. As technology advanced, it became helpful to define more precisely what "autonomous" might mean. In 2013, the U.S. Department of Transportation released a policy on automated vehicle development that included five levels of automation.[16] The Society of Automotive Engineers (SAE) released its own report the following year with six levels, drawing also on work done by the German Federal Highway Research Institute.[17] The SAE report has been updated twice, most recently in 2018, and the six levels are now generally regarded as the industry standard.[18]

---

[12] Robert Veal and Michael Tsimplis, *The Integration of Unmanned Ships into the Lex Maritima*, 2017 LLOYD'S MAR. & COM. L.Q. 303, 308-14 (2017).

[13] This is not limited to mechanical vehicles. In some jurisdictions, for example, horses are "vehicles" for the purposes of road transportation law only when a rider is present. *See generally* BRENDA GILLIGAN, PRACTICAL HORSE LAW: A GUIDE FOR OWNERS AND RIDERS 106-12 (2002).

[14] Convention on Road Traffic, done at Vienna, Nov. 8, 1968 (in force May 21, 1977), art 8.

[15] *See, e.g., "Phantom Auto" Will Tour City*, MILWAUKEE SENTINEL, Dec. 8, 1926 (describing a "driverless" car controlled via radio from another vehicle). For a description of 1980s research funded by DARPA, *see* Dean A. Pomerleau, *ALVINN: An Autonomous Land Vehicle in a Neural Network*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 305 (David S. Touretzky ed., 1989).

[16] U.S. Department of Transportation Releases Policy on Automated Vehicle Development (Department of Transportation, Washington, DC, May 30, 2013), *at* http://www.transportation.gov/briefing-room/us-department-transportation-releases-policy-automated-vehicle-development.

[17] Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems (Society of Automotive Engineers, Warrendale, PA, 2014), *at* http://www.sae.org/standards/content/j3016_201401.

[18] *See, e.g.,* Automated Vehicles: A Joint Preliminary Consultation Paper (Law Commission, London, Consultation Paper No. 240; Scottish Law Commission, Discussion Paper No. 166, 2018), *at* http://www.lawcom.gov.uk/project/automated-vehicles, para 2.6.

At level zero (no automation), the human driver is in complete control and performs all the driving functions; at level five (full automation), the vehicle is entirely self-driven and requires no human input whatsoever. Between these extremes, increasing amounts of control are handed off to the driving system. Level one denotes driver assistance through technologies such as cruise control, which maintains speed even as the driver remains in charge of the vehicle. In practical terms, this means the driver keeps his or her hands on the wheel. At level two, partial automation may enable the vehicle to take control of accelerating, braking, and steering, but the driver must monitor the driving environment. Though sometimes described as "hands off" mode, the driver must be ready to resume control at any time.

Level three, conditional automation, marks an inflection point. Now the driving system is primarily responsible for monitoring the environment and controlling the vehicle; the human driver may direct his or her attention elsewhere, but is expected to respond to a request to intervene. High automation, level four, removes the need for the human driver to respond to a request with the ability to bring the vehicle to a stop in the event that the human does not take control. Level three is sometimes described as "eyes off" the road, while level four is colloquially known as "mind off." At level five, no human intervention would be required at all, leading to its characterization as "steering wheel optional."[19]

The importance of that inflection point between levels two and three is apparent when it comes to liability, though where level two ends and level three begins may not always be clear. In theory, the Uber test vehicle described in the opening of this article was a level two vehicle, but its "driver" appears to have acted as though it were level three. That divergence highlights one of the significant dangers of increased autonomy if it relies on the presence of a driver ready to seize control of the vehicle at any moment. Though satisfying the legal fiction that there is a "driver," the reality is that humans not actively engaged in a task such as driving—that is, when their hands are off the wheel—are unlikely to maintain for any length of time the level of attention necessary to serve the function of backup driver in an emergency.[20] For this reason, several car manufacturers have announced that they plan to skip SAE level three completely.[21]

Many observers believe that autonomous vehicles will eventually be far

---

[19] Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems (revised) (Society of Automotive Engineers, Warrendale, PA, 2018), *at* http://www.sae.org/standards/content/j3016_201806.

[20] Raja Parasuraman and Dietrich Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52(3) HUMAN FACTORS 381 (2010).

[21] Paresh Dave, *Google Ditched Autopilot Driving Feature After Test User Napped Behind Wheel*, REUTERS, Oct. 31, 2017; *Why Car-Makers Are Skipping Sae Level-3 Automation?*, M14 INTELLIGENCE, Feb. 20, 2018.

safer than human drivers and ultimately replace them.[22] Presently, more than a million people die each year in traffic accidents around the world,[23] with the vast majority of these deaths caused by driver error.[24] As autonomous vehicles become more common, continued reliance on the fiction that there is a driver may become further and further divorced from the reality of transportation. A British Law Commission discussion paper has proposed the concept of a "user-in-charge," designating a person who might be required to take over in specified circumstances.[25] That intermediary step between a true driver and a mere passenger helpfully focuses attention on the grey zone of responsibility, but it does not resolve the questions to which it gives rise if something goes wrong.

A. Civil Liability

For the purposes of civil liability—the obligation to compensate another person that is injured, for example—existing rules can largely accommodate autonomous vehicles. Presently, if someone carelessly drives over your foot, say, the driver may be required to pay for your medical expenses. If your foot is injured because the car explodes due to a defective petrol tank, then the manufacturer may be liable. Insurance helps to allocate these costs more efficiently and many jurisdictions already require minimum levels of cover or remove questions of fault from personal injuries due to traffic accidents by providing compulsory coverage.[26] These possibilities of a suit for the tort of negligence, product liability, and statutory requirements for insurance will address most of the harms associated with autonomous vehicles.

---

[22] *See, e.g.,* Neal Katyal, *Disruptive Technologies and the Law*, 102 GEO. L.J. 1685, 1688 (2014); Jeffrey K. Gurney, *Driving Into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles*, 5 WAKE FOREST J.L. & POL'Y 393, 402 (2015); Tracy Hresko Pearl, *Fast & Furious: The Misregulation of Driverless Cars*, 73 N.Y.U. ANN. SURV. AM. L. 24, 35-39 (2017). *Cf.* LIM, *supra* note 5, at 1-2 (arguing that claims of autonomous vehicle safety have been greatly exaggerated).

[23] *See, e.g.,* Road Traffic Injuries (World Health Organization, Geneva, Dec. 7, 2018), *at* http://www.who.int/mediacentre/factsheets/fs358/en (estimating annual fatalities as a result of road traffic crashes at 1.35 million).

[24] The U.S. National Motor Vehicle Crash Causation Survey (NMVCCS), conducted from 2005 to 2007, assigned the critical reason—the last event in the crash causal chain—to the driver in 94 percent of crashes. In about 2 percent, the critical reason was assigned to a vehicle component's failure or degradation, and in another 2 percent it was attributed to the environment (slick roads, weather, etc.). Of driver errors, recognition errors accounted for about 41 percent, decision errors 33 percent, and performance errors 11 percent of the crashes. *See* Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey (National Highway Traffic Safety Administration, Washington, DC, 2015), *at* http://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115.

[25] Automated Vehicles: Consultation Paper, *supra* note 18, para 1.42.

[26] *See infra* notes 39-47.

In terms of negligence, a preliminary question is whether a duty of care is owed to those who might be harmed. In general, the driver of a car owes a duty of care to other road users.[27] On SAE levels zero, one, and two, this duty of care clearly applies. In some cases, the driver's employer may also assume such a duty. After the incident described in the introduction, for example, Uber reached an undisclosed settlement with the family of Ms. Herzberg—implicitly recognizing liability.[28] At levels three and four, however, even if a duty of care were found to exist on the part of the "driver," the standard of care owed would diminish as the responsibility for controlling the vehicle is assumed by the manufacturer.[29] At level five, there may be no driver at all.

A key question is how responsibility for actions on the part of an AI system—in this case an autonomous vehicle—is to be determined. Though it is conceivable that such a system might itself be regarded as having sufficient legal personality to be capable of committing a tort,[30] the more likely scenario is that potential gaps in accountability under civil law will be filled by product liability and by statute. Volvo's CEO made headlines in 2015 when he announced that the Swedish company would accept full liability for accidents when its cars are in autonomous mode.[31] This was somewhat disingenuous given that various jurisdictions already imposed high standards of care on manufacturers through product liability.

In Britain, drivers are generally responsible for the roadworthiness of their vehicles unless the problem is latent and not discoverable through the exercise of reasonable care.[32] As vehicles become more complex, it is less

---

[27] ROBERT M. MERKIN AND JEREMY STUART-SMITH, THE LAW OF MOTOR INSURANCE 186-88 (2004).

[28] Bernie Woodall, *Uber Avoids Legal Battle with Family of Autonomous Vehicle Victim*, REUTERS, Mar. 29, 2018. Her family subsequently sued the city of Tempe and the state of Arizona for negligence: Tim Gallen, *Arizona, Tempe Sued by Family of Woman Killed by Self-Driving Uber Vehicle* PHOENIX BUS. J., Mar. 20, 2019.

[29] Jonathan Morgan, *Torts and Technology*, *in* THE OXFORD HANDBOOK OF LAW, REGULATION, AND TECHNOLOGY 522, 538 (Roger Brownsword, Eloise Scotford, and Karen Yeung eds., 2017). Reference to the "manufacturer" here may be complicated by diverse parties involved in production and maintenance of such vehicles, but these are not new problems in product liability. *See infra* note 48.

[30] On legal personality of AI systems, *see generally* SAMIR CHOPRA AND LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS (2011); JOHN FRANK WEAVER, ROBOTS ARE PEOPLE TOO: HOW SIRI, GOOGLE CAR, AND ARTIFICIAL INTELLIGENCE WILL FORCE US TO CHANGE OUR LAWS (2014); GABRIEL HALLEVY, LIABILITY FOR CRIMES INVOLVING ARTIFICIAL INTELLIGENCE SYSTEMS (2015); LEGAL PERSONHOOD: ANIMALS, ARTIFICIAL INTELLIGENCE AND THE UNBORN (Visa A.J. Kurki and Tomasz Pietrzykowski eds., 2017).

[31] Jim Gorzelany, *Volvo Will Accept Liability for Its Self-Driving Cars*, FORBES, Oct. 9, 2015.

[32] Road Traffic Act 1988 (U.K.), s. 50A. A leading case from 1970 held that the failure of brakes on a truck did not provide a defense unless the defect was not reasonably

realistic to expect drivers to guard against latent defects. In its 2018 consultation paper, the Law Commission noted that drivers' insurers currently pay claims where it would be difficult to distinguish between driver fault and vehicle defects. In the case of autonomous vehicles, that distinction may become clearer if there is no prospect of a driver being aware of a defect in the system—or if there is no driver at all.[33]

In the United States, manufacturers and retailers can be held liable if a product is defective, either due to manufacturing or design, or if there was a failure to warn users of a non-obvious danger. In the case of a design defect, it is necessary to show that the foreseeable risks of harm could have been reduced or avoided by the adoption of a "reasonable alternative design."[34] That standard has proved problematic in cutting-edge technology cases, but the difficulties posed are surmountable.[35] To avoid uncertainty and guard against uncompensated loss, some authors have argued in favor of imposing strict liability for autonomous vehicles.[36] An alternative would be to expand the no-fault regimes for accidents already in place in New Zealand, Israel, Sweden, and a dozen U.S. states as well as parts of Australia and Canada.[37]

If the claimed benefits of autonomous vehicles in terms of safety prove true,[38] the costs associated with injuries due to traffic accidents should decline in the long term. In the medium term, however, as drivers cede control of vehicles to AI systems, the proportion of claims against drivers will drop as compared to claims against manufacturers. Those costs will therefore move from insurance premiums paid by drivers to product liability insurance on the part of manufacturers—and then back to drivers through increased prices for vehicles.[39]

---

discoverable: *Henderson v Harry Jenkins & Sons*, [1970] AC 282 (1970); MERKIN AND STUART-SMITH, *supra* note 27, at 201.

