Journal of Business & Technology Law

Volume 15 | Issue 1 Article 4

Machine Performance and Human Failure: How Shall We Regulate Autonomous Machines?

Horst Eidenmüller

Machine Performance and Human Failure: How Shall We Regulate Autonomous Machines?

HORST EIDENMÜLLER*©

"But the mind that had once rebelled against the gods
was about to dethrone itself by way of its own fabulous reach.
In the compressed version, we would devise a machine a little
cleverer than ourselves, then set that machine to invent another
that lay beyond our comprehension. What need then of us?"[1]

ABSTRACT

Machines powered by artificial intelligence ("AI") are on the rise. In many use cases, their performance today already exceeds human capabilities. In this essay, I explore fundamental regulatory issues related to such "autonomous machines." In doing so, I adopt an analytical perspective that highlights the importance of what this article refers to as the "deep normative structure" of a particular society for crucial policy choices with respect to autonomous machines. This paper makes two principal claims. First, the jargon of welfare economics appears well-suited to analyze the chances and risks of innovative new technologies, and it is also reflected in legal doctrine on risk, responsibility and regulation. A pure welfarist conception of "the good" will tend to move a society into a direction in which autonomous systems will eventually take a prominent role. However, such a conception assumes more than the welfarist calculus can yield, and it also ignores the categorical difference between machine and human characteristic of Western legal systems. Second, taking the "deep normative structure" of Western legal systems seriously

* Statutory Professor of Commercial Law, University of Oxford; Member of the BBAW and ECGI Research Associate. I have benefited from helpful comments from participants following presentations of prior versions of this work at the AI for English Law Launch Conference (Oxford, March 18-19, 2019), the Third Annual Toronto-Oxford-UCLA Colloquium on Legal, Moral and Political Philosophy (Los Angeles, June 27-20, 2019) and on occasion of a special lecture delivered at the University of Würzburg ("Würzburger Vorträge zur Rechtsphilosophie, Rechtstheorie und Rechtssoziologie", Würzburg, July 11, 2019). Special thanks for detailed comments go to Calvin Normore, UCLA Department of Philosophy, and to Nikita Aggarwal, University of Oxford. I also thank Conor McLaughlin and Tilmann Frobenius for outstanding research assistance. The usual disclaimers apply.

1. IAN MCEWAN, MACHINES LIKE ME 80 (2019).

leads to policy conclusions regarding the regulation of autonomous machines that emphasize this categorical difference. Such a humanistic approach acknowledges human weaknesses and failures and protects humans. It is characterized by fundamental human rights and by the desire to achieve some level of distributive justice. Welfaristic pursuits are constrained by these humanistic features, and the severity of these constraints differs from jurisdiction to jurisdiction. The argument is illustrated with legal applications taken from various issues in the field of contract and tort.

INTRODUCTION

Machines powered by artificial intelligence ("AI") are on the rise.[2] In many use cases, their performance today already exceeds human capabilities. However, machines do not operate flawlessly—defects and accidents do occur. In 2018, an article on "Robotic Rules of the Road" in The Economist discussed the legal regime for self-driving cars.[3] In particular, it raised the question of when such cars would be allowed to participate in regular traffic. The author suggested that "… [autonomous vehicles] will always be held to higher safety standards than human drivers."[4] We learn that scholars and practitioners ponder whether they should be 10%, 90% or even 99.9% safer before being allowed to cruise the roads.[5]

But why should this be so? Why should self-driving cars not be allowed on our roads once they are as safe as human drivers or just *marginally* safer? Would this not be an improvement compared to the status quo? Indeed, taking the "safety logic" seriously appears to suggest an even more radical question: *When do we prohibit humans from driving cars?* Elon Musk pondered this question as early as 2015. He was reported to think "… that once self-driving cars become widely used, traditional human-driven vehicles may need to be banned. 'It's too dangerous. You can't have a person driving a two-tonne death machine,' said Musk during an appearance at Nvidia's annual developers conference, where he discussed Tesla's ambitions for autonomous-cars."[6] Applying the same logic, one may wonder whether superior machine performance should not have a feedback effect on the level of care required from humans during the intermediate stage when humans are still allowed to drive. This too has already been suggested in the literature: "…

2. The literature on the subject is of course vast. *See*, *e.g.*, ERIK BRYNJOLFSSON & ANDREW MCAFEE, THE SECOND MACHINE AGE: WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES (2014); MARTIN FORD, THE RISE OF THE ROBOTS: TECHNOLOGY AND THE THREAT OF MASS UNEMPLOYMENT (2015); THOMAS RID, RISE OF THE MACHINES: THE LOST HISTORY OF CYBERNETICS (2016).

3. *Robotic Rules of the Road*, THE ECONOMIST, May 12, 2018, at 70, 71.

4. *Id.* at 71.

5. *Id.*

6. Stuart Dredge, *Elon Musk: Self-Driving cars could lead to ban on Human Drivers*, THE GUARDIAN (Mar. 18, 2015, 3:22 PM), https://www.theguardian.com/technology/2015/mar/18/elon-musk-self-driving-cars-ban-human-drivers.

[O]nce it becomes practical to automate, and once doing so is safer, a computer should become the 'reasonable person' or standard of care."[7]

This essay explores fundamental regulatory issues related to "autonomous machines."[8] Regulation refers to regulation promulgated by states—i.e. not self-regulation by private actors—with machines and human behavior as objects of regulation as opposed to technology as a possible regulatory tool. "Autonomous machines" means artifacts which are ultimately designed and built by humans to perform specific functions without human intervention.[9] I adopt an *analytical* perspective that highlights the importance of the "deep normative structure" of a particular society for crucial policy choices with respect to autonomous machines.[10] In other words, I do not attempt to develop and defend new normative principles for the regulation of autonomous machines. Rather, I seek to proceed on the basis of and relative to a "deep normative structure" that I take as a given. By "deep normative structure" I mean the fundamental normative principles that are constitutive for the normative fabric of a particular jurisdiction at a particular point in time, guiding legal policy-making in that jurisdiction.

I make two principal claims. First, a naive form of utilitarianism may indeed take us down the policy road envisaged by Elon Musk and also reflected in the statement above on the influence of machine performance on the standard of care required from humans. It is a naive form of utilitarianism because it assumes more than the utilitarian calculus can yield. It also ignores the categorical difference between machines and humans characteristic of Western legal systems. Second, taking the "deep normative structure" of Western legal systems seriously leads to policy conclusions regarding the regulation of autonomous machines that emphasize this categorical difference. Most importantly, only humans enjoy (fundamental) *human* rights, and humans are treated according to their human—not super-human—faculties. Indeed, maintaining current levels of human activities and intercourse may even require us to relax behavioral standards in order to prevent a crowding out of the former by machine action.

Section I begins with a brief overview of what AI is, and what it can and cannot do, including in the realm of ethical decision-making ("machine ethics"). Against this background, the challenges of regulating autonomous machines and of

7. Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1, 36 (2018).

8. The essay builds on a previous blog post. *See* Horst Eidenmüller, *Machine Performance and Human Failure*, OXFORD BUS. L. BLOG (Nov. 2, 2018), https://www.law.ox.ac.uk/business-law-blog/blog/2018/11/law-and-autonomous-systems-series-machine-performance-and-human.

9. *See* NIKITA AGGARWAL, HORST EIDENMÜLLER ET AL., AUTONOMOUS SYSTEMS AND THE LAW 1 (2019). "Ultimate design" leaves open the possibility that an autonomous machine which is built by a human designs and builds another autonomous machine. "Without human intervention" can be further specified as implying that the artefact relies on its own perceptions rather than on the prior knowledge of its designers.

