



Jeanne Gaakeer*

Minding the Shift: Some Thoughts on the Human and the Not Too Human Mind

<https://doi.org/10.1515/pol-2023-2003>

Abstract: Software systems that claim to think for themselves may prove to be Frankenstein’s nightmare come true. When Amazon used algorithms to read digital letters of job applicants, it turned out that mentioning women sports in the category of one’s favourite leisure activities was detrimental to getting the job, i.e., the algorithm favoured male applicants. What if AI becomes able to and will in fact take over cars, elevators and other electronic devices with devastating results for their human owners or operators? In this article, I aim to address the differences between, and the possible clashes of the human and the AI minds (and their souls and hearts) by means of some examples, including a.o., drones, ADS (automatic driving systems) and our recent Artificial Friend, Kazuo Ishiguro’s fictional robot Klara, and I do so by focussing on the narratives involved, on the view that new technologies need a legal hermeneut. This article is inspired, firstly, by the novelist Samuel Butler and the today not so fictional question that he raised, i.e., whether “because mechanical life is a very different thing from ours, therefore it is not life at all?” (Samuel Butler, *Erewhon* [1872]. Harmondsworth UK: Penguin, 1954, 175), because Butler adds that if it is considered life it could have a will of its own as well and that may lead to dire consequences also from a point of view of law; secondly, because the topic of AI also invites us to consider the interdisciplinary question why new technologies need the humanities. As the philosopher Edmund Husserl claimed, a determination of the world by the positive sciences as took place in the second half of the nineteenth century occasioned “an indifferent turning-away from the questions which are decisive for a genuine humanity. Merely fact-minded sciences make merely fact-minded people” (Edmund Husserl, *The Crisis of European Sciences and Transcendental Phenomenology, An Introduction to Phenomenological Philosophy* [1954], trans. David Carr. Evanston, Ill.: Northwestern University Press, 1970, 6).

Keywords: A.I., “black box” mentality, automatic driving systems, automatic weapon systems, criminal liability, human and A.I. agency, remote control, Ishiguro, Ellul

*Corresponding author: Jeanne Gaakeer, Erasmus University Rotterdam, Rotterdam, Netherlands, E-mail: gaakeer@law.eur.nl

Open Access. © 2023 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

“Philosophy is not to relieve one of decisions, but to confront [...] how superfluous any existence if ultimately the world involved no contradiction and life involved no decision!”¹

1 Posthuman Technologies

After the turn to interpretation in law and the social sciences and the turn to literature and narrative in law we have increasingly been forced to face the computational turn in law, also known as artificial legal intelligence or data-driven law.² Our posthuman lives of and with modern technologies are filled to the brim with algorithms, i.e., “any instruction, such as a computer code, that carries out a set of commands,” or “encoded procedures for transforming input data into the desired output, based on specific calculations.”³ With algorithms in legal surroundings come huge challenges varying from privacy issues and other human rights intrusions, to questions of due process and good governance. The year 2020 introduced novel forms of government surveillance in the form of cyber-tracking during Covid-19 of the whereabouts of people via telecom data and GPS systems and smartphones in order to check whether they stuck to their quarantine or to order them to go into quarantine, for example in Hong Kong, South Korea, and Israel.⁴ That should make us wonder whether monitoring has not become self-monitoring as well. The introduction of other apps to trace both Covid-19 virus-infected humans and those recovered from the virus in their spatial connections is a case in point, as is the Covid-19 check app necessary to get access to restaurants, museums and other public spaces throughout Europe.⁵ Elsewhere I have argued that one risk of modern technologies is that if it can be done, somewhere, someone, at some point in time will actually do it,

1 Gustav Radbruch, *The Legal Philosophies of Lask, Radbruch, and Dabin*, the 20th century legal philosophy series vol. IV, trans. Kurt Wilk and intro. Edwin W. Patterson (Cambridge (Mass.): Harvard University Press, 1950), 47–224, 112.

2 Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (Cheltenham: Edward Elgar Publishing, 2015).

3 See Article 19 Group, “Privacy and Freedom of Expression in the Age of Artificial Intelligence,” April 2018, accessed October 16, 2021, <https://privacyinternational.org/sites/default/files/2018-04/Privacy%20and%20Freedom%20of%20Expression%20in%20the%20Age%20of%20Artificial%20Intelligence.pdf>.