[33] Automated Vehicles: Consultation Paper, *supra* note 18, para 6.12.

[34] Restatement (Third) of Torts: Product Liability (American Law Institute, Philadelphia, 1998), §2.

[35] Mark A. Geistfeld, *The Regulatory Sweet Spot for Autonomous Vehicles*, 53 WAKE FOREST L. REV. 101, 124-25 (2018).

[36] *See, e.g.,* Adam Rosenberg, *Strict Liability: Imagining a Legal Framework for Autonomous Vehicles*, 20 TUL. J. TECH. & INTELL. PROP. 205 (2017); LIM, *supra* note 5, at 105.

[37] *See, e.g.,* Maurice Schellekens, *No-Fault Compensation Schemes for Self-Driving Vehicles*, 10(2) LAW, INNOVATION & TECH. 314 (2018/07/03 2018), *at* http://https://doi.org/10.1080/17579961.2018.1527477. For a critical account of no-fault insurance in the United States, *see* JAMES M. ANDERSON, PAUL HEATON, AND STEPHEN J. CARROLL, THE U.S. EXPERIENCE WITH NO-FAULT AUTOMOBILE INSURANCE: A RETROSPECTIVE (2010); Nora Freeman Engstrom, *When Cars Crash: The Automobile's Tort Law Legacy*, 53(2) WAKE FOREST L. REV. 293, 309-14 (2018) (discussing early enthusiasm for no-fault personal injury protection and its decline).

[38] *See supra* note 22.

[39] Daniel A. Crane, Kyle D. Logue, and Bryce C. Pilz, *A Survey of Legal Issues Arising*

Such a role for insurance has been a feature of the automobile industry from its earliest days. In Britain, for example, compulsory third-party insurance has been a requirement for anyone using a motor vehicle since 1930.[40] Basic car insurance is now mandatory in most major jurisdictions, including almost all U.S. states.[41] In Germany and Japan, strict liability on the part of owners of traditional vehicles is complemented by a mandatory insurance regime, though there is ongoing debate as to whether liability should shift towards manufacturers.[42] China introduced a requirement for mandatory insurance in 2003, but only has limited third-party coverage.[43]

For the vast majority of cases, responsibility for insurance falls to the driver or the driver's employer.[44] Britain's Automated and Electric Vehicles Act 2018 extended that insurance requirement to cover vehicles operating autonomously. Victims (including the "driver") of an accident caused by a fault in the vehicle itself will be covered.[45] Failure to have such insurance is a criminal offence, with liability in such circumstances being borne by the owner of the vehicle.[46] Moving forward, it is likely that the increased use of autonomous vehicles will see greater standardization of laws requiring that vehicles be insured rather than drivers, where that is not already the case.[47]

Other complications in attributing responsibility for civil law purposes include the many discrete components in an autonomous vehicle that might be defective, notably the various sensors—though these are practical rather than conceptual difficulties.[48] Similarly, the possibility of a hacker

---

from the Deployment of Autonomous and Connected Vehicles, 23 MICH. TELECOMM. & TECH. L. REV. 191, 256-59 (2017).

[40] Road Traffic Act 1930 (U.K.), s. 35.

[41] A few U.S. states require that a bond be posted as an alternative, while New Hampshire and Virginia are outliers in not requiring insurance at all. (Virginia residents must pay a fee of $500 if they do not have insurance, though this does not provide any coverage.) *See* Engstrom, *supra* note 37, at 306.

[42] Frank Henkel et al., Autonomous Vehicles: The Legal Landscape of DSRC in Germany (Norton Rose Fulbright, Munich, July 2017), *at* http://www.nortonrosefulbright.com/en/knowledge/publications/e77157b8/autonomous-vehicles-the-legal-landscape-of-dsrc-in-germany; Gen Goto, *Autonomous Vehicles, Ride-share Services, and Civil Liability: A Japanese Perspective*, FORTHCOMING (2019).

[43] 中华人民共和国道路交通安全法 [Law of the People's Republic of China on Road Traffic Safety] 2003 (China); XU Xian and FAN Chiang-Ku, *Autonomous Vehicles, Risk Perceptions and Insurance Demand: An Individual Survey in China*, TRANSP. RES. PART A: POL'Y & PRAC. https://doi.org/10.1016/j.tra.2018.04.009 (2018).

[44] Automated Vehicles: Consultation Paper, *supra* note 18, para 6.7.

[45] Automated and Electric Vehicles Act 2018 (U.K.), s. 2(1).

[46] *Id*., s. 2(2).

[47] *Cf.* Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 CAL. L. REV. 1611 (2017).

[48] Crane, Logue, and Pilz, *supra* note 39, at 262 ("This web of technologies at work in an ACV means there is a web of potential defendants in a lawsuit regarding an ACV's alleged

interfering with software and thereby causing a crash is a novel challenge for liability, but not materially different from a case in which an unknown person cuts the brake cables on a traditional automobile.[49] Given the foreseeability of cybersecurity issues in autonomous vehicles, it is likely for tortious purposes that reasonable safeguards against such interference would fall within the duty of care owed by the driver (to update software, for example) and the manufacturer (to provide reasonable protection against viruses and hackers).[50] Alternatively, the imposition of strict liability standards would make clear the manufacturer's responsibility to take adequate precautions. A more challenging example would be where the owner or driver of a vehicle him- or herself makes changes to an autonomous vehicle—for example, overriding security protocols or enabling it to exceed speed limits—that contribute to a crash.[51] If such situations are not covered by statute,[52] the law of contributory negligence could apportion blame as it does in other cases.[53] Further adaptations may be required if the business model of transportation changes—for example, if vehicles come to be seen as a service to be used rather than a thing to be owned.[54]

Autonomous vehicles thus pose important challenges to ensure that their presumed benefits in terms of road safety and efficiency do not come at the cost of unfair or disproportionate allocation of risk. In terms of how the civil law allocates those risks, amendments to reflect a shift of responsibility from drivers to manufacturers and software providers may be necessary, but the fundamental legal concepts are sound.

---

defect."); LIM, *supra* note 5, at 23-25.

[49] Crane, Logue, and Pilz, *supra* note 39, at 248-49.

[50] *See* Araz Taeihagh and Hazel Si Min Lim, *Governing Autonomous Vehicles: Emerging Responses for Safety, Liability, Privacy, Cybersecurity, and Industry Risks*, 39(1) TRANSP. REV. 103 (2019).

[51] An extreme case would be the potential use of an autonomous vehicle as a weapon: Tim Bradshaw, *Self-Driving Cars Raise Fears over "Weaponisation,"* FINANCIAL TIMES, Jan. 14, 2018 (discussing concerns raised by the head of self-driving cars at Baidu, speaking at the 2018 Consumer Electronics Show).

[52] *See, e.g.,* Road Traffic (Amendment) Act 2017 (Singapore) (creating a new offence of interfering without reasonable excuse with "any equipment or device in or on an autonomous motor vehicle").

[53] *See generally* Vadim Mantrov, *A Victim of a Road Traffic Accident Not Fastened by a Seat Belt and Contributory Negligence in the EU Motor Insurance Law*, 5(1) EUR. J. RISK REG. 115 (2014); Noah M. Kazis, *Tort Concepts in Traffic Crimes*, 125(4) YALE L.J. 1131, 1139-41 (2016); James Goudkamp and Donal Nolan, *Contributory Negligence in the Twenty-First Century: An Empirical Study of First Instance Decisions*, 79(4) MOD. L. REV. 575 (2016).

[54] James Arbib and Tony Seba, Rethinking Transportation 2020-2030: The Disruption of Transportation and the Collapse of the Internal-Combustion Vehicle and Oil Industries (RethinkX, San Francisco, 2017), *at* http://www.rethinkx.com/transportation.

B. Criminal Law

Not so in relation to criminal law. Criminal law is concerned less with allocating costs than apportioning blame for the purposes of deterrence and punishment. The regulation of road traffic relies heavily on criminal offences, with the majority of those offences directed at the human driver of a motor vehicle. These include responsibility not merely for the speed and direction of the vehicle, but its roadworthiness and his or her own fitness to drive. Drivers may also be required to have adequate insurance, to report accidents, and in some cases to control the behavior of passengers (such as requiring children to wear seatbelts).[55] Identification of the driver in question may be aided by a presumption that it is the person in whose name a vehicle is registered. If such a vehicle is caught by a speed camera, for example, that person may be presumptively responsible unless it possible to point to the responsibility of another person.[56]

Because of the centrality of drivers, the various jurisdictions that have allowed autonomous vehicles on open roads initially provided that a human "driver" had to be behind the wheel and alert. The first truly driverless cars on open roads were authorized in Arizona by executive order of the Governor in April 2018. The order provided, among other things, that any traffic citation or other penalty arising from infractions by the vehicle would be issued to the person "testing or operating the fully autonomous vehicle."[57] In practice, however, backup drivers remained in the various cars.[58] In the same month, California's Department of Motor Vehicles modified state regulations to allow applications for driverless testing permits.[59] Where a human backup driver is not present, a remote operator holding the appropriate license is required to "continuously supervise the vehicle's performance of the dynamic driving task."[60] Other jurisdictions have similarly expanded the concept of a "driver" to include a remote operator deemed to be in charge of the vehicle, despite not being seated within it.[61]

Singapore, like Arizona, has made provision for truly autonomous

---

[55] Automated Vehicles: Consultation Paper, *supra* note 18, para 7.1.

[56] *See, e.g.,* Road Traffic Act 1961 (Cap 276, 2004 Rev. Ed., Singapore) s. 81(1B).

[57] Executive Order 2018-04: Advancing Autonomous Vehicle Testing and Operation; Priotizing Public Safety 2018 (Arizona), para 3(c).

[58] Alex Davies, *Waymo's So-Called Robo-Taxi Launch Reveals a Brutal Truth*, WIRED, May 12, 2018.

[59] Autonomous Vehicles in California (Californai Department of Motor Vehicles, Sacramento, 2018), *at* http://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/bkgd.

[60] Testing of Autonomous Vehicles 2018 (California).

[61] *See, e.g.,* Experimenteerwet zelfrijdende auto's 2018 (Netherlands) (Dutch law allowing the use of driverless vehicles on public roads, though requiring them to be controlled remotely by a human operator).

vehicles "without the active physical control of, or monitoring by, a human operator," but this provision adopted in 2017 is limited to enabling the Minister to make rules for trials of autonomous vehicles.[62] In Germany, 2017 amendments to the Road Traffic Act allow the use of autonomous technology comparable to level three, but require that the driver remain "*wahrnehmungsbereit*" [mentally alert] at all times and able to take control of the vehicle when prompted to do so or when "*offensichtlicher Umstände*" [obvious circumstances] require it.[63]

China adopted regulations for autonomous vehicle testing in April 2018, with detailed requirements for backup drivers who would remain personally liable for any traffic violations as well as a requirement that the entity conducting the test be registered in China and have adequate civil compensation capacity for personal and property losses.[64] It remains an important jurisdiction, having overtaken the United States as the largest market for automobiles in 2009.[65] A challenge is the non-standardized road signage in China, which adds to the training time for autonomous systems. Such constraints may be offset by the more tolerant regulatory regime, far lower levels of litigation, and a willingness to embrace new technologies quickly and with higher acceptance of risk. The Chinese government has created test zones for autonomous vehicles in 14 cities, the largest being in Beijing and Shanghai.[66] This has been accompanied by large investments on the part of companies like Alibaba, Baidu, and Tencent.[67]

In terms of the SAE levels mentioned earlier, the criminal law in these jurisdictions typically continues to assume that no vehicle is operating above level two, with a human driver bearing ongoing responsibility for its operation.[68] As autonomous vehicles become more sophisticated, this position will become untenable.