10. *See* Horst Eidenmüller, *The Rise of Robots and the Law of Humans*, 25 ZEITSCHRIFT FÜR EUROPÄISCHES PRIVATRECHT [ZEuP] 766, 774–77 (2017) (Ger.).

potential feedback effects on the standards governing human conduct become much clearer. Section II presents a utilitarian dystopia which might be the end-result of the naive utilitarianism inspiring much of the pro-machine policy rhetoric cited above. What makes this dystopia not an unrealistic prospect is the fact that utilitarian thinking in the form of simple welfare economics has a significant influence on legal policy-making, especially in the Anglo-American world. Relying on welfare economic concepts alone as guideposts for such policy-making might put us on a slippery slope which, at some point, could lead us to ignore the categorical difference between humans and machines. Section III contrasts this utilitarian dystopia with a richer and, I submit, more accurate account of the "deep normative structure" of Western societies. This account emphasizes the categorical difference between machines and humans, provides a convincing justification for human rights, acknowledges the failures of humans and protects them in situations of vulnerability, and it does not endeavor to make humans machine-like. I illustrate my argument with legal applications taken from various issues in the field of contract and tort law.

I. AI AND AUTONOMOUS MACHINES

AI aims at building artificial systems that function as well as, or better than, a human—in a domain requiring intelligence. The classic assessment of whether a system functions as well as a human is the "Turing test." In this test, a human is asked to engage in a conversation with messages sent through a mechanism that does not reveal whether the party on the other side is human or not.[11] If a human participant cannot distinguish the communications of an artificial system from a human, then the test is passed by that system. To pass a Turing test without any constraints around the type of conversation that could be had, the machine would need to exhibit "Artificial General Intelligence" ("AGI"); that is, intelligence that is as good as human in every dimension of intelligence.[12] Modern AI systems do not come anywhere near AGI. Rather, the AI deployed today only has (super)human-level capability in respect of narrowly-defined functions, such as image recognition, driving vehicles in straightforward surroundings, or the classification of documents.

Over the years, various methodological and technical approaches have been pursued within AI research.[13] The most recent development in AI has related primarily to *machine learning* ("ML"). This is an approach to computing in which the solution to an optimization problem is not coded in advance, but is derived inductively by reference to data.[14] In a sense, ML turns the logic of expert systems

11. Alan M. Turing, *Computing Machinery and Intelligence*, 49 MIND 433, 434 (1950).

12. *See* Richard Walters, *The Billion-Dollar Bet to Reach Human-Level AI*, FINANCIAL TIMES (Aug. 3, 2019), https://www.ft.com/content/c96e43be-b4df-11e9-8cb2-799a3a8cf37b (describing Artificial General Intelligence as "a level of cognition that would match its makers").

13. *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 16–28 (3rd ed. 2010).

14. For a comprehensive treatment see, for example, ETHEM ALPAYDIN, MACHINE LEARNING: THE NEW AI (2016).

upside down. Instead of deriving answers from rules and data, rules are developed from data and answers. The technique relies on applying computing power to very large amounts of data, the availability of which has blossomed in recent years.

Progress since 2012 has largely been in a particular type of ML known as deep learning, which involves running multiple layers of representation of the data in series.[15] A typical deep learning setup consists of an input and an output layer, with multiple hidden layers in between that lie at different levels of abstraction and are linked to each other ("artificial neural networks"). The various (hidden) layers represent different stages of abstraction of a thought process. For example, if the question is to identify at the output layer whether an image is a door, the first level would be pixels, the second would be edges, then corners, and so on up to the output layer: door, or not. The learning process of the algorithm takes place via so-called back-propagation: In the course of training the algorithm, new information is fed back from the output layer over the various hidden levels and recalibrates the settings or weights of the individual neurons with the aim of improving the accuracy of results.

The greatest practical successes with ML to date have been in the use of supervised learning techniques.[16] This refers to a process that begins with a dataset that is classified or labelled by humans according to the dimension of interest ("training data"). The system analyses this dataset and determines the best way to predict the relevant outcome variable (classified by the experts) by reference to the other available features of the data. The trained model—that is, the algorithm with the set of parameters that optimized performance on the training dataset—is then put to work on a new test dataset, to see how effective it is at predicting outside the original training sample.[17]

AI has also brought significant advances to the ability of machines to engage in ethical reasoning. Machines can be trained to act as if they have undertaken complex ethical considerations. In this sense, intelligent machines can be said to

15. See, for example, FRANÇOIS CHOLLET, DEEP LEARNING WITH PYTHON 8–11 (2018).

16. Approximately 95% of all ML applications today are based on this method. MARTIN FORD, ARCHITECTS OF INTELLIGENCE: THE TRUTH ABOUT AI FROM THE PEOPLE BUILDING IT 186 (2018).

17. Two other approaches to ML that are subject of much current research activity have yet to see the same level of practical application. These are unsupervised learning and reinforcement learning. *See* ALPAYDIN, *supra* note 14, at 111–23, 125–39. Unsupervised learning relies on the model itself to identify patterns in the data. In contrast to supervised learning, this does not necessitate a labelled training data set. However, the setup of the model's parameters becomes correspondingly more important in order to ensure that the resulting patterns are open to meaningful interpretation. In a reinforcement learning setup, the algorithm learns by trial and error. It receives a "reward" by finding the correct answer to a specified problem. A practical example is the training process relating to board games such as Go. AlphaGo by DeepMind has achieved superhuman performance within a very short time by playing very many games against itself. While reinforcement learning has much promise as a technique, its applications are currently limited to contexts in which the entire payoff structure can be specified with clarity. In most real-world settings, as opposed to game-playing, the complexity of the payoff structure rapidly exceeds the capability of hardware for the application of reinforcement learning techniques.

exhibit "moral agency", and "machine ethics" has emerged as a new discipline at the intersection of philosophy, computing and robotics.[18] The methods and technologies to teach machines to act ethically vary.[19] ML based on stories of real-world human behavior in situations involving, often complex, ethical decision-making is one such method.[20] By contrast to such a bottom-up approach, top-down approaches seek to formalize and automate ethical rules based on theories such as utilitarianism or Kantian principles.[21] Interestingly, this allows us to develop systems according to which machines apply ethical rules in a much more systematic and stable manner than humans. McEwan makes the point well when writing that "[h]umans were ethically flawed – inconsistent, emotionally labile, prone to biases, to errors in cognition, many of which were self-serving."[22]

AGI is not on the horizon anytime soon.[23] The near future will be characterized by (a) "autonomous AI" (autonomous systems/machines) for certain limited functions, and (b) by applications that support and improve human decision-making (assisted AI) or enable humans to do new things (augmented AI).[24] As a consequence, we will see a lot of collaborations between (smart) humans and machines—in all kinds of professional and private domains (semi-autonomous systems/machines). However, for the purposes of the following discussion, I will focus on fully autonomous systems/machines. This has the advantage of more clearly revealing the fundamental policy choices societies face with respect to regulating autonomous machines and the impact these choices have on humans and human behavior.

II. A WELFARIST DYSTOPIA

How should we regulate AI systems, especially autonomous AI systems? When should we, for example, allow self-driving cars to take part in regular traffic? What should be the liability regime if a self-driving car causes an accident? And should that liability regime influence the regime applicable to human drivers? Should fully autonomous cars have legal personality so that the car itself is liable?

These are important policy questions. It is clear that these questions cannot be answered without some normative conception of regulation/lawmaking. The

18. *See, e.g.*, CATRIN MISSELHORN, GRUNDFRAGEN DER MASCHINENETHIK (3rd ed. 2018) (Ger.).

19. *See, e.g.*, Simon Parkin, *Teaching Robots Right from Wrong*, THE ECONOMIST 1843, (June/July 2017), https://www.1843magazine.com/features/teaching-robots-right-from-wrong.

20. *Id.*

21. *See* MISSELHORN, *supra* note 18, at 97–114.

22. MCEWAN, *supra* note 1, at 86.

23. Naveen Josi *How Far are We from Achieving Artificial General Intelligence*, FORBES: COGNITIVE WORLD (June 10, 2019, 12:36 AM), https://www.forbes.com/sites/cognitiveworld/2019/06/10/how-far-are-we-from-achieving-artificial-general-intelligence/#12aba8af6dc4.

24. *See* THOMAS H. DAVENPORT, THE AI ADVANTAGE: HOW TO PUT THE ARTIFICIAL INTELLIGENCE REVOLUTION TO WORK 133–137, 190–192 (2018).

Following section demonstrates that much of the discourse on these issues is heavily influenced by a welfarist conception of "the good" which has its roots in utilitarianism. It has a significant traction particularly in Anglo-American legal systems, especially in so far as commercial activities in a broad sense are concerned. At the same time, it isolates and radicalizes one element of the "deep normative structure" of Western societies, and it may lead to policy conclusions which are squarely at odds with other elements of that structure.