4 See Edan Ring, “On Pegasus,” *London Review of Books* (4 November, 2021): 32–33, for an account of how Israel infiltrated the mobile telephones of its citizens via Pegasus software with “zero click” capabilities.

5 Cf., Claire Wrobel, “Negotiating Dataveillance in the Near Future: Margaret Atwood’s Dystopias,” *Commonwealth Essays and Studies* [Online] 43.2, last modified 23 July 2021, accessed November 4, 2021, <http://journals.openedition.org/ces/7718>.

and that these technologies can lead to inequalities in daily life and at a political level, to undesired forms of “social sorting”.⁶

What is more, when software systems start to think for themselves they may prove to be Frankenstein’s nightmare come true. When Amazon used algorithms to read digital letters of job applicants, it turned out that mentioning women sports in the category of one’s favourite leisure activities was detrimental to getting the job, i.e., the algorithm favoured male applicants. Obviously, this bias was built-in. Computers governing credit card transactions do so on the basis of AI agents and they operate “in a designed environment that allows them to make their decisions.”⁷ From human resource management to border control, from weather forecasts to penal law, e.g., COMPAS used in the U.S. American criminal law system and the OxRec algorithm in the Netherlands used to predict recidivism, machine learning (ML) as a form of AI, more specifically supervised ML, reigns.⁸ We have come a long way since the 2001 issue addressed by Cass Sunstein and others, “How Computers Think Like Lawyers” – note the semantics: *how*, not *do* – because to the average jurist and lay person alike, the complexity and with it the opacity of AI systems including the opacity of the goals and aims of both engineers and entrepreneurs, has increased exponentially, as has our dependence upon them.⁹ Thus computational law “forms a potential challenge of a distinctively hermeneutical kind to the protections afforded by the rule of law,” as Emilie van den Hoven points out, referring to Miranda Fricker’s work on discriminatory epistemic injustice, more specifically hermeneutical injustice that “affects the individual’s capacity for social understanding.”¹⁰ In the discursive situation of court surroundings, this demands from the judge not only technological expertise but also the practical wisdom to recognize this phenomenon

6 See Jeanne Gaakeer, “‘Select before you collect’: uses and abuses of profiling and data mining in law and literature,” *Pólemos* 14.1 (2020): 57–71.

7 Colin Allen, “Artificial life, artificial agents, virtual realities: technologies of autonomous agency,” in *The Cambridge Handbook of Information and Computer Ethics*, ed. Luciano Floridi (Cambridge: Cambridge University Press, 2010): 219–233, 222.

8 Machine learning refers to the ability of computer algorithms to “learn” and improve their performance of some task over time. The learning concerns learning from the patterns in the data that have been gathered. See Access Now and Lindsay Andersen, “Human Rights in the Age of Artificial Intelligence,” accessed October 1, 2021, [AI-and-Human-Rights.pdf](#) (accessnow.org); John Zerilli with others, *A Citizen’s Guide to Artificial Intelligence* (Cambridge [Mass.]: MIT Press, 2021), 2.

9 Cass Sunstein et al., “Symposium: Legal Reasoning and Artificial Intelligence: How Computers Think Like Lawyers,” *University of Chicago Law School Roundtable* 8.1 (2001): 1–17. For a more recent discussion of what “robojudges” may achieve, see Kiel Brennan-Marquez and Stephanie E. Henderson, “Artificial Intelligence and Role-Reversible Judgment,” *The Journal of Criminal Law & Criminology* 109 (2019): 137–164.

10 Emilie van den Hoven, “Hermeneutical injustice and the computational turn in law,” *Journal of Cross-Disciplinary Research in Computational Law, CRCL online-first* (March 2021):1–13, 2 and 3.

because it puts someone at an unfair advantage.¹¹ Furthermore, this not only suggests the question whether or not our humanity will be able to set limits to artificial intelligence,¹² but it also shows that hermeneutical capabilities as taught in the humanities are much needed.