A preliminary matter is that some laws as they stand may render certain forms of autonomous driving inherently unlawful. Specific requirements that cars have a "driver," for example, or that prohibit leaving a vehicle

---

[62] Road Traffic (Amendment) Act, *supra* note 52.

[63] Strassenverkehrsgesetz (StVG) [Road Traffic Act] 1909 (Germany), §1b.

[64] 智能网联汽车道路测试管理规范（试行） [Intelligent Network Linked Vehicle Road Test Management Regulations (Trial)] 2018 (People's Republic of China).

[65] Luca Pizzuto et al., How China Will Help Fuel the Revolution in Autonomous Vehicles (McKinsey Center for Future Mobility, Beijing, 2019), *at* http://www.mckinsey.com/industries/automotive-and-assembly/our-insights/how-china-will-help-fuel-the-revolution-in-autonomous-vehicles.

[66] FAN Feifei, *Autonomous Vehicles Gaining More Ground*, CHINA DAILY, Jan. 15, 2019.

[67] Chris Udemans, *Mapping Out the Road Ahead for China's Autonomous Vehicles*, TECHNODE, Feb. 7, 2019, *at* http://technode.com/2019/02/07/china-av-roadmap/.

[68] The Arizona executive order is unusual in explicitly mentioning levels four and five: Executive Order 2018-04, *supra* note 57, para 1(d).

unattended, would be incompatible with fully autonomous taxi services.[69] Unless the focus on human drivers changes, it could also result in blameless passengers being held responsible if a vehicle makes a mistake. In addition to being unfair, this could discourage public acceptance of driverless technology.[70]

Various U.S. states have experimented with different answers to the question of "driverless" cars.[71] One possibility is to continue to focus on a natural person riding in the vehicle, or controlling it remotely, as in Arizona and California.[72] This remains the most common legal position across the various jurisdictions that explicitly allow autonomous vehicles on public roads. A second approach is to focus on the "operator" of the vehicle, akin to the "user-in-charge" proposed by the British Law Commission.[73] In Georgia this means the person who "causes" the vehicle to move.[74] Thirdly, the burden can rest on the owner of the vehicle. That is the approach adopted in Texas.[75]

A fourth possibility, thus far adopted only in Tennessee, is to define the "automated driving system" (ADS) itself as the "driver." The definition of "person," in the same 2017 amendment to the State Code, was expanded to mean "a natural person, firm, co-partnership, association, corporation, *or an engaged ADS*."[76] The definition only applies to provisions of the Code concerning motor vehicles, however, and it does not appear to have been invoked for the purposes of civil liability or criminal sanction.[77] Law reform bodies in other jurisdictions, notably Australia and Britain, have suggested the concept of an automated driving system entity (ADSE), but this refers to

---

[69] *See, e.g.,* Road Vehicles (Construction and Use) Regulations 1986 (U.K.), reg 107.

[70] Michael Cameron, Realising the Potential of Driverless Vehicles: Recommendations for Law Reform (New Zealand Law Foundation, Wellington, NZ, 2018), *at* http://www.lawfoundation.org.nz/wp-content/uploads/2018/04/Cameron_DriverlessVehicles_complete-publication.pdf, at 9.

[71] The U.S. National Conference of State Legislatures maintains a list of enacted legislation on self-driving cars *at* www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx.

[72] *See supra* notes 57-60.

[73] *See supra* note 25.

[74] Official Code of Georgia Annotated §40-1-1(38) (2017): "'Operator' means any person who drives or is in actual physical control of a motor vehicle or who causes a fully autonomous vehicle to move or travel with the automated driving system engaged."

[75] Texas Transportation Code §545.453(a)(1) (2017): "the owner of the automated driving system is considered the operator of the automated motor vehicle solely for the purpose of assessing compliance with applicable traffic or motor vehicle laws, regardless of whether the person is physically present in the vehicle."

[76] Tennessee Code Annotated §55-8-101 (2017) (emphasis added): "'Driver' means: … (B) For purposes of an ADS-operated vehicle and when the context requires, the ADS when the ADS is engaged."

[77] *Cf.* the discussion of "personal" liability of AI systems discussed *supra* note 30.

the legal entity responsible for the vehicle rather than a novel category of legal person.[78]

A more utopian vision is that driverless cars may be so superior to human drivers that there is no need to provide for criminal responsibility at all.[79] That seems unrealistic, but it does raise the question of the function that road traffic laws are intended to play, and the purpose of punishing proscribed conduct. The two basic aims of road traffic law are promoting safety and order on the roads.[80] As Australia's National Transport Commission (NTC) has observed, existing penalties focus on influencing the behavior of human drivers. An individual who breaches the rules may be punished, or his or her license may be suspended or revoked. In the case of autonomous vehicles, monetary and custodial punishments may be less appropriate than seeing enforcement as part of a feedback loop to train the system. This could take the form of improvement notices and enforceable undertakings to increase safety.[81] In more serious cases, withdrawing the authorization to drive on the roads may be sufficient to protect other road users, while traditional penalties could be applied to natural or legal persons if there is evidence of wrongdoing that rises to the level of a crime.

Larger questions of whether and how AI systems themselves might be "punished" are beyond the scope of the present article.[82] For present purposes, what is interesting is that, in its application to autonomous vehicles, the criminal law sheds its deontological overtones in favor of instrumentalism—rather than moral failings to be corrected, violations may come to be seen as errors to be debugged.

---

[78] Changing Driving Laws to Support Automated Vehicles (Policy Paper) (National Transport Commission, Melbourne, May 2018), *at* http://www.ntc.gov.au/Media/Reports/(B77C6E3A-D085-F8B1-520D-E4F3DCDFFF6F).pdf, para 1.5; Automated Vehicles: Consultation Paper, *supra* note 18, para 4.107. Similarly, the Dutch Infrastructure Minister suggested that the Netherlands might introduce driving licences for cars rather than drivers. *Speech by Cora van Nieuwenhuizen, Minister of Infrastructure and Water Management* (The Hague, Global Entrepreneurship Summit, June 4, 2019).

[79] Jay L. Zagorsky, *Cops May Feel Biggest Impact from Driverless Car Revolution*, THE CONVERSATION, Mar. 16, 2015 (arguing that autonomous vehicles may reduce the need for police by half); Robin Washington, *Autonomous Vehicles Will Mean the End of Traffic Stops*, WIRED, Sept. 30, 2016 (suggesting that driverless cars will transform law enforcement on the roads as those within the vehicle would not be responsible for its actions).

[80] SALLY CUNNINGHAM, DRIVING OFFENCES: LAW, POLICY AND PRACTICE 1-6 (2008).

[81] Changing Driving Laws, *supra* note 78, para 8.2.1. *See* also Automated Vehicles: Consultation Paper, *supra* note 18, paras 7.33-7.34. It would depend, of course, on why a given violation took place. If the violation was due to an override by the driver/user-in-charge, for example, traditional penalties might apply.

[82] *See, e.g.,* GABRIEL HALLEVY, WHEN ROBOTS KILL: ARTIFICIAL INTELLIGENCE UNDER CRIMINAL LAW (2013); Ying Hu, *Robot Criminals*, 52 U. MICH. J.L. REFORM 487 (2019).

C. Ethics

The possibility that autonomous vehicles will—eventually—be significantly better drivers than humans has invited much speculation about how they can and should behave in limit situations, such as an impending crash. Human drivers are generally held to the standard of the "reasonable driver."[83] If a child runs onto a street, for example, swerving to avoid him or her might be a violation of the road rules—but unlikely to be one that is prosecuted. Swerving to avoid a rat, by contrast, may not be excused.[84] An autonomous vehicle might respond more swiftly, but lacks the moral compass expected to guide a human. That must be programmed in or learned through experience.[85]

A common illustration of the dilemmas that can arise is the trolley problem used by ethicists. A single-carriage train is heading towards five people and will kill them all. If a lever is pulled, the train will be diverted onto a siding but will kill someone else. Do you pull the lever? Though many people would do so, there is no "right" answer to this question. When confronted with an analogous situation in which five people are going to die and the only way to stop the train is by pushing someone into its path, most people tend to hold back. The first scenario reflects a utilitarian approach that looks to the consequences of an action (one death versus five). The second *feels* different because we know intuitively that pushing a person to their death is wrong—even though the choice is still between one person and five people dying.[86]

Researchers at MIT developed a Moral Machine that offers these and a dozen other scenarios that might confront driverless cars.[87] Should two passengers be sacrificed if it would save five pedestrians? Does it matter if the pedestrians were jaywalking? If they were criminals? In real life, faster reaction times mean that braking would almost certainly be the best choice,

---

[83] Jeffrey K. Gurney, *Imputing Driverhood: Applying a Reasonable Driver Standard to Accidents Caused by Autonomous Vehicles*, *in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 51 (Patrick Lin, Keith Abney, and Ryan Jenkins eds., 2017)

[84] *See* Filippo Santoni de Sio, *Killing by Autonomous Vehicles and the Legal Doctrine of Necessity*, 20(2) ETHICAL THEORY AND MORAL PRACTICE 411 (2017) (discussing the difficulty of transposing ethical principles and the legal doctrine of necessity to autonomous vehicles).

[85] Ivó Coca-Vila, *Self-Driving Cars in Dilemmatic Situations: An Approach Based on the Theory of Justification in Criminal Law*, 12 CRIM. L. & PHIL. 59 (2018).

[86] *See generally* DAVID EDMONDS, WOULD YOU KILL THE FAT MAN? THE TROLLEY PROBLEM AND WHAT YOUR ANSWER TELLS US ABOUT RIGHT AND WRONG (2013); THOMAS CATHCART, THE TROLLEY PROBLEM; OR, WOULD YOU THROW THE FAT GUY OFF THE BRIDGE? A PHILOSOPHICAL CONUNDRUM (2013).

[87] *See* http://moralmachine.mit.edu.

but for the purposes of the experiment one is to assume that the brakes have failed and that the vehicle cannot stop. In an unusual sampling method, they abandoned standard academic survey approaches to deploy a "viral online platform"—raising problems of self-selection, but enabling them to gather data from millions of people all over the world.[88]

Among the interesting findings were clear global preferences for sparing human lives over animals, sparing more lives, and sparing young lives.[89] The first of these is consistent with rules proposed by the German Ethics Commission on Automated and Connected Driving; the last, however, runs directly counter to a proposed prohibition on making distinctions based on personal features such as age.[90] In a subsequent interview about the paper, one of its authors was asked about the implicit prejudices disclosed in the results—sparing professionals over the homeless, the healthy over the obese, dogs over criminals, and so on. "That suggests to us that we shouldn't leave decisions completely in the hands of the demos," he replied.[91]

In practice, it should be noted, such "dilemma situations" are overly reductive. They posit the false dichotomy of exactly one out of two results, when the reality of any actual road incident is that there are a great many possible outcomes.[92] That is especially true of those scenarios in which a vehicle must either kill its occupants or kill pedestrians. In any event, executives at Mercedes-Benz are on record saying that they would prioritize the lives of passengers in its cars.[93] A paper published in *Science* supports this commercial decision: while many people approve of autonomous vehicles sacrificing a passenger to save other people in theory, they are unlikely to buy or ride in a car programmed that way in practice.[94]

Regulators, for their part, have emphasized the importance of safety in a general sense, but without weighing in on specific choices to be made by autonomous vehicles in such limit situations. While human drivers

---

[88] Edmond Awad et al., *The Moral Machine Experiment*, 563 NATURE 59, 63 (2018).