A. Utilitarianism and Welfarism

As an ethical theory, utilitarianism still appears to hold some appeal, also in the context of AI regulation. In a well-known textbook on machine ethics, for example, utilitarianism is discussed as the first top-down approach for teaching autonomous machines to behave morally.[25] This may be because of the individualistic starting-point of utilitarianism as a behavioral theory—every individual attempts to maximize pleasure over pain—and the fact that programming machines to maximize utilities for humans appears to be an innocuous enough societal goal.

At the same time, the deficiencies of utilitarianism as a *regulatory* theory are plainly obvious:[26] Utilitarianism does not tell us how to measure pleasures and pains—the utilitarian effects of a contemplated regulation. It is not able to provide a solid foundation for fundamental human rights which we believe humans should enjoy,[27] and it does not care about how utilities are distributed, violating widely held views on the minimum content of a (distributively) just society. Finally, utilitarianism simply does not answer the question why it is a desirable thing to maximize utility in a given society. As a normative proposition this does not, of course, follow from the fact that individuals engage in such a maximization exercise (assuming that they do).

For these reasons, utilitarianism has never been a serious candidate to guide legal policy-making. However, a "modern" version of utilitarianism, namely the economic analysis of law, has become just that, and for a simple reason: By substituting the utilitarian calculus with a cost/benefit assessment of the real-life consequences of legal rules and regulations, it appears to solve at least the measurement problem that plagues utilitarianism. Some economists pay lip service

25. MISSELHORN, *supra* note 18, at 97–101.

26. *See, e.g.,* HORST EIDENMÜLLER, EFFIZIENZ ALS RECHTSPRINZIP: MÖGLICHKEITEN UND GRENZEN DER ÖKONOMISCHEN ANALYSE DES RECHTS [EFFICIENCY AS A LEGAL PRINCIPLE: POSSIBILITIES AND LIMITS OF THE ECONOMIC ANALYSIS OF LAW] 187–234 (4th ed. 2015).

27. *See* H.L.A. Hart, *The Shell Foundation Lectures, 1978-1979 Utilitarianism and Natural Rights*, 53 TUL. L. REV. 663, 670, 672–73 (1979) (explaining the inherent conflict between utilitarianism and individual rights). Rule utilitarianism does slightly better than act utilitarianism in this respect. *See* Mirko Bagaric, *A Utilitarian Argument: Laying the Foundation for a Coherent System of Law*, 10 OTAGO L. REV. 163, 167–68 n. 15 (2002) (contrasting act utilitarianism with rule utilitarianism). However, even under a system of rule utilitarianism, the status of human rights is precarious in the sense that such rights are always *contingent* on whether they increase societal utility (or not). *Id.*

to a broader welfare conception that goes beyond monetizable effects,[28] but when it comes to practical legal-economic analyses, immaterial benefits, "psychic costs" or "moralisms" are conveniently ignored.[29] Fundamental human rights do not appear to be a central concern at least if the scope of such analyses is restricted to commercial and corporate activities. Distributive goals are claimed to be more efficiently pursued by the tax and transfer system.[30] The question "Why efficiency?"[31] is answered with a simple "Because it is better for everybody to live in a richer society—at least in the long run—as benefits and costs will be randomly (evenly) distributed."[32]

B. Welfarism and AI

Many scholars and practitioners who engage in the discussion on policy-making regarding autonomous systems and machines do so with little appreciation for the need to specify and defend a regulatory goal and to demonstrate why and how that goal could be achieved by some proposed regulation. Some, even though highly critical of the prospect of "superintelligence" and the "existential risks" it poses, stop short of considering legal intervention at all and confine themselves to promoting "best practices" among AI researchers—a puzzling or even contradictory position.[33] Others proceed on the basis of vague statements such as the following: "We've developed laws to incentivize and facilitate *cooperation*, so if AI can *improve* our legal and governance systems, then it can enable us to cooperate more successfully than ever before, bringing out the very *best in us*."[34] Even those who ambitiously set out to provide "… a roadmap for a new set of regulations, asking not just what the rules should be but—more importantly—who should shape them and how can they be upheld"[35] end up with a hotchpot of pragmatic, utilitarian, economic, humanistic and ontological arguments that are presented in a random fashion to justify the result the author thinks is appropriate.

Hence, it is not surprising that lawmakers around the world struggle to identify some clear guideposts for AI regulation. The European Parliament is a good

28. STEVEN SHAVELL, FOUNDATIONS OF ECONOMIC ANALYSIS OF LAW 2 (2004).

29. Frank I. Michelman, *Norms and Normativity in the Economic Theory of Law*, 62 MINN. L REV. 1015, 1036 (1978); Duncan Kennedy, *Cost-Benefit Analysis of Entitlement Problems: A Critique*, 33 STAN. L REV. 387, 398 (1981).

30. Steven Shavell, *A Note on Efficiency vs. Distributional Equity in Legal Rulemaking: Should Distributional Equity Matter Given Optimal Income Taxation?*, 71 AM. ECON. REV. PAPERS AND PROC., 414, 414 (1981).

31. Ronald Dworkin, *Why Efficiency?*, 8 HOFSTRA L. REV. 563, 563 (1980).

32. *See, e.g.*, A. Mitchell Polinsky, *Probabilistic Compensation Criteria*, 86 Q.J. OF ECON. 407, 408-09 (1972).

33. NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 218–219 (2014).

34. MAX TEGMARK, LIFE 3.0: BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE 105 (2017).

35. JACOB TURNER, ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE 3 (2019). See also RESPONSIBLE AI—A GLOBAL POLICY FRAMEWORK (Charles Morgan ed. 2019), at 25 (suggesting "Ethical Purpose and Societal Benefit", "Accountability", "Transparency and Explainability", "Fairness and Non-Discrimination", "Safety and Reliability", "Open Data and Fair Competition", "Privacy", "AI and Intellectual Property").

example for this. In a resolution adopted on February 16, 2017, "with recommendations to the Commission on Civil Law Rules on Robotics",[36] the Parliament sets out "General Principles" for AI regulation from which *any* particular outcome could be deduced. It appears that the middle ground between computer scientists and philosophers on the one hand and law-makers on the other is not occupied. Coherent and convincing principles for AI-related lawmaking are needed, but it seems they are not yet available.

In a sense this is to be expected. AI is a relatively new field, certainly for lawyers. Prudent legal scholars and lawmakers would first try to understand the key issues, problems and risks before suggesting normative concepts or principles on how to deal with these. At the same time, jurisdictions worldwide have to make certain crucial policy decisions—such as, for example, whether to admit self-driving cars to practice—*now*; and for these decisions a principled approach is necessary or at least highly desirable. This is even more important as any adopted regulatory approach is likely to create path-dependencies, i.e. determine, to a certain extent, the steps that will follow in the future.

Those who do attempt to come up with coherent principles, or even a theory, for AI regulation often resort to welfare economics—either implicitly or more explicitly. For example, in a paper calling for a "global solution" on the issue of AI regulation the authors pose the following question: "Do these applications really make human society more *efficient, better, or safer*?"[37] "Efficient" has a precise meaning, "better" does not, and "safer" refers to risks or costs that influence the efficiency calculus. Another scholar argues that the liability regime he proposes for autonomous machines "… would benefit the general *welfare* . . ."[38] Even more clearly, another author writes this about the appropriate "framework of liability for autonomous systems:" lawmakers should devise rules "… with a view to maximize the net surplus for society by minimizing the costs associated with personal injury and property damage."[39] Finally, in a paper on "Regulating Artificial Intelligence Systems" we are reminded of the efficiency-enhancing path of the common law system and the need "… to find mechanisms for internalizing the costs associated with AI."[40] Many more statements such as these could be cited.

However, one should certainly not exaggerate. AI regulation is by no means just about welfare economics or efficiency, and there is an extensive literature by scholars concerned with other fundamental issues of AI regulation such as, for

36. Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics EUR. PARL. DOC. (INL 2015/2103) (2017).

37. Olivia J. Erdelyi & Judy Goldsmith, *Regulating Artificial Intelligence: Proposal for a Global Solution*, *in* 2018 AAAI/ACM CONF. ON AI, ETHICS, AND SOC'Y 1, 2 (2018) (emphasis added).