However useful these algorithms may seem, they or rather their ways of reasoning are hard to understand for humans. This pertains more specifically those of deep learning and DNNs, i.e., deep neural networks, that “simulate human ways of thinking and rules for acting. Here, the software is capable of enhancing digital programming created by humans, and thus of evolving independently of such programming.”¹³ Such learning, however, is not equivalent to what in humans is called thinking,¹⁴ even though the human brain uses a comparable computational technique and it, too, “is a collection of neurons, linked together by synapses.”¹⁵ What is more, like juries in court, DNNs are black boxes that give no reason or justification for what they do and decide. As Frank Pasquale notes,

There is a strange vogue among defenders of ‘black box’ artificial intelligence (AI) systems to say that they are no worse than human beings at making decisions, because human reasoning is merely a veil thrown over actions taken on inscrutable grounds. Though framed as an open-minded epistemological pluralism, meant to welcome robots and AI within the fold of ‘thinkers’ and intentional agents, this diminution of human thought is dangerous. It paves the way to a world in which computational systems can simply enforce law’s commands, without understanding what the enforcement means to the person who experiences the force of law.¹⁶

Such critique points to the need for transparency of AI systems, on the one hand, and, on the other hand, it urges us not to consider deep learning as a form of intelligence yet. Here too the reflective, critical attitude of the humanities can help form a counterweight.

¹¹ For an analysis of Fricker in court surroundings, see Jeanne Gaakeer, *Judging from Experience. Law, Praxis, Humanities* (Edinburgh: Edinburgh University Press, 2019), 154–155.

¹² Tania Sourdin and Richard Cornes, “Do Judges Need to Be Human? The Implications of Technology for Responsive Judging,” in *The Responsive Judge*, eds Tania Sourdin and Archie Zariski (Singapore: Springer, 2018): 87–119, 103.

¹³ Wolfgang Hoffmann-Riem, “Legal Technology/Computational Law, preconditions, opportunities and risks,” *Journal of Cross-Disciplinary Research in Computational Law, CRCL online-first* (November 2020), 1–13, 3.

¹⁴ Cf., Zerilli, *Citizen’s Guide*, 55, on the artificially intelligent chatbot Tay, a learning algorithm launched by Microsoft in 2016 and designed to talk like a nineteen-year-old American girl. Tay’s career was short-lived because within 24 h his language became so offensive that Tay was shut down. Microsoft had overseen the possibility of an attack by trolls who had indeed fed Tay with abusive language and information: “In short, Tay was a learner but not a *thinker*” [italics in the original].

¹⁵ Zerilli, *Citizen’s Guide*, 12.

¹⁶ Frank Pasquale, “The Substance of Poetic Procedure: Law & Humanity in the work of Lawrence Joseph,” *Law & Literature* 32.1 (2021): 1–46, 32 and 33, endnote omitted.

Now one could ask: Do we need to understand anything or anyone before we can trust them? Not necessarily so. *Reliabilism* as an epistemology obviously has very good cards. In daily life we trust many other things we don't understand, for example the diagnoses made by our doctors. Generally speaking, if we deem a decision of a doctor, or a judge for that matter, reliable, it is because we assume, consciously or not, that the procedure to reach that decision is trustworthy and that the agent is competent. The same goes for ML. Yet, at the same time we suffer from anthropomorphism when it comes to AI. On 23 April 2014 I watched a CNN broadcast in which the reporter spoke on the subject of the Malaysia Airlines plane then gone missing since early March 2014, to a member of an Australian team specialised in the delicate task of extracting data from flight recorders, the interior of which was referred to as "the brain within." Familiar appliances such as Siri and Alexa, not to mention the human look-alike robot Sophia that was granted Saudi Arabian citizenship¹⁷ are other obvious anthropomorphic examples, as is Helen in Richard Powers 1995 novel *Galatea 2.2*.¹⁸

So one question would be: What if AI becomes able to and will in fact take over cars, elevators and other electronic devices with devastating results for their human owners or operators? Such a frightening view of the future in which AI "takes over" is found in the 2020 series *NeXT*, a creation of Manny Coto.¹⁹ More down to earth is this legal example: when a proposal for new Dutch legislation with respect to more far-reaching powers for the police to hack a suspect was discussed, one high police officer suggested these powers should include hacking a suspect's car to make it stop to ensure easy arrest. The Secretary of State then quickly had to reassure both Chambers that such goal was not politically intended.²⁰ With respect to autonomous vehicles we have yet to consider the real and present dangers of cybercrime as a safety risk. A team of Chinese hackers, for example, was able to take remote control of a Tesla Model S, move the seats back and forth, and open the boot while the car was driving.²¹ A (jealous?) man wanting to follow his girlfriend's movements was able to

17 Jeanne Gaakeer, "Select before you collect": uses and abuses of profiling and data mining in law and literature," *Pólemos* 14.1 (2020): 57–71, 66, note 20.