[89] *Id*. at 60. The authors also identified, however, a split between cultures broadly identified as individualistic and collectivistic, with the latter demonstrating a significantly weaker preference for sparing younger characters in the various scenarios. *Id*. at 62.

[90] Christoph Luetge, *The German Ethics Code for Automated and Connected Driving*, 30(4) PHIL. & TECH. 547 (2017).

[91] Caroline Lester, *A Study on Driverless-Car Ethics Offers a Troubling Look Into Our Values*, NEW YORKER, Jan. 24, 2019 (quoting Azim Shariff).

[92] Tom Michael Gasser, *Fundamental and Special Legal Questions for Autonomous Vehicles*, *in* AUTONOMOUS DRIVING: TECHNICAL, LEGAL AND SOCIAL ASPECTS 523, 533-34 (Markus Maurer, et al. eds., 2016).

[93] MIchael Taylor, *Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians*, CAR AND DRIVER, Oct. 8, 2016, *at* http://www.caranddriver.com/news/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians.

[94] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, *The Social Dilemma of Autonomous Vehicles*, 352(6293) SCIENCE 1573 (2016).

predominate on the roads, the standard of the reasonable driver is likely to persist and autonomous vehicles will be measured against that. If and when those proportions are switched, new standards may be required, with a corresponding move from licensing the skills of a driver to certifying the safety of a product.

II. KILLER ROBOTS AND THE MORALITY OF OUTSOURCING

Autonomous vehicles raise concerns about how they fit into existing models of civil and criminal liability, as well as how AI systems should take decisions in life-and-death situations such as an imminent crash. These are, in many ways, problems to manage through technical improvement and regulatory tweaks. The prospect of truly autonomous weapon systems, by contrast, has led to calls for a moratorium or an outright ban.[95]

In one sense, this may seem irrational. Much as autonomous vehicles offer the prospect of significantly reducing the number of deaths and injuries caused by driver error behind the wheel, reducing mistakes and excesses on the battlefield has the potential to lessen the human costs of warfare. Many "dumb" devices are, in any case, already automated after a fashion. An anti-personnel landmine or an improvised explosive device (IED) operates without additional human control, though it is not selective in its targeting.[96] Heat-seeking missiles are an example of a weapon that, when launched, follows a program, but is not in a meaningful sense selective in its targeting. Further along the spectrum is a new generation of Long Range Anti-Ship Missiles (LRASM), which are launched with targeting parameters but able to search for and identify enemy warships within those parameters.[97]

As with autonomous vehicles, the key distinction in autonomous weapons is the degree to which the system makes decisions independently. According to the U.S. Department of Defense, an autonomous weapon system is one that, once activated, can select and engage targets without further intervention by a human operator.[98] Similarly, the International Committee of the Red

---

[95] Losing Humanity: The Case Against Killer Robots (Human Rights Watch, New York, 2012), *at* http://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots; Michael Press, *Of Robots and Rules: Autonomous Weapon Systems in the Law of Armed Conflict*, 48(4) GEO. J. INT'L L. 1337, 1344 (2017) (discussing arguments for a prohibition of autonomous weapon systems).

[96] Though anti-personnel mines were banned in many countries by the 1997 Ottawa Convention, they remain in use by many military powers. *See* Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and Their Destruction (Ottawa Convention), done at Oslo, Sept. 18, 1997 (in force Mar. 1, 1999).

[97] Sebastien Roblin, *LRASM: The Navy's Game Changer Missile Russia and China Should Fear?*, NAT'L INT., Apr. 21, 2018.

[98] Autonomy in Weapon Systems (Department of Defense, Washington, DC, Directive

Cross (ICRC) has emphasized that autonomy in this context should focus on the critical functions of selecting and attacking targets, as opposed to movement or navigation.[99]

There tends to be less concern about purely defensive systems. Close-in weapon systems (CIWS), such as the U.S. Navy's Phalanx CIWS, were first deployed in the 1970s as the last line of defense against an attack on a ship at sea. Such systems automatically detect and destroy incoming missiles or enemy aircraft at close proximity.[100] Land-based ballistic missile defense systems also have varying degrees of automation—most prominently the U.S. Patriot Missile and Israel's "Iron Dome," which identify and attempt to destroy rockets and artillery shells.[101] Stationary anti-personnel weapons, such as sentry guns, have been deployed in the Demilitarized Zone between North and South Korea, though their true degree of autonomy is disputed.[102]

Offensive autonomous weapons have yet to be widely deployed, but the technology is rapidly advancing in that direction. Various unmanned aerial vehicles, or drones, have the capacity for independent targeting; some are also able to suggest targets as well as angles of attack, though decisions to engage remain the positive responsibility of their operators.[103] Other land and sea-based combat vehicles have been developed with varying degrees of autonomy. Typically, these have been remote-controlled—though there are periodic breathless reports of killer robots deployed in theatre, as when the U.S. experimented in Iraq with a machine-gun tank system called

---

Number 3000.09, Nov. 21, 2012) (defining autonomous weapon system as a weapon system "that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation").

[99] Towards Limits on Autonomy in Weapon Systems (International Committee of the Red Cross, Geneva, Apr. 9, 2018), *at* http://www.icrc.org/en/document/towards-limits-autonomous-weapons ("the functions of weapon systems that are most relevant to legal obligations and ethical concerns within the scope of the Convention on Certain Conventional Weapons, namely autonomy in the critical functions of selecting and attacking targets. Autonomy in other functions (such as movement or navigation) would not in our view be relevant to the discussions").

[100] Similar systems include Russia's Kaftan CIWS and China's Type 730 CIWS.

[101] *See, e.g.,* Michael J. Armstrong, *Modeling Short-Range Ballistic Missile Defense and Israel's Iron Dome System*, 62(5) OPERATIONS RES. 1028 (2014).

[102] Ian Kerr and Katie Szilagyi, *Evitable Conflicts, Inevitable Technologies? The Science and Fiction of Robotic Warfare and IHL*, 14(1) LAW, CULTURE & HUM. 45, 52 (2018) (citing blog posts from 2006 and 2010); *Trying to Restrain the Robots; Autonomous Weapons*, ECONOMIST, Jan. 19, 2019 (stating that the system is no longer deployed).

[103] Kenneth Anderson and Matthew Waxman, Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can (Hoover Institution, Stanford, 2013), *at* http://media.hoover.org/sites/default/files/documents/Anderson-Waxman_LawAndEthics_r2_FINAL.pdf, at 4.

"SWORDS" in 2007.[104]

A decade later, the U.S. Army sparked controversy in 2019 when it put out a request for vendors to help build its Advanced Targeting and Lethality Automated System (ATLAS). The initial call said that the hope was to develop combat vehicles with the ability to "acquire, identify, and engage targets at least 3X faster than the current manual process." After news headlines announced that the Pentagon was about to turn its tanks into "AI-powered killing machines," the announcement was modified to emphasize that there had been no change in Department of Defense policy on autonomy in weapon systems.[105] That policy remains that autonomous weapon systems must allow commanders and operators to "exercise appropriate levels of human judgment" over the use of force.[106]

Many commentators accept that an increasing degree of autonomy on the battlefield is inevitable, and that the superiority of autonomous weapon systems over humans is inevitable also.[107] Yet the view that the finger on the trigger must be flesh and blood rather than metal and silicon is widely held, and points to something qualitatively different from debates over autonomy in transportation.

A. International Humanitarian Law

In contrast to many of the legal regimes considered in this article, international humanitarian law explicitly provides for its application to new and emerging technologies. This provision takes the form of the Martens Clause, named after the Russian delegate who introduced it at the 1899 Hague Peace Conference. The text made its way into the preamble of the Convention on the Laws and Customs of War in the following form:

> Until a more complete code of the laws of war is issued, the High Contracting Parties think it right to declare that in cases not included in the

---

[104] Noah Shachtman, *First Armed Robots on Patrol in Iraq (Updated)*, WIRED, Aug. 2, 2007. The Special Weapons Observation Reconnaissance Detection System (SWORDS) was essentially a repurposed remote-controlled bomb disposal unit.

[105] Industry Day for the Advanced Targeting and Lethality Automated System (ATLAS) Program (Department of the Army, Belvoir, Solicitation Number: W909MY-19-R-C004, Feb. 11, 2019). *See* Justin Rohrlich, *The U.S. Army Wants to Turn Tanks into AI-Powered Killing Machines*, QUARTZ, Feb. 26, 2019.

[106] Autonomy in Weapon Systems, *supra* note 98, para 4a.

[107] As early as 2001, for example, the U.S. Congress set the goal of making one-third of combat aircraft unmanned by 2010 and one-third of combat ground vehicles unmanned by 2015: Floyd D. Spence National Defense Authorization Act for Fiscal Year 2001, §220. *See* also Ian Kerr and Katie Szilagyi, *Asleep at the Switch? How Killer Robots Become a Force Multiplier of Military Necessity*, *in* ROBOT LAW 333, 334 (Ryan Calo, A. Michael Froomkin, and Ian Kerr eds., 2016).

> Regulations adopted by them, populations and belligerents remain under the protection and empire of the principles of international law, as they result from the usages established between civilized nations, from the laws of humanity, and the requirements of the public conscience.[108]

Over the subsequent decades, text that was originally a cunning diplomatic maneuver to break a deadlock came to be invested with far greater significance—at times treated as though it created a new source of law, rather than an interpretive tool to be applied in cases of uncertainty.[109] When the International Court of Justice was asked to consider the legality of the threat or use of nuclear weapons, for example, it noted that the Martens Clause—now enshrined in article 1(2) of the First Additional Protocol to the Geneva Conventions[110]—made clear that the "principles and rules of humanitarian law" applied to such weapons, notwithstanding the lack of a specific treaty to that effect.[111] Indeed, it found that the clause had proved to be an effective means of addressing the rapid evolution of military technology.[112]

Though some writers have argued that the Martens Clause itself is a basis for outlawing autonomous weapons,[113] that goes well beyond its normal use as an interpretive tool to address uncertainty or *lacunae* in the law. The use of such weapons would be subject to the principles and rules of humanitarian law, but it goes too far to conclude that they are unlawful as such because of the clause alone.[114]

Applying those principles and rules to new technology is not a simple task. It is sometimes argued that computers should not be empowered to make life and death decisions because of the "infinite number of possible

---

[108] Convention (II) with Respect to the Laws and Customs of War on Land and Its Annex: Regulations Concerning the Laws and Customs of War on Land (1899 Hague Regulations), done at The Hague, July 29, 1899, preamble.

[109] Antonio Cassese, *The Martens Clause: Half a Loaf or Simply Pie in the Sky?*, 11 EUR. J. INT'L L. 187, 212-14 (2000).

[110] Protocol Additional to the Geneva Conventions of Aug. 12, 1949, and relating to the Protection of Victims of International Armed Conflicts (Additional Protocol I), June 8, 1977, *at* http://www.icrc.org/ihl, art 1(2): "In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience."

[111] *Legality of the Threat or Use of Nuclear Weapons* (International Court of Justice, Advisory Opinion, July 8, 1996), *at* http://www.icj-cij.org, para 87.