38. Abbott, *supra* note 7, at 5 (emphasis added).

39. GERHARD WAGNER, *Robot Liability*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3198764 (last visited on June 9, 2019).

40. Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 391 (2016).

example, privacy concerns and data protection or questions of "algorithmic discrimination."[41] At the same time, it can be anticipated that the rhetoric of welfare maximization is going to become more and more important, shaping the debate about AI regulation and achieving a crucial role for AI-related policy-making. A combination of factors and considerations leads to this assessment.

First, AI is about developing artificial systems that are at least as intelligent, rational, and productive as humans and, in many instances, even more so. Hence, AI appears to increase welfare and wealth of societies because we humans are, with the help of AI, able to accomplish tasks more rationally, systematically, and efficiently. There is a natural nexus between the nature of AI and the rhetoric of welfare economics.

Second, the debate about AI regulation is shaped to a significant degree by those who have an informational advantage with respect to the (ab)uses to which certain AI applications may be put and their real-world effects: the entrepreneurs who develop these applications in the first place.[42] Clearly these entrepreneurs are more than happy to use the rhetoric of welfare economics to help sell their products. Self-driving cars are safer, allow us to use driving time more productively, free up parking space, and mobilize the elderly.[43] Algorithmic credit scoring provides access to credit for many who were denied credit in the past[44] while smart medical applications diagnose illnesses faster and more precisely, treat patients better and, as a consequence, allow us to lead a longer and more fulfilling live.[45] Even where and to the extent negative effects of autonomous systems are discussed, the debate is framed in terms of risks, i.e. expected costs, and how these

41. For two seminal pieces of scholarship on these issues *see, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTIONS: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016).

42. *See also* Alan J. Dignam, Artificial Intelligence: The Very Human Dangers of Dysfunctional Design and Autocratic Corporate Governance, page 1, 30–32 (2019) (unpublished legal research paper) (on file with Queen Mary University of London School of Law).

43. *See All Tesla Cars Being Produced Now Have Full Self-Driving Hardware*, TESLA (Oct. 19, 2016), https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware?redirect=no ("Self-driving vehicles will play a crucial role in improving transportation *safety* and accelerating the world's transition to a *sustainable future.* Full autonomy will enable a Tesla to be *substantially safer* than a human driver, *lower the financial cost of transportation* for those who own a car and provide *low-cost on demand mobility* for those who do not.") (emphasis added).

44. *See* ZEST FINANCE, https://www.zestfinance.com/ (last visited June 6, 2019) ("ZAML® is the only way to get your ML credit models from the lab to product. *Reduce losses. Increase approvals. Fast*.") (emphasis added).

45. DEEPMIND, https://web.archive.org/web/20190606125832/https://deepmind.com/applied/deepmind-health/ (last visited June 6, 2019) ("Nurses and doctors in the NHS and elsewhere across the globe simply don't have the tools to instantly analyse each test result, determine the right treatment, and make sure that every single patient who needs complex or urgent care is escalated to the right specialist immediately. Many people think that new technology could help clinicians with *more accurate analyses*, and ultimately get *faster treatment* to the patients who need it most. We're committed to working with those on the frontline of healthcare to build technological solutions to these problems. We work with hospitals on mobile tools and AI research to help get patients from test to treatment *as quickly and accurately as possible*.") (emphasis added).

could be reduced. So, AI regulation is just about benefits and costs, it seems—familiar welfare economic terrain.

Third, one of the most heavily researched areas within the economic analysis of law is risk and the regulation of risk in contract and tort.[46] This research has led to a highly sophisticated conceptual apparatus that appears well-suited to answer key questions of AI regulation such as: When should we allow the use of autonomous systems? Who should be liable if something goes wrong, and under what standard? Should the liability regime applicable to autonomous machines influence the liability standard applicable to humans and, if so, how?

What is more, in most Western legal systems technological advances have an influence on the standard of care required from humans, for example with respect to medical diagnosis and treatment. In the United Kingdom, the so-called "*Bolam* test" is one of the foundational stones of the modern law of professional negligence. The test dates back to the *Bolam* case[47] in which the court held that "… a medical man [cannot] obstinately and pig-headedly carry on with some old technique if it has been proved to be contrary to what is really substantially the whole of informed medical opinion."[48] Hence, more efficient techniques must be used if there is a consensus on their superior properties by professionals working in the field.[49] This is "fertile ground" for the advent of smart AI-powered systems. Also, the classic Hand formula[50] for determining negligent behavior has made its way, often in a revised form that considers *marginal* costs/benefits instead of absolutes, into court decisions[51] and leading textbooks and commentaries on tort in many common and civil law jurisdictions.[52] Hence, the welfare economics-inspired doctrinal apparatus to deal with autonomous systems is in place.

C. Regulatory Consequences

On this basis, a series of mostly risk-related policy-recommendations for the regulation of autonomous systems can be deduced which will shape the debate about AI regulation in the years to come. Some of these recommendations are

46. *See, e.g.*, SHAVELL, *supra* note 28, at 175–385.

47. Bolam v. Friern Hospital Management Committee [1957] 1 WLR 582 (QB).

48. *Id.* at 587.

49. *See*, *e.g.*, Jason Millar & Ian Kerr, *Delegation, Relinquishment, and Responsibility: The Prospect of Expert Robots, in* ROBOT LAW 102, 117 (Ryan Calo et al. eds., 2016) ("Once there are expert robots, it will be easier to argue in some instances that they *ought* to be used to their full potential, because the evidence will suggest that in those instances they will, on average, deliver better results than human experts.") (emphasis added).

50. United States v. Carroll Towing Co., Inc., 159 F.2d 169, 173 (2d Cir. 1947).

51. *See, e.g.*, United States Fidelity & Guaranty Co. v. Plovidba, 683 F.2d 1022, 1027 (7th Cir. 1982); Bundesgerichtshof [BGH] [Federal Court of Justice] Nov. 29, 1983, NEUE JURISTISCHE WOCHENSCHRIFT [NJW] 1, 801–03, 1984 (Ger.)

52. *See* JOHN FREDERIC CLERK & WILLIAM HARRY BARBER LINDSELL, CLERK & LINDSELL ON TORTS 8–185 (Michael A. Jones et al. eds., 22nd ed. 2018); Gerhard Wagner, *MÜNCHENER KOMMENTAR ZUM BÜRGERLICHEN GESETZBUCH BAND 5*, *in* SCHULDRECHT BESONDERER TEIL III Vor § 823 ¶¶ 47–57 (Franz Jürgen Säcker et al. eds., 6th ed. 2013) (Ger.).

probably quite uncontroversial, others less so. Some may potentially bring about a fundamental change of our legal and political order. At the same time, it will be difficult to argue why this should not happen if one buys into the logic of welfare economics—it is a slippery slope that may well lead to unforeseen and "upsetting" consequences.

The following analysis will be illustrated mostly with references to self-driving cars for several reasons. Specifically, self-driving cars will become a reality affecting our lives fairly soon—they are not science fiction; transport by autonomous vehicles has a huge economic and political importance; and the regulatory challenges posed by self-driving cars are representative of the kinds of regulatory challenges by innovative and complex autonomous systems.[53]

Taking the issue of admitting self-driving cars to regular traffic first, the key welfare economic variable appears to be *safety*. People are concerned about potential damage caused by accidents, especially injury to humans, and so the scholars and practitioners cited in The Economist's article in the Introduction engage in discussions and speculations on the "safety margins" we should require from autonomous cars.[54] In the context of automobile traffic, safety can be measured by expected accident costs. If self-driving cars lower these costs, they are safer than human-driven cars. One reason why safety plays such a prominent role in the economic analysis of whether and when to allow self-driving cars on our roads is an "availability bias." Data on accidents involving cars is readily available,[55] so this is a hard factor in any argument about traffic regulation.