18 For an analysis of the novel, see Mireille Hildebrandt, "From *Galatea 2.2* to Watson – And Back?," in *Human Law and Computer Law: Comparative Perspectives*, eds Mireille Hildebrandt and Jeanne Gaakeer (Dordrecht: Springer, 2013): 23–45.

19 For an early discussion of machines that carry out crimes, see Carson Reynolds and Masatoshi Ishikawa, "Robotic Thugs," in *2007 ethicomp proceedings* (Tokyo: Global e-SCM Research Center & Meiji University): 487–492.

20 The Dutch *Wet Computercriminaliteit III*, implemented 1 March 2019.

21 See <https://www.theguardian.com/technology/2016/sep/20/tesla-model-s-chinese-hack-remote-control-brakes>, accessed December 13, 2020.

install an app so that the position of her car was sent to his email address.²² The examples point to another legal problem, i.e., that of “*how* personal data are processed, in what context and towards what end.”²³ Thus we must face the future nightmare of totalitarian states connecting their citizen’s automated devices to systems of deterrence and surveillance.²⁴ This is also to say that while new technology-induced perspectives on human and non-human agency and responsibility matter even more in legal environments, they are not as easily categorised as before. So another, equally crucial question would be whether algorithms are “carefully calibrated to their specific role?”²⁵ Both questions suggest the need to reflect on new forms of responsibility and accountability in legal surroundings.²⁶

In this article, I highlight a few differences between, and the possible clashes of the human and the AI mind, assuming for the purpose that both have minds although not always of their own.²⁷ I do so by means of three examples of AI: ADS (automatic driving systems), drones and Automatic Weapon Systems (AWS) and our new Artificial Friend (AF), Ishiguro’s fictional robot Klara. My argument is inspired by the view that new technologies need a legal hermeneut, as Bruno Latour noted when he claimed that the despite the process of postmodernization the humanistic elements such as “the hermeneut” remain indispensable as the objects of reflection and research, and that this is even more so because they have been abandoned or even

22 See <https://www.whichcar.com.au/car-news/man-found-guilty-stalking-with-car-app>, accessed December 20, 2020.

23 Simone van der Hof and Corien Prins, “Personalisation and its Influence on Identities, Behaviour and Social Values,” in *Profiling the European Citizen, Cross-Disciplinary Perspectives*, eds Mireille Hildebrandt and Serge Gutwirth (Dordrecht: Springer, 2008): 111–124, 118. See also “The key principles of vehicle cyber security for connected and automated vehicles,” 6 August 2017, <https://www.gov.uk/>, accessed August 1, 2019.

24 See also Gaakeer, *Judging from Experience*, ch. 13, for a discussion of Juli Zeh’s dystopian novel *The Method* (orig. *Corpus Delicti*). See also Roger Brownsword and Alon Harel, “Law, liberty and technology: criminal justice in the context of smart machines,” *International Journal of Law in Context* 15 (2019): 107–125.

25 Günther Teubner, “Digital Personhood? The Status of Autonomous Software Agents in Private Law,” *Ancilla Iuris* (2018): 35–78, 36.

26 See also Sabine Gless, “AI in the Courtroom: A Comparative Analysis of Machine Evidence in Criminal Trials,” *Georgetown Journal of International Law* 51 (2020): 195–253; Bart Custers and Lonneke Stevens, “The Use of Data as Evidence in Dutch Criminal Courts,” *European Journal of Crime, Criminal Law and Criminal Justice* 29 (2021): 25–46.