[112] *Id*., para 78.

[113] Advancing the Debate on Killer Robots: 12 Key Arguments for a Preemptive Ban on Fully Autonomous Weapons (Human Rights Watch, New York, May 2014), *at* http://www.hrw.org/news/2014/05/13/advancing-debate-killer-robots, at 14-17.

[114] Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837, 1879-81 (2015).

scenarios" in which such decisions might be made.[115] This is one of the weaker arguments against autonomy, as the underlying concern is not the ability of an AI system to respond to limitless scenarios, but of a human to be able to program them in advance. Indeed, some commentators argue that AI systems may be *more* capable of compliance with the laws of war than their human counterparts.[116] Unlike humans, who must be trained, autonomous weapon systems could have these rules programmed in and be required to act on them without emotion. Many war crimes arise not from conscious decisions to violate the rules of engagement, but as a result of fatigue, fear, or anger—precisely the qualities that machines are built to avoid.[117]

Another set of concerns recall the nontrivial possibility that a truly intelligent system in the sense of general AI might decide that humans were its enemy.[118] The prospect of an autonomous weapon system turning on its creator is one of the more visceral images of the threat of AI—epitomized and immortalized in the various *Terminator* movies. Though nothing quite so dramatic has yet occurred on the battlefield, there have been incidents of friendly fire by autonomous systems that experienced targeting errors[119] or engaged friendly craft that came within the system's engagement envelope.[120]

In some cases, those involved in the development of AI systems have expressed a simple aversion to being involved in military projects at all. When Google's role in the U.S. Department of Defense's Project Maven was revealed, thousands of employees signed a letter demanding that Google withdraw from the project and commit that neither the company nor its

---

[115] Shaking the Foundations: The Human Rights Implications of Killer Robots (Human Rights Watch, New York, 2014), *at* http://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots.

[116] *See, e.g.,* Kenneth Anderson, Daniel Reisner, and Matthew Waxman, *Adapting the Law of Armed Conflict to Autonomous Weapon Systems*, 90 INT'L L. STUD. 386, 411 (2014); PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD 280 (2015).

[117] Ronald C. Arkin, *The Case for Ethical Autonomy in Unmanned Systems*, 9(4) J. MIL. ETHICS 332 (2010); Kenneth Anderson and Matthew C. Waxman, *Debating Autonomous Weapon Systems, Their Ethics, and Their Regulation Under International Law*, *in* THE OXFORD HANDBOOK OF LAW, REGULATION, AND TECHNOLOGY 1097, 1108-10 (Roger Brownsword, Eloise Scotford, and Karen Yeung eds., 2017).

[118] *See, e.g.,* Matthew Sparkes, *Top Scientists Call for Caution over Artificial Intelligence*, THE TELEGRAPH, Jan. 13, 2015; Peter Holley, *Bill Gates on Dangers of Artificial Intelligence: "I Don't Understand Why Some People Are not Concerned,"* WASH. POST, Jan. 29, 2015; Peter Holley, *Elon Musk's Nightmarish Warning: AI Could Become "An Immortal Dictator From Which We Would Never Escape,"* WASH. POST, Apr. 6, 2018.

[119] A.J. Plunkett, *Iwo Jima Officer Killed in Firing Exercise*, DAILY PRESS, Oct. 12, 1989.

[120] Philip Shenonjune, *Japanese Down Navy Plane In an Accident; Crew Is Safe*, N.Y. TIMES, June 5, 1996.

contractors would ever build "warfare technology."[121] It would be tempting to dismiss this as the conceit of employees working for a company whose slogan was once "don't be evil,"[122] but such "twinges of indignation"[123] are apparent in many aspects of the autonomous weapon systems debates.

B. Human-out-of-the-Loop?

Central to many of the worries expressed is that dissociation from the choice of whom to kill weakens the moral dilemma that should accompany all such decisions.[124] One could argue that this applies to other sanitized military operations—from launching a cruise missile against faceless targets to the drone operator at an army base who goes home for dinner.[125] The distinction of truly autonomous weapon systems, however, is that in addition to being physically absent from the battlefield, handing over life-and-death decisions to algorithms would mean that human operators would be psychologically absent also.[126] In a 2018 speech to the General Assembly, UN Secretary-General António Guterres denounced this prospect as "morally repugnant."[127]

With regard to lethal force, it is often argued, the decision whether to use it should be made by a human—and it should be possible to hold that human accountable for his or her actions afterwards. This view is based on the conception of warfare itself as an intimately human institution. As Michael

---

[121] Scott Shane and Daisuke Wakabayashi, *"The Business of War": Google Employees Protest Work for the Pentagon*, N.Y. TIMES, Apr. 4, 2018. Project Maven focuses on computer vision, using machine learning and deep learning to extract objects of interest from moving or still imagery: Cheryl Pellerin, Project Maven to Deploy Computer Algorithms to War Zone by Year's End (Department of Defense, Washington, DC, July 21, 2017).

[122] KEN AULETTA, GOOGLED: THE END OF THE WORLD AS WE KNOW IT 20 (2009); STEVEN LEVY, IN THE PLEX: HOW GOOGLE THINKS, WORKS, AND SHAPES OUR LIVES 144 (2011) (describing the origin of the slogan).

[123] Helen Nissenbaum, *Protecting Privacy in an Information Age: The Problem of Privacy in Public*, 17 LAW & PHIL. 559, 583 (1998) (describing the manner in which intuition as to social norms can provide a basis for some forms of regulation).

[124] *See, e.g.,* WENDELL WALLACH, A DANGEROUS MASTER: HOW TO KEEP TECHNOLOGY FROM SLIPPING BEYOND OUR CONTROL 213-34 (2015); Tetyana Krupiy, *Of Souls, Spirits and Ghosts: Transposing the Application of the Rules of Targeting to Lethal Autonomous Robots*, 16 MELB. J. INT'L L. 145, 201 (2015).

[125] For an early exploration of this question, *see* JEAN BAUDRILLARD, THE GULF WAR NEVER HAPPENED (1995).

[126] Christof Heyns, *Autonomous Weapons Systems: Living a Dignified Life and Dying a Dignified Death*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 3, 4 (Nehal Bhuta, et al. eds., 2016).

[127] António Guterres, *Address to the General Assembly* (New York, United Nations, Sept. 25, 2018), *at* http://www.un.org/sg/en/content/sg/speeches/2018-09-25/address-73rd-general-assembly.

Walzer has observed:

> It is one of the most important features of war, distinguishing it from the other scourges of mankind, that the men and women caught up in it are not only victims, they are also participants. All of us are inclined to hold them responsible for what they do.[128]

Were autonomous weapon systems to become widespread, the costs of war could be reduced—even more than they have been already in industrialized countries—to technical and material constraints. The juxtaposition of such systems with human adversaries, cold logic versus mortal fear, would, the argument continues, be corrosive of the equal dignity of humans.[129] It also suggests the likely progression of an autonomous weapons arms race: once such systems are deployed by one side, it would be difficult to justify sending human soldiers into battle against them.[130]

At the international level, opposition to autonomous weapon systems has tended to vary inversely with capacity. There is some support for a complete treaty ban among a handful of states,[131] but without the involvement of states possessing advanced technological and military capabilities that would be posturing at best. Scholars from the Military Law Institute at the China University of Political Science and Law, for example, have argued that states with advanced AI technology should play an "exemplary" role—going on to propose that a military commander or civilian official who employs a weapon system operating with "full autonomy" should bear personal responsibility for violations of IHL that ensue.[132]

---

[128] MICHAEL WALZER, JUST AND UNJUST WARS: A MORAL ARGUMENT WITH HISTORICAL ILLUSTRATIONS 15 (3rd ed. 2000).

[129] Nehal Bhuta, Susanne Beck, and Robin Geiβ, *Present Futures: Concluding Reflections and open Questions on Autonomous Weapons Systems*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 347, 355-56 (Nehal Bhuta, et al. eds., 2016).

[130] Leonard Kahn, *Military Robots and the Likelihood of Armed Conflict*, *in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 274, 283 (Patrick Lin, Keith Abney, and Ryan Jenkins eds., 2017). *See* also Ingvild Bode and Hendrik Huelss, *Autonomous Weapons Systems and Changing Norms in International Relations*, 44(3) REV. INT'L STUD. 393 (2018) (describing how the use of autonomous weapons may redefine notions of "appropriateness" in conflict).

[131] *See* the list maintained at Country Views on Killer Robots (Campaign to Stop Killer Robots, Washington, DC, Nov. 13, 2018), *at* http://www.stopkillerrobots.org/wp-content/uploads/2018/11/KRC_CountryViews13Nov2018.pdf.

[132] LI Qiang and XIE Dan, *Legal Regulation of AI Weapons Under International Humanitarian Law: A Chinese Perspective*, ICRC HUMANITARIAN LAW & POLICY BLOG, May 2, 2019, *at* http://blogs.icrc.org/law-and-policy/2019/05/02/ai-weapon-ihl-legal-regulation-chinese-perspective/ ("In the case of full autonomy of AI weapon systems without any human control, those who decide to employ AI weapon systems—normally senior military commanders and civilian officials—bear individual criminal responsibility for any

Two areas of ongoing discussion are in the context of weapons reviews and a possible requirement of "meaningful human control." Article 36 of the First Additional Protocol to the Geneva Conventions provides that the "study, development, acquisition or adoption of a new weapon, means or method of warfare" requires states parties to determine whether its use would violate international law.[133] This has been endorsed by the UN Group of Governmental Experts examining lethal autonomous weapon systems as a potential guiding principle in this area.[134] Though some have argued that Article 36 reflects customary international law,[135] the ICRC has held that such reviews of new weapons are necessary in any event as part of a "faithful and responsible" application of compliance with international law obligations.[136] The United States, for its part, introduced comparable processes three years before Protocol I came into force[137] and declined to develop blinding laser weapons in the 1990s after such a review.[138]

Key considerations in determining whether use of a weapon would violate international law tend to focus on the rules against weapons that are inherently indiscriminate or that cause unnecessary suffering or superfluous injury.[139] Though it has been argued that autonomous weapon systems are

---

potential serious violations of IHL. Additionally, the States to which they belong incur State responsibility for such serious violations which could be attributable to them.").

[133] Additional Protocol I, *supra* note 110, art. 36.

[134] Report of the 2018 Group of Governmental Experts on Lethal Autonomous Weapons Systems, UN Doc. CCW/GGE.2/2018/3 (2018), *at* http://undocs.org/en/CCW/GGE.1/2018/3, para. 26(d). *See* also Anderson and Waxman, *supra* note 117, at 1104-05.

[135] *See, e.g.,* Losing Humanity, *supra* note 95, at 21; Ryan Poitras, *Article 36 Weapons Reviews & Autonomous Weapons Systems: Supporting an International Review Standard*, 34(2) AM. U. INT'L L. REV. 465, 470-71 (2018).

[136] A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977 (International Committee of the Red Cross, Geneva, Jan. 2006), *at* http://www.icrc.org/en/doc/assets/files/other/icrc_002_0902.pdf, at 4.

[137] *Id*. at 4n6. *See* now The Defense Acquisition System (Department of Defense, Washington, DC, Directive Number 5000.01, May 12, 2003), *at* http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/500001p.pdf, para E1.1.15. ("The acquisition and procurement of DoD weapons and weapon systems shall be consistent with all applicable domestic law and treaties and international agreements …, customary international law, and the law of armed conflict (also known as the laws and customs of war). An attorney authorized to conduct such legal reviews in the Department shall conduct the legal review of the intended acquisition of weapons or weapons systems.")