Against this background, the question on the admission of autonomous cars to practice appears to have a straightforward and simple answer. Autonomous cars should be allowed to operate if the expected costs from accidents are as low as or marginally lower than those from cars driven by humans. Based on a welfare economic analysis, there seems to be no reason to require autonomous cars to be safer than humans at all—at least if we restrict the analysis to the issue of safety. At the same time, *if* autonomous cars are marginally safer, such an analysis appears to lead to the conclusion that *only* autonomous cars should be allowed to operate and that humans should be prevented from driving. Total accident costs would, it can be assumed, be lower if this policy were implemented.

This surely would be a radical measure and departure from the status quo. At the same time, market forces might drive societies to the very same end result even

53. *See* Michael Wayland, *GM, Lyft Waymo Want to Be Allowed to Remove Driver Controls on Autonomous Cars*, CNBC (Aug. 30, 2019, 2:26 PM), https://www.cnbc.com/2019/08/30/gm-lyft-urge-regulators-to-remove-driver-controls-on-autonomous-cars.html (presenting a policy issue currently facing federal regulators in regard to self-driving cars).

54. *See* Brynjolfsson & McAfee, *supra* note 2.

55. *See*, *e.g.*, *Global Health Observatory (GHO) Data*, WORLD HEALTH ORGANIZATION, https://www.who.int/gho/road_safety/mortality/en/ (last visited June 7, 2019) (providing various reports by the World Health Organization (WHO) on road traffic deaths and road safety).

if humans are still able to drive alongside autonomous cars. This requires us to consider the liability regime for autonomous cars: Who should be liable if an autonomous car causes an accident, and should the liability regime be fault-based or strict? Economic analysis suggests that the appropriate liability standard with respect to autonomous cars should be strict, for two reasons. First, it is exceedingly difficult to precisely define the efficient level of care in this context.[56] Second, only a strict liability regime regulates the "activity level" of the car which influences the likelihood of accidents.[57]

The primary liability addressee of this regime should be the car manufacturer. The manufacturer is best positioned to control the risks and balance the benefits and costs of the technologies that are "driving" autonomous cars.[58] This is clearly so if the manufacturer develops the relevant AI applications. But even if the AI device producer is different from the car manufacturer, the car manufacturer controls the overall system, including all component parts. Hence, the car manufacturer probably is the "cheapest cost avoider."[59] As far as tort liability vis-à-vis third parties is concerned, it should therefore be the only liability addressee.[60] Interestingly, this "solution" to the liability problem seems to be the one toward which the market and private contracting practice are moving. Late in 2015, for example, Volvo announced that it would take responsibility for the actions of its self-driving cars.[61]

---

56. *See* THE ECONOMIST, *supra* note 3, at 70: "It will take years rather than months for the industry to cohere around a standard." On uncertainty in the finding of negligence see SHAVELL, *supra* note 28, at 224–28.

57. EIDENMÜLLER, *supra* note 10, at 771–73. A strict liability regime is also not necessarily bad for incentivizing innovation; indeed, studies have shown that more liability can *increase* investments in innovations that make products safer. *See* Alberto Galasso & Hong Luo, *Punishing Robots: Issues in the Economics of Tort Liability and Innovation in Artificial Intelligence*, *in* THE ECONOMICS OF ARTIFICIAL INTELLIGENCE: AN AGENDA 493, 493-99 (Ajay Agrawal et al. eds., 2019).

58. One problematic aspect of a strict tort liability regime that holds the car manufacturer liable for accidents caused is the activity level of *owners* (and operators) of autonomous cars. Just think of a taxi company on the one hand and a private car owner on the other hand—the activity levels and associated accident risks differ significantly. One can think of various potential "solutions" to this problem. One would be co-liability of owners depending on their activity profile. Another could be tying the sale of the car to liability insurance with the premium determined by (i) the manufacturer, (ii) the type of car, and (iii) the owner/user profile. Such personalized insurance is available already today. *See, e.g.*, BLACKBOX INSURANCE, http://www.blackboxinsurance.com/ (last visited June 9, 2019).

59. If parties were able to bargain for the applicable liability rule at zero (transaction) costs, they would contract for liability of the party that is best positioned to avert the expected accident costs at the lowest costs (Coase Theorem). *See* R. H. Coase, *The Problem of Social Cost*, 3 J.L. & ECON. 1, 27 (1960).

60. Of course, the car manufacturer could seek indemnity from the device producer based on their contractual relationship if a defective AI device ultimately caused an accident. Holding the car manufacturer strictly liable involves the risk of the manufacturer falling insolvent and therefore not being able to pay up. To cover this risk, manufacturers should be required by the law to purchase product liability insurance.

61. *See* Kirsten Korosec, *Volvo CEO: We will Accept all Liability when our Cars are in Autonomous Mode*, FORTUNE (Oct. 7, 2015), http://fortune.com/2015/10/07/volvo-liability-self-driving-cars/.

If car manufacturers were strictly liable for accidents involving a fully autonomous car produced by them, damage claims would still depend on whether claimants can prove that the car caused the accident. For causation to be established, some product defect and its impact on the chain of events would need to be identified and proven by the claimant.[62] In this, important but limited, sense, issues of fault would continue to play a role even under a strict liability regime.

Car manufacturers would insure against the additional liability risk, maybe because of a new legal requirement to do so,[63] and this would drive up the price for self-driving cars. Consequently, those who purchase and own self-driving cars will internalize the expected accident costs of their operation. This mirrors the current situation with human-driven cars. Hence, the owners and operators of cars, either autonomous or human-driven, always bear the expected accident costs of the respective car type. And *if* autonomous cars are safer than human-driven cars, they will be cheaper to operate, which is an incentive to substitute the latter with the former.[64] It therefore does not matter that much whether the law does or does not allow humans to drive cars once we admit autonomous cars to practice. Because autonomous cars will be cheaper to operate, market participants will have an economic incentive to use them—and to quit driving themselves.

This kind of market dynamic will play out in many settings in which AI-powered machines will come to be used. For another example, think of medical treatment and liability for malpractice. In most common and civil law jurisdictions, such liability is not governed by strict liability but by a negligence standard. If human doctors, autonomous machines, and human doctors assisted by smart machines are held to the exact same standard of care, superior machine performance will translate into an expected liability cost advantage and associated price advantage of machines or machine-assisted human doctors because these will be able to escape liability with greater certainty.[65] The effect will be even more pronounced if, following the *Bolam*-logic, human doctors are supposed to meet a higher standard of care which might be considered justified because of the superior performance of autonomous machines.[66]

---

62. This can be tricky. To address this problem, legal systems will consider reversing the burden of proof so that the car manufacturer would have to establish that there was no defect that could have caused the accident. *See, e.g.*, Wagner, *supra* note 39, at 13–14.

63. If there is no legal requirement to buy insurance from a third party, (large) car manufacturers might also "self-insure" as they also benefit from the law of large numbers. In terms of additional costs—and a higher price for purchasers—this does not make much difference.

64. Of course *expected liability costs* are only a fraction of *total operating costs*, and regular cars differ significantly from self-driving cars in many cost-relevant aspects. Self-driving cars will likely cost more than regular cars when first available for consumers because *scale* will be a key cost factor. Once these cars are produced and sold at significant scale, the situation might look quite different. In any event, *at the margin*, expected liability costs will certainly be a relevant factor affecting price.

65. On the effect of liability costs on total costs, see Coase, *supra* note 59.

66. Bolam v. Friern Hospital Management Committee [1957] 1 WLR 582 (QB).

The consequences could be dramatic and result in a rapid crowding out of human activity in our daily lives. Superior machine performance translates into a cost advantage for autonomous machines, and this translates into a competitive advantage for AI-powered goods and services. From a welfare economic perspective, this development is as inevitable as it is welcome. If the goal is to maximize net economic welfare in a society, then surely the law should set incentives to substitute less safe conduct and technology with safer conduct and technology. To put it differently, if machines operate more efficiently than humans, the liability system should also contribute to a process of substituting humans with machines.

Such a crowding out of human activity in our daily lives would not be confined to tasks humans generally find burdensome or displeasing. Rather, it would affect all domains of human life, including pursuits which humans very much enjoy. The ultimate result could be that also those activities which we view as constitutive for the "human condition," such as child-rearing or communicating with each other, are eliminated. Machines would be literally everywhere, and they would be better at what they do than any human could ever be.