27 Zerilli, *Citizen’s Guide*, 1, “A famous definition of AI states that AI is the science of making computers produce behaviours that would be considered intelligent *if done by humans*” [italics in the original], and, 72, “Systems that automate well-defined processes are relatively predictable, whereas systems that learn from their experiences and exhibit behaviors going beyond their original programming can *seem* like they have a ‘mind of their own’” [my italics].

orphaned by those studying scientific-technological developments.²⁸ The hermeneut is necessary because it is one thing to understand how forms of AI work but quite another to evaluate what they do and, having done so, reflect on whether what they do is morally and legally acceptable. Practically speaking, in the end it is a matter of “responsibly assign(ing) specific kinds of work to appropriate configurations of people, machines, and processes.”²⁹ To be able to do so ethically, we need to probe our hermeneutic relation to the world by means of the artifact and ask what it is a representation of.³⁰ With Heidegger in “The Question Concerning Technology,” we must ask how things technological are revealed and emerge from their concealment so that we can understand their cause and effect. Heidegger claims that technology is a way of revealing in the sense of *poiesis*, a bringing-forth, a bringing into appearance in the world, literally as much as artistically or poetically. This means that technology is “no mere means”, and, “Since man drives technology forward, he takes part in ordering as a way of revealing.”³¹ While in referring to Heidegger I do not wish to promote the “tragic strand” in juridical-technological thought that Kieran Tranter claims originates in Heidegger,³² but, following Verbeek, I would insist on viewing technology as an activity rather than as something out there that is merely a means to an end, so that we also need to address the morality of things.³³ Humans remain responsible for their artifacts. Insofar as we are still grappling for hermeneutic tools to understand new technologies, here too the humanities offer both theory and practice to preclude universal, utilitarian thought.

²⁸ Bruno Latour, *We Have Never Been Modern* (New York and London: Harvester Wheatsheaf, 1993), 136.

²⁹ Brian Caldwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge [Mass.]: MIT Press, 2019), 5.

³⁰ Peter-Paul Verbeek, *What Things Do: Philosophical Reflections on Technology, Agency, and Design* (University Park, Pennsylvania: Pennsylvania State University Press, 2005), 126.

³¹ Martin Heidegger, “The question concerning technology,” [1954] in *The Question Concerning Technology and Other Essays*, ed. W. Lovitt (New York: Harper & Row, 1977), 3–35, 12 and 18. Cf. Verbeek, *What Things Do*, 56.

³² Kieran Tranter, *Living in Technical Legality. Science Fiction and Law as Technology* (Edinburgh: Edinburgh University Press, 2018), 2, “Within thinking about technology and society, there is a tragic strand that regards modern technological existence as the end of true human life, its vitality replaced by empty exchange. The theoretical foundation for this literature is Martin Heidegger’s account of technology.”

³³ Cf., Verbeek, *What Things Do*, 57, on Heidegger. Cf., Rosi Braidotti, *The Posthuman* (Cambridge: Polity, 2013), 41, that Verbeek’s view is a form of analytic post-humanism, i.e., “qualified by a profoundly humanist and thus normative approach to technology itself.”

2 The Deficit of Lady Lovelace's Objection

2.1 Let Me Drive Your Car? AI Technology and ADS

One good reason to consider AI in relation to law is that the pace with which the technologies both in the real and the digital, virtual world develop is in sharp contrast with the infrastructural developments in the world of tangible things. An illustrative example from the Netherlands is that, in 1970 – when GPS was non-existent – the Department of Waterways and Public Works announced the construction of a new part of a motorway, the A27/A12, one of the most complex and most congested motorways in the country. From day one, the plan met with fierce resistance from environmentalists. They went to court. Long story short, in 2017 the Council of State posed prejudicial questions to the European Court of Justice in Luxemburg and so far the legal issues haven't been resolved and the necessary road reconstructions – by now partly finished and partly amended because of increasing road traffic during this prolonged decision-making process – are still underway. The protests were still going strong end of 2020.³⁴ Meanwhile the US have launched military satellites with GPS, laptops, smartphones and Teslas have been developed, and Google and Amazon have become global rather than local players.

When we consider the development of ADS, civil, criminal and administrative law lags behind in numerous ways.³⁵ As Jacques Ellul had noted already in 1954, the continuity of the technological process on a world-wide scale has huge consequences: “The first is that the problems created by mechanical technique will be heightened to a degree as yet incalculable, as a result of the application of technique to administration and to all spheres of life.”³⁶ Since law has to tackle new developments with existing codes that usually date back several decades and with conceptual frameworks unsuited for the job, or develop new legislation while at the same time technologies develop at a far more rapid pace, law as a technique and a system is not well-equipped to react adequately to the technological determinism of forms of AI that are by now already integrated into our societies. This is also to say that while the means are there, we still need to consider the ends, lest we give ourselves over to the

34 See the Dutch Planbureau voor de Leefomgeving (Netherlands Environmental Assessment Agency) report, *Mobiliteit en electriciteit in het digitale tijdperk. Publieke waarden onder spanning* (The Hague: 2017), 1–111, 80 and 82.