[138] Anderson and Waxman, *supra* note 103, at 10.

[139] *See, e.g.,* A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977, *supra* note 136, at 15-16 (describing treaty and customary law obligations); Anderson and Waxman, *supra* note 103, at 10 (describing reviews over the past two decades); Poitras, *supra* note 135, at 473-75 (adding a third consideration of whether a weapon causes widespread, long

necessarily indiscriminate because they lack the human qualities necessary to identify combatants and assess the intentions of other humans,[140] these are practical challenges to the sensory and analytical capabilities of such systems.[141] Similarly, it has been argued that a machine will be unable to distinguish incapacitated or surrendering enemies from legitimate targets and thus will cause unnecessary suffering.[142] Again, this appears to be a surmountable problem[143]—comparable, perhaps, to some of the challenges facing autonomous vehicles navigating among human drivers and pedestrians.[144]

Of greater importance, in the context of autonomous weapon systems, is not the capabilities of the machines but the absence of humans. The ICRC issued a statement in 2018 that emphasized the importance of human involvement—not because of a superior capacity to identify or understand other humans, but to grapple meaningfully with the moral dilemma of whether force should be used and to take responsibility if it is:

> International humanitarian law (IHL) requires that those who plan, decide upon and carry out attacks make certain judgements in applying the norms when launching an attack. Ethical considerations parallel this requirement – demanding that human agency and intention be retained in decisions to use force. Humans therefore bear responsibilities in the programming, development, activation and operational phases of autonomous weapon systems.
>
> Mindful of the potential adverse humanitarian consequences of the loss of human control over weapons and the use of force, the ICRC has posited that a minimum level of human control is necessary from both a legal and ethical perspective. In the view of the ICRC, a weapon system beyond human control would be unlawful by its very nature.[145]

Despite resistance to an outright ban on autonomous weapons, calls for

---

term, and severe damage to the environment).

[140] Losing Humanity, *supra* note 95, at 30-32.

[141] Poitras, *supra* note 135, at 486-89.

[142] Losing Humanity, *supra* note 95, at 34-35.

[143] Poitras, *supra* note 135, at 482-86.

[144] *Cf.* Noel Sharkey, *Staying in the Loop: Human Supervisory Control of Weapons*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 23, 24-27 (Nehal Bhuta, et al. eds., 2016) (outlining difficulties in overcoming technical challenges to autonomous weapon systems having sufficient capacity to comply with the laws of war).

[145] Towards Limits on Autonomy, *supra* note 99. *See* also Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach (International Committee of the Red Cross, Geneva, June 6, 2019), *at* http://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach, at 7-10 (outlining the legal and ethical basis for ongoing human involvement).

"meaningful human control" have gained traction—even though such control may be inconsistent with a weapons system that is truly autonomous. At present, the lowest common denominator appears to be a possible ban on fully autonomous weapons that operate in such a manner that their mission, once started, cannot be aborted. The prospect of such truly "human-out-of-the-loop" machines running loose even after the conclusion of hostilities appears sufficient—for the time being, at least—to outweigh the benefits of such weapons being on the battlefield.[146]

C. Lessons from Mercenaries

As in many other aspects of regulating AI systems, there has been a tendency to view the problems posed by autonomous weapon systems as new and unique. This overlooks important analogies that can be drawn from other activities that have raised similar concerns. In particular, lessons may be drawn from efforts over the past three decades to regulate the outsourcing of warfighting capacity not to machines but to mercenaries.

Modern wariness about mercenaries and their corporate cousins, private military and security companies (PMSCs)—in particular their ability to use lethal force—stems from a belief that such decisions should be made within a framework that allows not merely legal but also political and moral accountability.[147] Today it is "common sense" that the control and use of violence should be limited to states. But it was not always so. The Pope, for example, is still protected by a private Swiss regiment first hired in 1502. Echoes of the past acceptability of mercenarism also live on in our language. The term "freelance," for example, now means a casual worker, but historically it referred literally to a free agent in possession of a lance.[148]

Interestingly, the popularity of or disdain for mercenaries has depended on the shifting importance of military skill and military numbers, with a major influence being emergent technology. The introduction of the musket two centuries ago vastly reduced the time it took to train an effective soldier, with the result that quantity soon mattered more than quality. In such

---

[146] Amitai Etzioni and Oren Etzioni, *Pros and Cons of Autonomous Weapons Systems*, MIL. REV., May-June 2017, 71, 79-80. For a discussion of "human-out-of-the-loop" and other decision-making paradigms, *see infra* notes 156-158.

[147] *See* FROM MERCENARIES TO MARKET: THE RISE AND REGULATION OF PRIVATE MILITARY COMPANIES (Simon Chesterman and Chia Lehnardt eds., 2007).

[148] *See generally* SARAH PERCY, MERCENARIES: THE HISTORY OF A NORM IN INTERNATIONAL RELATIONS (2007). On the decline of the Weberian notion of the state as enjoying a monopoly over the legitimate use of force, *see* Kevin A. O'Brien, *What Should and What Should Not Be Regulated?*, *in* FROM MERCENARIES TO MARKET: THE RISE AND REGULATION OF PRIVATE MILITARY COMPANIES 29, 33 (Simon Chesterman and Chia Lehnardt eds., 2007).

circumstances, national conscription offered a more efficient means of raising a large army. Such military and economic shifts were then reinforced by politics and culture, with the result that mercenaries "went out of style" in the nineteenth century.[149] Reliance on mercenaries soon came to be seen not only as inefficient but suspect: a country whose men would not fight for it lacked patriots; those individuals who would fight for reasons other than love of country lacked morals.[150]

The subversive role of mercenaries in Africa during the twentieth century led to efforts to ban them completely. A 1989 treaty sought to do just that, but foundered on a lack of signatures and problems of definition. A mercenary was defined as someone "motivated to take part in the hostilities essentially by the desire for private gain."[151] The difficulty of proving such motivation led one writer to suggest that anyone convicted of an offence under the Convention should be shot—as should his lawyer.[152]

This approach may be contrasted with an initiative led by the Swiss Government and the ICRC, which focused not on imposing criminal liability on the mercenary but highlighting ongoing obligations of the state. A series of intergovernmental meetings led to the drafting of the Montreux Document, named after the town on Lake Geneva at which government experts met over three days in September 2008. It stresses the non-transferability of state obligations under international law, which encompasses ongoing responsibility for outsourced activities—and a prohibition on outsourcing some activities completely.[153]

A better and more useful distinction to be drawn, then, and of relevance to the discussion here, is that some functions are "inherently governmental" and cannot be transferred to contractors, machines, or anyone else.[154]

---

[149] Deborah Avant, *From Mercenary to Citizen Armies: Explaining Change in the Practice of War*, 54 INT'L ORG. 41 (2000).

[150] *See generally* DEBORAH AVANT, THE MARKET FOR FORCE: THE CONSEQUENCES OF PRIVATIZING SECURITY (2005); PERCY, *supra* note 148.

[151] International Convention Against the Recruitment, Use, Financing, and Training of Mercenaries (Convention on Mercenaries), Dec. 4, 1989 (in force Oct. 20, 2001), *at* http://www.un.org/documents/ga/res/44/a44r034.htm, art. 1(1)(b).

[152] Geoffrey Best, quoted in DAVID SHEARER, PRIVATE ARMIES AND MILITARY INTERVENTION 18 (1998).

[153] The Montreux Document on Pertinent International Legal Obligations and Good Practices for States Related to Operations of Private Military and Security Companies During Armed Conflict (Swiss Federal Department of Foreign Affairs & International Committee of the Red Cross, Montreux, Sept. 17, 2008), *at* http://www.eda.admin.ch/psc.

[154] *See* Simon Chesterman, *"We Can't Spy… If We Can't Buy!": The Privatization of U.S. Intelligence Services and the Limits of Outsourcing "Inherently Governmental" Functions*, 19 EUR. J. INT'L L. 1055 (2008).

III. BLACK BOX DECISION-MAKING AND LEGITIMACY

Autonomous actions by AI systems are not limited to their physical interactions with the world. Though driverless cars and killer robots conjure the visceral image of machines displaying independence, underlying that autonomy is a capacity to gather data and take decisions with far wider applications. As ever more commercial and governmental activity moves online, vast numbers of routine tasks can be managed without human involvement. A growing number of decisions are now made essentially by algorithms, either reaching conclusive determinations or presenting a proposed decision that may be rarely if ever questioned by the human notionally responsible.[155]

As in the case of autonomous vehicles, it is useful to distinguish between levels of autonomy in decision-making. A commonly used metaphor here is of a human being in, over, or out of a decision-making process referred to as a "loop." At one extreme is fully human decision-making without computer support. Recalling the SAE levels for autonomous vehicles discussed earlier, this would be akin to level zero.[156] "Human-in-the-loop" refers to decision-making supported by the system, for example through suggesting options or recommendations, but with the human taking positive decisions. That may correspond to SAE level one or two ("hands on the wheel"). "Human-over-the-loop" denotes a process in which the human can oversee the process and make interventions as necessary, corresponding to SAE level three or four.[157] "Human-out-of-the-loop" means the process runs with minimal or no human intervention, akin to SAE level five.[158]

Another distinction can be made between algorithmic processes broadly comparable to deductive as opposed to inductive reasoning. The first is the application of pre-programmed, human-authored rules. At its most basic, this could include simple computation, such as the totaling of a grocery bill at an automated checkout; or it could be the application of a set of variables to

---

[155] *Cf.* Tarleton Gillespie, *Algorithm*, *in* DIGITAL KEYWORDS: A VOCABULARY OF INFORMATION SOCIETY AND CULTURE 18, 26 (Benjamin Peters ed., 2016): "What is central is the commitment to procedure, and the way procedure distances its human operators from both the point of contact with others and the mantle of responsibility for the intervention they make."

[156] *See supra* text accompanying note 18.

[157] Austin Graham et al., *Formalizing Interruptible Algorithms for Human Over-the-Loop Analytics* (Boston, MA, 2017 IEEE International Conference on Big Data (Big Data), 2017).

[158] *Cf.* Natasha Merat et al., *The "Out-of-the-Loop" Concept in Automated Driving: Proposed Definition, Measures and Implications*, 21(1) COGNITION, TECHNOLOGY & WORK 87 (2019). *See* also Karen Yeung, *Algorithmic Regulation: A Critical Interrogation*, 12 REG. & GOVERNANCE 505, 508 (2018) (developing a taxonomy of algorithmic regulation).

determine eligibility for government benefits or the interest rate for a loan. Such rules-based decision-making would not normally be seen as truly "autonomous." An alternative form of decision-making is the use of tools to make inferences or predictions based on historical data, such as through machine learning.[159] As those tools become more complex, the difficulty of understanding or explaining the reasons behind decisions may raise problems of opacity.[160] In this article, the focus is on the autonomy with which those tools reach conclusions that cannot be attributed back directly to a human author.

The manner in which the algorithm is constructed matters also. For rules-based processing, those rules must be interpreted. If they are based on a law that says "if circumstances A and B are satisfied, then conclusion C follows," this may be unproblematic. Laws are rarely so simple, however.[161] In Australia, for example, a 2015 program referred to as "Robo-debt" sought to calculate and collect debts owed because of welfare overpayments. Though it applied rules systematically, these rules were incomplete transcriptions of complex provisions in the law and resulted in around one in five people being incorrectly served with debt notices.[162]

In the case of machine learning, the AI system relies upon data that itself may or may not be reliable. Basing future decisions on the assumption that past decisions were correct runs the risk that errors or biases will become policies.[163] In one well-known example, Amazon developed a résumé screening algorithm trained on ten years of data, but had to shut it down when programmers discovered that it had "learned" that women's applications were to be regarded less favorably than men's.[164]

For many cases, the use of algorithms to support or replace human

---

[159] *Cf.* Monika Zalnieriute, Lyria Bennett Moses, and George Williams, *The Rule of Law and Automation of Government Decision-Making*, 82 MOD. L. REV. 425, 427 (2019). For a definition of machine learning, *see supra* note 4.