Taking this analysis a step further might also yield an answer to the key policy question of granting fully autonomous machines legal personality and the power to hold property, conclude contracts, etc. This has been suggested by several scholars,[67] and it has also already become a point of consideration for lawmakers. In the European Parliament's resolution of February 16, 2017, "with recommendations to the Commission on Civil Law Rules on Robotics,"[68] the Parliament calls on the European Commission to consider "… creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently."

A welfare economic analysis of the problem would look to the consequences of such a policy and assess its effects, i.e. it would be based on a functional account of the contemplated rule change. If autonomous machines act like humans, exemplified by passing the Turing test, and if they perform better than humans in their respective domain, according them legal personality and giving them rights appears to be a justifiable policy move because it would reflect the fact that autonomous machines are functionally equivalent or even superior to humans. What is more, it would also allow these machines to participate in commercial transactions, creating more and more wealth for themselves and, if appropriately

67. *See, e.g.*, JERRY KAPLAN, HUMANS NEED NOT APPLY: A GUIDE TO WEALTH AND WORK IN THE AGE OF ARTIFICIAL INTELLIGENCE 79–92 (2015); Jens Kersten, *Menschen und Maschinen*, 70 JURISTENZEITUNG 1, 7 (2015).

68. *See supra* note 36.

taxed, also for humans or so it could seem.[69] Humans would come to be a new leisure class: jobless, taskless, and enjoying the benefits of machine labor—but to do what?

III. HUMANISM AND MACHINE REGULATION

To some, the welfarist regulatory approach to autonomous machines discussed in the previous Section might be an attractive prospect. To others it might reflect a dystopian future. But what exactly are the flaws and shortcomings of the argument presented in Section II? The welfarist regulatory approach to AI is seriously defective for three reasons. It assumes much more than the economic calculus can deliver; it turns a blind eye to widely held considerations of distributive justice, and it is not able to offer a convincing justification for fundamental *human* rights. This does not mean that welfare-economic analyses of AI-related regulatory problems are useless. It is the radicalization and de-contextualization of the welfarist viewpoint which creates serious problems. Central to the "deep normative structure" of Western societies is a *humanistic* approach that stresses the categorical difference between humans and machines. It protects humans where they suffer from human failures, and it seeks to avoid a crowding out of humans by machines.

A. A Critique of AI-related Welfarism

1. The Limits of the Welfarist Calculus

Analyzing the costs and benefits of regulatory measures targeting AI applications in monetary terms suggests that there exists a precise calculus that can guide policy-making on a welfare-economic basis. However, this is an illusion. The best that such an analysis can do is to highlight how certain measurable variables probably impact the efficiency effects of a proposed regulatory measure, and often even this modest goal will not be reached.

Take self-driving cars again as an example. First, even if one restricts the analysis to safety and expected accident costs, a meaningful analysis would have to factor in the likelihood and potential damage triggered by a global cyberattack on networked cars, for example. It is obvious that, similar to the global financial crisis, such an incident could inflict extraordinary monetary and personal damage on thousands or even millions of people, far exceeding any safety benefit that self-driving cars might otherwise have. But how should we assess the likelihood and

69. For this to happen we would also need to solve what is called the "goal alignment problem" in AI research: the challenge is to determine the objective function of an application and to ensure its correspondence with desirable human values. *See, e.g.*, BOSTROM, *supra* note 33, at 132–134. Further, the AI application would have to have the property of "corrigibility": it must be able to be switched off, and it must be able to be corrected if it is doing something humans don't like. *See, e.g.*, RUSSELL, *in*: FORD, *supra* note 16, at 67.

damage size of such an incident? We have no reliable data, and because we do not have data, the issue is ignored.

Second, the same will happen to other cost/benefit-relevant factors which we find difficult or even impossible to quantify. Just think of the increase in "quality" time that some will be enjoying when being driven in an autonomous car or, on the negative side, the loss in "sheer driving pleasure" others will bemoan if they can no longer drive themselves.[70] Even if they can, such pleasure might be negatively affected by the number of autonomous cars on the road—how much pleasure do we get from driving if surrounded by driverless cars?[71] And what kind of price tag do we put on the "quality time" if being driven or on the loss of "driving pleasure" when not driving? Asking humans how much they would request for giving up an entitlement to "quality time" or "driving pleasure" might yield dramatically different results to asking them what they are prepared to pay to have such an entitlement.[72] There is no way to assess the monetary value of these effects in an "objective" manner. Hence, it is absolutely unclear how the rising presence of autonomous cars will affect the overall net welfare in a society. The calculus is indeterminate with respect to the regulatory question of the safety requirements such cars have to fulfil before we allow them to operate. It is also indeterminate with respect to the question of when we should disallow humans from driving.

A similar indeterminacy will limit the usefulness for welfare economic considerations in all contexts of AI regulation in which many factors must go into the cost/benefit analysis, and we have hard monetary data on only a few. I fear that this will be the rule, not the exception.

2. Concerns about Distributive Justice

Welfare-economic analyses are blind to considerations of distributive justice. All that is important is aggregate welfare or maximizing the "size of the pie." This is justified by the claim that, in the long run, everybody can expect to benefit from a wealthier society. In this respect, economic welfarism is structurally identical to utilitarianism.

However, the argument that everybody can expect to be better off in the long run under a welfare economic policy is not convincing, for a variety of reasons.[73] First, those who have more now will likely end up having much more in the future—

70. *See* BMW, FACEBOOK https://www.facebook.com/BMW/photos/sheer-driving-pleasure-/10154742975092269/ (last visited June 8, 2019). Delivering "Sheer Driving Pleasure" happens to be the key marketing slogan of one of the world's premium automobile manufacturers. *Id.*

71. *See* BMW Welt I BMW Museum, *BMW WELCOMES. ARTIFICIAL INTELLIGENCE* at 1:58:40, YOUTUBE, (Apr. 21, 2016), https://www.youtube.com/watch?v=ELEaIdy7boU.

72. *See* Kennedy, *supra* note 29, *passim*; Mark Kelman, *Consumption Theory, Production Theory, and Ideology in the Coase Theorem*, 52 S. CAL. L. REV. 669, 669–79 (1979); EIDENMÜLLER, *supra* note 26, at 118–133. The reason for these discrepancies are income effects and endowment effects.

73. *See* EIDENMÜLLER, *supra* note 26, at 243–51, 281–83.

because of income effects in assessing cost/benefits of policy measures and because of their higher leverage to influence the political process. Second, those who argue in favor of the "compensation thesis" do not claim that everybody actually will be better off. The claim is that a random distribution of wins and losses under various policy measures will generate a positive *expected* value. However, even a positive expected value for everybody, which is unlikely, is perfectly compatible with significant *actual* losses for some. Third, the long run may be much too long for many. Issues of distributive justice of policy measures must be addressed convincingly *now*.

Crucial AI-related policy issues can be used to illustrate these points. As is well-known, Big Data and AI tools enable firms to personalize offers, i.e. to calibrate them to the individual, idiosyncratic preferences of specific consumers.[74] This relates to features of products and services, but it also relates to price. For a long time, first-degree price discrimination, i.e. setting prices based on the preferences and reservation values of *individual* consumers, existed only in textbooks. With Big Data and smart algorithms, it has become a reality.[75]

The total welfare effects of first-degree price discrimination are unclear. On the one hand, it gives certain individuals access to goods and services at lower prices, allowing them to buy goods and services that they were not able to buy before. On the other hand, privacy concerns may lead consumers to avoid price-discriminating firms, and rent-seeking investments by firms to capture as much value as possible. The defensive tactics employed by consumers to protect their privacy might lead to huge deadweight welfare losses.

The one thing that is clear about first degree price discrimination is that it massively redistributes transaction surplus from consumers to producers. If it works perfectly, consumers' surplus in fact shrinks to zero—consumers pay exactly their respective reservation price. At the same time, producers' surplus is maximized. Hence, it is not surprising that studies have found that consumers emphatically view individual price discrimination as unfair.[76] These fairness perceptions regarding the allocation of welfare/rents are clearly part of the "deep normative structure" of Western societies. Of course, different jurisdictions will draw the line between unacceptable and acceptable consequences, in regard to

74. Daniel Newman, *How Marketers are Using AI and Machine Learning to Grow*, FORBES (June 4, 2019, 11:20 AM), https://www.forbes.com/sites/danielnewman/2019/06/04/how-marketers-are-using-ai-and-machine-learning-to-grow-audiences/#6dc76bc21c0b.