[160] On the problem of opacity in AI systems, *see* PASQUALE, *supra* note 10; Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1 (2016); Sandra Wachter, Brent Mittelstadt, and Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31(2) HARV. J.L. & TECH. 841 (2018).

[161] *See* Mireille Hildebrandt and Bert-Jaap Koops, *The Challenges of Ambient Law and Legal Protection in the Profiling Era*, 73(3) MOD. L. REV. 428, 429 (2010) (arguing that law might need to be remodeled so as to form part of an expanding socio-technological infrastructure).

[162] *See* Zalnieriute, Bennett Moses, and Williams, *supra* note 159, at 446.

[163] Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, [2016] BIG DATA & SOC'Y, 7-8 (2016).

[164] Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS, Oct. 10, 2018; Maude Lavanchy, *Amazon's Sexist Hiring Algorithm Could Still Be Better than a Human*, THE CONVERSATION, Nov. 1, 2018.

decision-making is uncontroversial. In addition to efficiency, automated processing may help ensure consistency and predictability. Indeed, in some situations it may be preferable to the arbitrariness that often characterizes human decision-making—whether that is due to conceptual limitations, carelessness, or corruption. At the same time, abdicating responsibility for decisions to a machine raises the possibility of other problems, ranging from latent discrimination to a lack of due process or procedural fairness. In between lies the question of how discretion should be exercised and whether, comparable to the debate over autonomous weapons, there are some decisions that simply should not be made by machine alone.

A. Private Sector

Vast numbers of routine commercial transactions now take place without any human intervention whatsoever, from purchasing items online to arguing with chatbots if those items do not arrive or are defective. The push to automate decision-making processes is greatest in areas that are high volume and low risk. In addition to online purchases, this has extended to areas such as small loans, retail insurance, and recruitment screening, where varying degrees of automation have introduced efficiencies for businesses.[165] A growing number of companies use automated dispute resolution systems, with eBay said to resolve more than 60 million such disputes annually.[166]

Much of the regulatory intervention in this space has focused on the problem of discrimination and will be discussed in the context of European efforts to limit the impact of automated processing.[167] Such efforts seek to prevent automation violating rights in a manner that would be impermissible if those decisions were being taken by a human. In this Part, the focus is on novel challenges posed by the autonomy of the algorithms.

As in the case of autonomous vehicles,[168] most private law questions involving algorithms can be resolved using existing laws and principles. Occasionally, however, there may be odd results when those methods are applied to new fact patterns. Increased reliance on algorithmic trading software, for example, has given rise to the phenomenon of computer programs concluding deals with one another that may move beyond their initial parameters. The validity of such contracts is not especially

---

[165] Stefanie Hänold, *Profiling and Automated Decision-Making: Legal Implications and Shortcomings*, *in* ROBOTICS, AI AND THE FUTURE OF LAW 123, 127-28 (Marcelo Corrales, Mark Fenwick, and Nikolaus Forgó eds., 2018).

[166] PABLO CORTÉS, THE LAW OF CONSUMER REDRESS IN AN EVOLVING DIGITAL MARKET: UPGRADING FROM ALTERNATIVE TO ONLINE DISPUTE RESOLUTION 8 (2017).

[167] *See* Section III.C.

[168] *See* Part I.

complicated,[169] though high-frequency trading may pose practical challenges to implementation.[170]

A novel problem directly tied to autonomy did, however, arise in a 2019 case before the Singapore International Commercial Court. The parties, Quoine and B2C2, used software programs that executed trades involving the cryptocurrencies Bitcoin and Ethereum, with prices set according to external market information. The case focused on seven trades that were made when a defect in Quoine's software saw it execute trades worth approximately $12m at 250 times the prevailing exchange rate.[171] Quoine claimed that this was a mistake and attempted to reverse the trades, reclaiming its losses. B2C2 argued that the reversal of the orders was a breach of contract, while Quoine argued that the contract was void or voidable, relying on the doctrine of unilateral mistake.[172]

At common law, a unilateral mistake can void a contract if the other party knows of the mistake.[173] If it cannot be proven that the other party *actually* knew about the mistake, but it can be shown that he or she *should have*, the contract may be voidable under equity.[174] What became crucial in this case was the judge's finding that the computer programs in question were incapable of "knowing" anything:

> The algorithmic programs in the present case are deterministic, they do and only do what they have been programmed to do. They have no mind of their own. They operate when called upon to do so in the pre-ordained manner. They do not know why they are doing something or what the external events are that

[169] *See* FAYE FANGFEI WANG, LAW OF ELECTRONIC COMMERCIAL TRANSACTIONS: CONTEMPORARY ISSUES IN THE EU, U.S. AND CHINA (2nd ed. 2014).

[170] *See, e.g.,* Megan Woodward, *The Need for Speed: Regulatory Approaches to High Frequency Trading in the United States and the European Union*, 50 VAND. J. TRANSNAT'L L. 1359 (2011); E. Wes Bethel et al., *Federal Market Information Technology in the Post Flash Crash Era: Roles for Supercomputing*, 7(2) J. TRADING 9 (2012); MICHAEL LEWIS, FLASH BOYS: A WALL STREET REVOLT (2014); Andrei Kirilenko et al., *The Flash Crash: High-Frequency Trading in an Electronic Market*, 72 J. FIN. 967 (2017); David R. Meyer and George Guernsey, *Hong Kong and Singapore Exchanges Confront High Frequency Trading*, 23(1) ASIA PAC. BUS. REV. 63 (2017); Tilen Čuk and Arnaud van Waeyenberge, *European Legal Framework for Algorithmic and High Frequency Trading (Mifid 2 and MAR) A Global Approach to Managing the Risks of the Modern Trading Paradigm*, 9 EUR. J. RISK REG. 146 (2018).

[171] Gary Low and Terence Tan, Unilateral Mistake and Algorithmic Trading (Drew & Napier, Singapore, Mar. 25, 2019).

[172] *B2C2 Ltd v Quoine Pte Ltd*, [2019] SGHC(I) 3 (2019).

[173] John Cartwright, *Unilateral Mistake in the English Courts: Reasserting the Traditional Approach*, [2009] SING. J.L.S. 226 (2009).

[174] YEO Tiong Min, *Unilateral Mistake in Contract: Five Degrees of Fusion of Common Law and Equity*, [2004] SING. J.L.S. 227, 231-33 (2004).

cause them to operate in the way that they do.[175]

As a result, the question of knowledge rested with the original programmer of B2C2's software, who could not have known about Quoine's subsequent mistake. Quoine was therefore liable to pay damages to B2C2.[176]

The finding was consistent with existing law, but the judge was careful to confine himself to the facts at hand, noting that the law may need to develop with technology—in particular, if a future computer could be said to have "a mind of its own."[177] He clearly viewed this as an incremental process, however, citing with approval the somewhat optimistic statement of Lord Briggs in a UK Supreme Court decision the previous year: "The court is well versed in identifying the governing mind of a corporation and, when the need arises, will no doubt be able to do the same for robots."[178]

Knowledge also plays a role in the criminal law. Another curious example of automated decision-making is "Random Darknet Shopper," the brainchild of two Swiss artists. Given a budget of up to $100 per week in Bitcoin, this is an automated online shopping bot that randomly chooses and purchases items from the deep web that are mailed directly to an exhibition space. An interesting legal puzzle was created when it came to the attention of the St. Gallen police that the bot's meandering through the unindexed portions of the Internet had led it to purchase a bag of ecstasy pills. The entire exhibition was seized, but the public prosecutor later decided that the incident was "within the realm of art" and disposed of the drugs without pressing charges.[179]

B. Public Authorities

Like the private sector, many governments have sought efficiencies through automation. The difference is that the exercise of public authority typically requires not only efficiency in its outcomes but legitimacy in its processes. The most basic problems have arisen when rules-based processing

---

[175] *B2C2 Ltd v Quoine Pte Ltd*, para 208.

[176] The judge viewed the software as carrying out actions that could have been carried out by a human and that it was necessary to look at the intention and knowledge of the operator or controller of the machine to determine the intention and knowledge of the machine. In this case, the programmer, a Mr. Tseung, believed that the computer would only force-close in circumstances that would have led to a margin call if everything was operating properly, hence there was no "mistake." *Id*., paras 210, 221-222.

[177] *Id*., para 206. *Cf.* Shawn Bayern, *Artificial Intelligence and Private Law*, *in* RESEARCH HANDBOOK ON THE LAW OF ARTIFICIAL INTELLIGENCE 144, 151-52 (Woodrow Barfield and Ugo Pagallo eds., 2018).

[178] *Warner-Lambert Co Ltd v Generics (U.K.) Ltd*, [2018] UKSC 56 (2018), para 165.

[179] Katie Grant, *Random Darknet Shopper: Exhibition Featuring Automated Dark Web Purchases Opens in London*, INDEPENDENT, Dec. 12, 2015.

does not correspond directly to the underlying legal basis for the activity, as in the case of the Robo-debt mentioned earlier.[180] For machine learning algorithms, the possibility of discrimination will be considered in the context of European responses below.[181] Distinct from private sector activities, the inclusion of non-discriminatory but irrelevant information may also undermine public authority decisions. A private company may choose to hire someone because his name is Jared,[182] for example, but that should not affect his ability to receive government benefits.

A preliminary issue is that with some AI systems it may not be possible to identify precisely the grounds of a decision.[183] This may render review of a decision meaningless if, for example, the correctness of underlying information and the weight attributed to it cannot be determined.[184] More generally, a decision made by algorithm may also have the effect of reversing the burden of proof by creating the illusion of an objective decision that a petitioner must argue against.[185]

In certain decisions by public bodies, legislation specifically requires the involvement of a human decision-maker. Under the English Taxes Management Act, for example, a notice to pay tax may be issued by "an officer of the Board."[186] A taxpayer charged with late filing objected on the basis that the notice sent to him was computer generated, lacking a signature or even a name. The judge concluded that the specific language required that the decision be made by "a real 'flesh and blood' officer, and not by [the tax authority] as a collective body. Nor is it a computerized decision."[187] Though such decisions were not themselves unlawful, in this case at least an identifiable public officer was required to make the determination.

Similarly, in most jurisdictions the judicial function must be carried out by a human officer of the court. Though online dispute resolution is becoming more common in small claims tribunals[188] and predictive algorithms

---

[180] *See supra* text accompanying note 162.

[181] *See infra* Section III.C.

[182] Stéphanie Thomson, Here's Why You Didn't Get that Job: Your Name (World Economic Forum, Geneva, May 23, 2017), *at* http://www.weforum.org/agenda/2017/05/job-applications-resume-cv-name-descrimination/.

[183] *See supra* note 160.

[184] Zalnieriute, Bennett Moses, and Williams, *supra* note 159, at 449.

[185] Mittelstadt et al., *supra* note 163, at 8.

[186] Taxes Management Act 1970 (England), s. 8.

[187] *Peter Groves v The Commissioners for Her Majesty's Revenue & Customs* (First-Tier Tribunal Tax Chamber, Appeal number: TC/2017/09024, June 15, 2018), (2018).