75. For an analysis see OREN BAR-GILL, *Algorithmic Price Discrimination: When Demand is a Function of Both Preferences and (Mis)perceptions*, 86 CHI. L. REV. 217, 217–18 (2019); GERHARD WAGNER & HORST EIDENMÜLLER, *Down by Algorithms? Siphoning Rents, Exploiting Biases, and Shaping Preferences: Regulating the Dark Side of Personalized Transactions*, 86 CHI. L. REV. 581, 585–92 (2019).

76. *See, e.g.*, Timothy J. Richards et al., *Personalized Pricing and Price Fairness*, 44 INT'L J. OF INDUST. ORG. 138, 140 (2016); Kelly L. Haws & William O. Bearden, *Dynamic Pricing and Consumer Fairness Perceptions*, 33 J. OF CONSUMER RES. 304, 306–07 (2006).

distribution, differently. But they will not ignore distributive questions completely—as the welfare economic perspective does.

3. Concerns about Fundamental Human Rights

Finally, in a welfare-economic policy conception for regulating autonomous systems, fundamental human rights are always *contingent* on whether or not they contribute to maximizing social welfare. They have no inherent value independent of that goal. In this respect, welfare economics suffers from the same problems as utilitarianism. At the same time, the contingent status of fundamental human rights is clearly at odds with widely held beliefs regarding the cornerstones of a just and free society.

A proponent of a welfare-economic approach to regulating autonomous systems might argue that human rights concerns should not be raised in many instances of AI systems regulations, at least not if the scope of such regulations is restricted to commercial or corporate activities. This argument is clearly disingenuous. Just think of algorithmic credit scoring and the ever-present issues of direct or indirect discrimination involved in the decision on whether to accept a credit application or not. Similar concerns will be present in almost all cases in which large amounts of personal data and ML applications are relevant for autonomous systems, and that means practically always.[77]

Reconsidering the issue of admitting autonomous cars to our roads, for example, one might be concerned about the privacy implications of all the personal data that is needed to train the underlying ML models powering autonomous cars, plus the data that these cars will increasingly collect as part of the Internet of Things. One might also be concerned about the liberty of human drivers in the sense of autonomously choosing how to drive and interact with other drivers. Being constrained by the liberty of other humans is one thing—being constrained by machines' actions is quite another. Clearly these concerns about liberty become much stronger still if humans are banned from driving cars.

On the other hand, context matters, and other issues of regulating autonomous cars might well be less sensitive from a human rights point of view. The privacy and liberty concerns mentioned above relate primarily to the question of whether autonomous cars should be allowed to travel on roads at all. The applicable liability regime in case defects or accidents occur does not appear to raise such fundamental concerns. Another way of putting this is to say that, for example, a

77. Arianna Dorschel, *Rethinking Data Privacy: The Impact of Machine Learning*, MEDIUM (Apr. 24, 2019), https://medium.com/luminovo/data-privacy-in-machine-learning-a-technical-deep-dive-f7f0365b1d60. Hence, it is not necessary to resort to extreme—and rather futuristic—examples such as "autonomous child-rearers" to demonstrate the human rights concerns triggered by autonomous machines (in the case of "autonomous child-rearers" those of the parents and the children affected). John C. Havens, *Will We Lose Our Rights as Parents Once Robots Are Better at Raising Our Kids?*, QUARTZ, (Jul. 10, 2019), https://qz.com/1650396/tech-for-kids-will-soon-automate-away-the-job-of-parents/.

Kantian probably does not hold strong views as to whether this liability regime should be strict or fault-based. If anything, he or she might be more inclined to argue for a strict liability regime as it appears to be more protective of human safety. By comparison, the marginal reduction of choice opportunities for humans, strict liability might lead to fewer autonomous cars on the road compared to a negligence regime, appears to carry less weight.

B. A Humanistic Approach

The final section of this essay briefly outlines a humanistic approach to the regulation of autonomous systems that better reflects the "deep normative structure" of Western societies than the welfarist conception criticized above. It should be stressed that the goal in this Section is analytical, not normative. I am attempting to capture what, I believe, makes up the normative fabric of our societies, and I am neither offering nor defending a new conception. I am also not claiming that these fundamental normative principles cannot or do not change over time or that they are or should be universally accepted. Rather, I set out what, I believe, currently is constitutive for the normative fabric of Western societies. Central to the following observations is the view that humans are categorically different from machines and that this difference is central to our legal systems.

1. Humans and Machines

There is no denying that AI-powered machines are becoming smarter and smarter, outperforming humans in many domains. Machines will pass the Turing test for more and more applications.[78] At the same time, machines are able to engage in complex ethical considerations more systematically and precisely than humans, exhibiting a "moral agency." There is evidence that humans develop feelings towards humanoid robots and treat them like humans.[79]

All this is true. At the same time, it is missing the crucial point. We should not commit "The Android Fallacy."[80] *Machines and humans are categorically different, and everybody knows this*. This knowledge is not so much rooted in specific human behavior as it is derived from the collective human knowledge of the history of humankind and of human life—how it begins and how it ends. Humans create machines but not the other way around.[81] The categorical difference between

78. Turing *supra* note 11, at 434.

79. See, for example, Kate Darling, *Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behaviour Towards Robotic Objects*, *in*: RYAN CALO ET AL., *supra* note 49, at 213, 226–29; *see also* Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 545–549 (2015) (discussing "social valence").

80. Neil M. Richards & William D. Smart, *How Should the Law Think About Robots?*, *in*: RYAN CALO ET AL., *supra* note 49, at 3, 18–21.

81. Reproductive technologies have not yet advanced to the point were "human input" could be completely dispensed with, and it is highly doubtful whether that point will ever be reached.

humans and machines is also deeply ingrained in the "deep normative structure" of Western legal systems. Key aspects of this structure reflect this point.

2. Human Weaknesses and Failure

The first relates to how we treat humans and machines as legal actors. Machines may work perfectly or close to perfectly in certain settings and circumstances—humans do not. We are not machines and we make mistakes.[82] Being able to make mistakes, and to, hopefully, learn from them, is a manifestation of our freedom. Research in cognitive psychology has demonstrated that the mistakes we make are not random events. Rather, we suffer from systematic irrationalities, especially when we are "thinking fast" and follow heuristics.[83] Humans are often weak and wish they were not. We engage in self-paternalism to strengthen those preferences we wish we had in the long run—but do not have in a moment of vulnerability or temptation.[84]

The legal system takes humans as they are, with all their irrationalities, vulnerabilities and weaknesses. It reacts to these with an elaborate set of protective rules. Among these are disclosure duties by better informed parties, mandatory rules on substantive contract provisions, for example guarantees or liability rules, and rights to withdraw from certain contracts such as online sales or doorstep sales within a specified time-period ("withdrawal rights") to name just a few. The general rationale for these rules is to assist us not to engage in transactions that are potentially harmful to us. Sophisticated AI-powered tools used by businesses have made the problem more severe as businesses systematically use these tools to exploit behavioral anomalies.[85]

In tort law, humans are held to an objective standard of care.[86] But "objectivity" means no more than eliminating the personal idiosyncrasies of a particular person. The reasonable care required from a potential tortfeasor reflects what an "average" human could and would do under the circumstances. This standard punishes below-average laxity and sets incentives to do more, better. But the law does not require us to function like machines. It is true that the law may also require us to use new technologies under certain circumstances to escape liability. But this does

82. *See also* NICK BOSTROM, *in*: MARTIN FORD, *supra* note 16, at 100 ("Humans are a mess. We don't have a particular goal from which all the other objectives we pursue are sub-goals. We have different parts of our minds that are pulling in different directions, and if you increase our hormone levels, we suddenly change those values. Humans are not stable in the same way as machines, and maybe don't have a very clean, compact description as goal-maximizing agents.").

83. *See, e.g.,* DANIEL KAHNEMAN, THINKING, FAST AND SLOW 12–13 (2011).

84. *See* EIDENMÜLLER, *supra* note 26, at 374–85.

85. *See* WAGNER & EIDENMÜLLER, *supra* note 75, at 592–97; ILLAH REZA NOURBAKHSH, ROBOT FUTURES 14-15, (2013) ("[p]erfectly manufactured desire"); Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995, 1007 (2014) ("Mass Production of Bias").

86. *See* CLERK & LINDSELL, *supra* note 52, at 8–151.

not call the categorical difference between humans and machines into question. We do not treat humans and machines as if they were literally the same "thing."

This categorical difference is also reflected in the standards we apply to human behavior in situations involving moral dilemmas such as accidents that will inevitably occur but will have different negative consequences for different persons depending on the action, or inaction, of a moral agent. Because machines are capable of speedier, more systematic and consistent calculations than humans, we demand more from them than from ourselves. Humans will not be held morally or legally responsible for actions or inactions in a dilemmatic situation if they have to decide in a split-second what to do and if their behavior is not considered to be completely unreasonable. We acknowledge an "all too human" reaction in a situation involving complex and difficult moral decision-making.

3. Human Rights

The second feature of the "deep normative structure" of Western societies that reflects the categorical difference between humans and machines is human rights. Currently, only humans enjoy such rights. As has been mentioned before, human rights have a precarious status under social conceptions such as utilitarianism or welfarism which attempt to maximize a social welfare function. Indeed, one prominent philosophical justification for human rights conceives of them as "trumps" in the hands of individuals to protect themselves against majority rule.[87]

Within Western legal systems, human rights have a firm and crucial status, putting them above domestic parliamentary decisions in many jurisdictions. Just think of the "Charter of Fundamental Rights of the European Union"[88] or the core international human rights treaties negotiated and concluded under the auspices of the United Nations.[89] With respect to the regulation of autonomous systems in particular, the European Union's General Data Protection Regulation ("GDPR")[90] has become crucially important for data processing and privacy protection.

As a starting point under these and similar rules and regulations, fundamental human rights are enjoyed by humans, i.e. natural persons, and by humans only. The Charter of Fundamental Rights of the European Union, for example, starts out in Article 1 with stipulating that "[h]uman dignity is inviolable. It must be respected and protected." Similarly, Article 1(1) of the GDPR stipulates that "[t]his Regulation lays down rules relating to the protection of natural persons with regard to the

87. *See* RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 231–38, 272–78 (2nd ed. 1978).

88. Charter of Fundamental Rights of the European Union, 2012 O.J. (C 326) 2.

89. *The Core International Human Rights Treaties*, OFFICE OF THE UNITED NATIONS HIGH COMM'R FOR HUMAN RTS., https://www.ohchr.org/Documents/Publications/CoreTreatisen.pdf (last visited June 11, 2019).

90. Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC ("General Data Protection Regulation"), 2016 O.J. (L 119).

processing of personal data and rules relating to the free movement of personal data."

It is true that jurisdictions world-wide also extend "human rights" protection to legal persons such as corporations. Just think of the famous, and controversial, US Supreme Court decision in *Citizens United v. FEC*,[91] granting free speech protection to corporations. More generally, Article 19(3) of the German Constitution ("Basic Law")[92] stipulates that "[t]he basic rights shall also apply to domestic legal persons to the extent that the nature of such rights permits." However, we extend "human rights" to legal persons because our legal systems grant personhood to legal persons as vehicles for humans to exercise their fundamental human rights such as liberty and freedom of entrepreneurial activity. A corporation is a legal fiction run by humans for the benefit of humans. Even if and to the extent "self-driving corporations" become a reality,[93] it will ultimately still be humans that devise these vehicles and benefit from them. It would be an altogether different matter to grant fully autonomous machines legal personality and "human rights" without there being any humans involved as beneficiaries or shareholders etc.

It is to be expected that the crucial status of human rights within the normative fabric of Western societies will become apparent in the role such rights play in the public discourse about crucial AI-related policy-choices. The welfarist narrative of innovation, growth and risks captures only parts of what is at stake and, arguably, not the most important parts. Whether autonomous cars or human drivers will be seen on our roads, for example, is going to involve a delicate balancing exercise of different entrepreneurial and personal rights regarding business activities, private pursuits, bodily integrity, personal data etc. Different jurisdictions will strike the balance differently at different times. But surely arguments about rights and their relative weight will figure prominently in the decision-making process.

4. Distributive Justice

Finally, the "deep normative structure" of Western societies is characterized by some conception of achieving a distributively just outcome. Maximizing the size of the pie in terms of welfare is not enough. Even in capitalist societies that are closest to the welfarist manifesto, a lot of taxing and redistribution on social grounds is going on. Nozick's minimal state has been implemented nowhere in the real world.[94] As an antidote to the perceived huge problems of "global capitalism,"

91. Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 372 (2010).

92. BASIC LAW FOR THE FEDERAL REPUBLIC OF GERMANY [CONSTITUTION], art. 19.

93. John Armour & Horst Eidenmüller, *Self Driving Corporations?* (Eur. Corp. Governance Inst. L. Working Paper No. 475, 2019) (forthcoming in HARV. BUS. L.R.).

94. ROBERT NOZICK, ANARCHY, STATE AND UTOPIA 26–27 (1977).

societies currently appear to shift towards more redistribution, not less. Prime Minister May promised a "Britain that works for everyone"[95] in a post-Brexit UK.

Crucial AI-related policy-choices will be affected by concerns of distributive justice regarding how any efficiency gains associated with the new technologies should be distributed in a given society. The strong views of consumers with respect to first-degree price discrimination discussed above are indicative of what is to be expected if autonomous systems severely subvert labor markets. The debate about the potential introduction of a universal basic income may still seem to be somewhat premature right now.[96] However, the issue may become a fierce political battleground rather sooner than later.

CONCLUSION

If one wanted to sum up the features of the humanistic approach discussed above, one has a societal conception that stresses the categorical difference between humans and machines, acknowledges human weaknesses and failures and protects humans, and is characterized by fundamental human rights and by the desire to achieve some level of distributive justice. Welfaristic pursuits are constrained by these humanistic features, and the severity of these constraints differs from jurisdiction to jurisdiction. Finding the right balance between the competing concerns is a key challenge for deliberative democratic decision-making.

Against this background, there is nothing wrong or problematic about, for example, requiring autonomous cars to be much safer than human drivers before we allow them to participate in regular traffic, and there is nothing wrong about allowing humans to drive cars even though their driving skills might fall much short of the level achievable by smart cars. There is also nothing wrong about applying different standards of care to humans and smart machines. In fact, societies probably will, and should, consider relaxing the standards applicable to humans. Applying the same standards to humans and to autonomous machines translates into a cost and price advantage of the latter and might contribute to humans being shut out of more and more domains of our daily lives such as driving a car or just going out for a walk.

For this is the "slippery slope" of all societies which are built on foundations which reflect not only deep humanistic values but also a commitment to free markets as the main form of organizing economic activities. The jargon of welfare economics appears well-suited to analyze the chances and risks of innovative new technologies, and it is also reflected in legal doctrine on risk, responsibility and regulation. However, the welfarist narrative has an inbuilt tendency to go to extremes and shake off the humanistic constraints discussed above. What seems

95. *Speech at Conservative Party Conference in Birmingham*, BBC (Oct. 5, 2016), https://www.bbc.com/news/av/uk-politics-37563510/conservative-conference-theresa-may-s-speech-in-full.

96. *See*, *e.g.*, TEGMARK, *supra* note 34, at 126–28.

to be clear is that a pure welfarist conception of "the good" will tend to move a society into a direction in which autonomous systems eventually will take a prominent role—by virtue of the law.

Hence, regulating autonomous systems is a challenge that requires us to take the "deep normative structure" of our societies seriously. Our laws are an expression of the human condition. They reflect what we believe lies at the heart of humanity, at the heart of what it means to be human. It simply and literally would be the dehumanizing of the world if we were to treat machines like humans, even though machines may be smart—possibly even much smarter than humans.[97]

97. *See* generally CURTIS WHITE, WE, ROBOTS: STAYING HUMAN IN THE AGE OF BIG DATA (2015) for a discussion on the threats for the human condition by Big Data. On how to counter these threats *see* BRETT FRISCHMANN & EVAN SELINGER, RE-ENGINEERING HUMANITY (2019).