[188] *See, e.g.,* Trish O'Sullivan, *Developing an Online Dispute Resolution Scheme for New Zealand Consumers Who Shop Online: Are Automated Negotiation Tools the Key to Improving Access to Justice?*, 24(1) INT'L J.L. & INFO. TECH. 22 (2016); Jeremy Barnett and Philip Treleaven, *Algorithmic Dispute Resolution: The Automation of Professional Dispute Resolution Using AI and Blockchain Technologies*, 61(3) COMPUTER J. 399 (2018); Jesse

increasingly assist judges in China,[189] with comparable systems being tested in the United States,[190] Europe,[191] and elsewhere, it seems unlikely in the short term that judges will be replaced by robots.[192]

C. EU Protections Against Automated Processing

The strongest protections against certain forms of algorithmic decision-making are found in Europe. As early as 1978, France adopted a law that prohibited administrative and private decisions based solely on automatic processing of data describing the "profile or personality" of an individual.[193] Though similar laws were adopted in Portugal[194] and Spain,[195] these remained outliers until the 1995 Data Protection Directive.[196] That required EU member states to grant individuals the right not to be subject to decisions based solely on automated processing of data evaluating them in areas such as "performance at work, creditworthiness, reliability, conduct, etc." Such processing was permissible only if it was part of a contractual relationship requested by the individual or if there were suitable measures to safeguard legitimate interests, such as arrangements allowing the individual "to put his [*sic*] point of view." An additional exception allowed for processing authorized by a law that also included measures to safeguard the individual's

---

Beatson, *AI-Supported Adjudicators: Should Artificial Intelligence Have a Role in Tribunal Adjudication?*, 31 CAN. J. OF ADMIN. L. & PRAC. 307 (2018); Ayelet Sela, *Can Computers Be Fair: How Automated and Human-Powered Online Dispute Resolution Affect Procedural Justice in Mediation and Arbitration*, 33(1) OHIO ST. J. ON DISP. RESOL. 91 (2018).

[189] Masha Borak, *China Embraces Tech in Its Courtrooms*, TECH NODE, Oct. 24, 2018; YU Meng and DU Guodong, *Why Are Chinese Courts Turning to AI?*, THE DIPLOMAT, Jan. 19, 2019.

[190] *See* Note, *State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing*, 130 HARV. L. REV. 1530 (2017).

[191] Nikolaos Aletras et al., *Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective*, 2:e93 PEERJ COMPUTER SCI. (2016)

[192] Frank Pasquale and Glyn Cashwell, *Prediction, Persuasion, and the Jurisprudence of Behaviourism*, 68(Supplement 1) U. TORONTO L.J. 63 (2018).

[193] Loi no 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés 1978 (France), art 2.

[194] Lei no 10/91, Lei da Protecção de Dados Pessoais face à Informática 1991 (Portugal), art 16.

[195] Ley Orgánica 5/1992, de 29 de octubre, de regulación del tratamiento automatizado de los datos de carácter personal 1992 (Spain).

[196] In 1981, a Council of Europe treaty was adopted on the topic, though it focused on privacy and data protection rights associated with automatic processing. Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108), done at Strasbourg, Jan. 29, 1981 (in force Oct. 1, 1985), ETS No 108, *at* http://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/108.

legitimate interests.[197]

The 2016 General Data Protection Regulation (GDPR) expanded both the possibilities for automated processing as well as the protections available. In addition to contractual arrangements, explicit consent can now be a basis for automated processing. Either basis, however, requires that safeguarding of interests goes beyond an opportunity to "put [one's] view" and includes the right to obtain "human intervention" to contest the decision.[198]

The question of whether the GDPR creates a "right to explanation"—meaning the ability to demand reasons as to how a particular decision was made—has been the subject of some debate, but is beyond the scope of the present article.[199] What is interesting in the present context is the rationale for prohibiting purely automated decision-making and the circumstances in which it can be allowed.

Early arguments put forward in the EU context focused on the need for individuals to be able to influence important decisions about themselves, as well as guarding against the abdication of human responsibilities to take those decisions in the face of a computer-approved outcome.[200] Safeguards against purely automated processing could have prohibited it entirely—requiring, for example, a "human-in-the-loop" approach that requires intervention prior to a decision being taken. That is unrealistic, as it would essentially render many widespread practices unlawful. In practice, it would also likely be ineffective, as routine human involvement to approve computer-prompted outcomes

---

[197] Directive 95/46/EC of the European Parliament and of the Council of Oct. 24, 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (EU Data Protection Directive) 1995 (EU), art 15.

[198] General Data Protection Regulation 2016/679 (GDPR) 2016 (EU), art 22. Recital 71 to the GDPR provides examples such as automatic refusal of an online credit application or e-recruiting practices without human intervention. Certain forms of sensitive personal data cannot be the basis for automated processing, unless it is necessary for reasons of "substantial public interest" or if the individual has given explicit consent. *Id*., art 9 (prohibiting the use of "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation").

[199] *See, e.g.,* Bryce Goodman and Seth Flaxman, *European Union Regulations on Algorithmic Decision Making and a "Right to Explanation,"* 38(3) A.I. MAG. 50 (2017); Gianclaudio Malgieri and Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7(4) INT'L DATA PRIVACY L. 243 (2017); Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7(2) INT'L DATA PRIVACY L. 76 (2017).

[200] Lee A. Bygrave, *Automated Profiling: Minding The Machine—Article 15 Of The EC Data Protection Directive And Automated Profiling*, 17(1) COMPUTER L. & SECURITY REP. 17, 18 (2001).

would quickly devolve into rubber-stamping or "quasi-automation."[201]

In general, for the purposes of private activities (based on contract or explicit consent) and public activities (based on legal authority), the requirement for "suitable measures" to protect the rights and interests of individuals makes it clear that automated processing can take place provided that there is a remedy if those rights or interests are violated—in particular, if a decision is based on impermissible forms of discrimination.[202] For decisions based on contract or consent, this is explicitly linked to the ability to challenge the decision and ensuring that such a challenge can be made to a human.[203]

Algorithmic decision-making thus poses an interesting counterpoint to the utilitarian approach to autonomous vehicles—where concerns are based largely on safety and ensuring accountability—and the deontic approach to autonomous weapons—where the concerns focus on the morality of allowing life and death decisions to be made at all. In the case of automated processing, decision-making by machine is tolerated provided that the legitimacy of such decisions can be ensured through the protection of rights and interests, in certain cases explicitly including the right to bring your concerns before another human being.

CONCLUSION: THE PROBLEM OF AUTONOMY

The rule of law is the epitome of anthropocentrism: humans are the primary subject and object of norms that are created, interpreted, and enforced by humans. Though legal constructs such as corporations may have rights and obligations, these are in turn traceable back to human agency in their acts of creation, even as their daily conduct is overseen to varying degrees by human agents.

True autonomy of AI systems challenges that paradigm. As we have seen, however, the challenge occurs in different ways. The emergence of autonomous vehicles is exposing gaps in the liability and criminal law regimes governing the roads, but these are ultimately practical problems to be addressed by amending those rules. The complicated nature of such amendments should not be underestimated, but the objective of managing risk is largely uncontroversial. Autonomous weapon systems, by contrast,

---

[201] Ben Wagner, *Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems*, 11(1) POL'Y & INTERNET 104 (2019).

[202] Yeung, *supra* note 158, at 515. *Cf.* Solon Barocas and Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016) (discussing profiling in the context of U.S. antidiscrimination law).

[203] *Cf.* Antoni Roig, *Safeguards for the Right not to Be Subject to a Decision Based Solely on Automated Processing (Article 22 GDPR)*, 8(3) EUR. J.L. & TECH., 6 (2017).

raise discrete moral questions—not *how* decisions by a machine should fit into our legal paradigms, but *whether* such decisions should be allowed in the first place. Algorithmic decision-making, at least for some decisions affecting the rights and obligations of individuals, runs the risk of treating human subjects as a means rather than an end. Unlike autonomous vehicles and weapons, the concern there is with the legitimacy of a decision made without human involvement.

These three types of concern—practicality, morality, legitimacy—are useful lenses through which to view the regulatory tools needed to address the larger challenges of AI, including those that are beyond our current horizon. Managing risk, preserving moral boundaries, and maintaining the legitimacy of public authority offer three strategies to help ensure that the benefits of AI do not come at unacceptable cost.

Yet, the nature of that cost is calculated differently in each case. Practical questions of minimizing harm reflect the utilitarian calculus of cost-benefit analysis. Moral questions of bright, non-negotiable lines suggest the duty-based ethics of deontology. The legitimacy of public authority, by contrast, points to issues of political theory. The aim here is not to reconcile these disparate conversations; rather, it is to highlight the complexity of the ostensibly simple notion of "autonomy."[204]

The history of the word itself embodies some of that complexity. Etymologically, "autonomy" comes from the Greek *autonomía*, combining *autos* (self) and *nomos* (law); its original use was confined almost exclusively to the political sphere, denoting civic communities with independent legislative authority.[205] It was only in the eighteenth century that Immanuel Kant applied the concept to humans, positing that morality requires a form of individual self-governance—that we ourselves legislate the moral law as rational beings.[206] Today, autonomy is also used in a looser sense of personal autonomy, meaning that a person acts in accordance with his or her own desires and values.[207]

None of these meanings corresponds fully to the AI systems discussed here. Though it is common for the "autonomy" of those systems to be described with reference to their ability to take decisions on their own, they do not have "desires" or "values" in any meaningful sense, nor are they

---

[204] *Cf.* Mike Ananny, *Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness*, 41(1) SCIENCE, TECH. & HUM. VALUES 93 (2016).

[205] John M. Cooper, *Stoic Autonomy*, *in* AUTONOMY 1 (Ellen Frankel Paul, Fred D. Miller, Jr.,, and Jeffrey Paul eds., 2010).

[206] JEROME B. SCHNEEWIND, THE INVENTION OF AUTONOMY: A HISTORY OF MODERN MORAL PHILOSOPHY 483 (1997).

[207] James Stacey Taylor, *Autonomy*, *in* ENCYCLOPEDIA OF MODERN POLITICAL THOUGHT 57 (Gregory Claeys ed., 2013).

"rational" in a way that Kant would have understood them to be.[208] On the contrary, what we typically mean when we describe an AI system as autonomous is not that it takes decisions "by itself," but that it takes decisions *without further input from a human*.

Understood in this way, the problem with autonomy is not some mysterious quality inherent in the AI system. Rather, it is a set of questions about whether, how, and with what safeguards human decision-making authority is being transferred to a machine. Algorithmic decision-making, for example, raises directly the question of the extent to which public authorities can outsource their responsibilities. Autonomous weapon systems have led many to argue that some decisions should not be outsourced at all. In the case of autonomous (*viz*. "driverless") vehicles, optimizing transportation does seem to be an area in which AI may be able to move people and goods more efficiently and—eventually—more safely than human drivers.

We are not there yet, of course. Almost a year after Elaine Herzberg died in Tempe, the Attorney for Yavapai County in Arizona, Sheila Polk, concluded that there was no basis for criminal liability on the part of Uber. She did, however, recommend that there should be further investigation of the backup driver, Ms. Vasquez, with a view to possible prosecution for manslaughter.[209] The Volvo XC90 itself, together with its onboard computer system, has been repaired and is, presumably, still on the road.[210]

---

[208] *Cf.* discussion of the anthropomorphic fallacy in in references cited *supra* note 30.

[209] David Meyer, *Uber Cleared Over Arizona Pedestrian's Self-Driving Car Death*, FORTUNE, Mar. 6, 2019.

[210] David Shepardson, *Uber Unveils Next-Generation Volvo Self-Driving Car*, REUTERS, June 12, 2019.