35 See Jeanne Gaakeer “The knowledge of causes and the secret motions of things’: autonomous vehicles, ‘responsibility gaps’, and criminal law,” in *Human-Robot Interaction – A Digital Shift in the Law?*, eds Sabine Gless et al. (Cambridge: Cambridge University Press, forthcoming 2024) for an extensive analysis of criminal responsibility and ADS.

36 Jacques Ellul, *The Technological Society* [1954], trans. John Ilkinson (London: Jonathan Cape, 1965), 12.

engineers and the programmers. Or, as John Caputo puts it, we *cannot* not desire a humanism in the sense of

a new post-human humanism, which would turn on a non-programmable programmability that was intent on keeping the future *open* to the coming of what we cannot see coming, which of course is a very risky hermeneutic business; thus, programmability is the new determinism.³⁷

The problem of law lagging behind is exacerbated in ADS by their refinement in terms of the various levels of autonomy and automation of vehicles as distinguished by the American Society of Automotive Engineers (SAE), from level 0 to 1 or cars as we know them without driving automation, or with only minimal driver assistance such as lane-centering and (adaptive) cruise control to levels 4 and 5, high driving automation and full automation, i.e., when the autonomous system controls the vehicle in all circumstances and no activity of the human passenger is required.³⁸ The word driver is then no longer appropriate here because level 5 vehicles can do without the steering wheel and the brake pedal. Since in most countries Traffic Codes still think in terms of the human driver this is a problem that requires new legislation on the concept of “driver”.

Allow me to give some examples to illustrate the point. When in the Dutch town of Appelscha, a pilot study was conducted with an autonomous electric mini-van (i.e., a vehicle without a driver, a steering wheel and pedals, maximum speed 15 km per hour) it drove on a bicycle path. This made for some hilarious pictures of confused cyclists,³⁹ but one cannot help but wonder what would have been the legal case in terms of civil and criminal responsibility when anyone would have been run over by the mini-van on the bicycle path, both given the rules of the current Dutch Traffic Act and the Traffic Rules and Road Signs Regulations, and given questions of corporate liability because the vehicle did not have a human driver. The same pertains to comparable test drives in the Netherlands with ATO (Automatic Train Operation, resembling the auto-pilot in airplanes, i.e., with an engine driver to intervene in case of an emergency). The AOT has the advantage that it can drive more closely to the train before it, so that more trains can be scheduled and more passengers transported safely (note that it will take at least another 5 years before it is truly operational). But one can question what exactly this new range of the train driver’s civil and criminal responsibilities will be when the computer is supposedly

³⁷ John D. Caputo, *Hermeneutics. Facts and Interpretation in the Age of Information* (London: Penguin Random House, 2018), 166 and 167, referencing Donna Haraway.

³⁸ See SAE International, <https://www.sae.org/>, accessed February 1, 2021. See also the U.S. department of transportation, ‘Federal Automated Vehicles Policy,’ <https://www.transportation.gov/AV>, accessed February 1, 2021.

³⁹ For one such picture, see the Dutch Planbureau, *Mobiliteit en electriciteit*, 45.

in control, so the union of train drivers was not amused when they were not consulted.⁴⁰ April 2021 saw the death of two passengers (i.e., nobody occupied the driver's seat) who were testing the Autopilot of their vehicle; Tesla has already claimed that the Autopilot was not engaged.⁴¹ Less tragic is the traffic accident that took place in 2018 in Orange County, Florida, Tesla driver Shawn Hudson sued both Tesla and Oscar Bustamante into whose disabled vehicle negligently left on the road Hudson had crashed, causing damages to his vehicle and multiple injuries to himself. Hudson claimed that the Tesla sales representative had assured him that he only needed to place his hand on the steering wheel now and then because the vehicle would do everything else.⁴² By way of aside, from 1975 onwards European countries have aspired after harmonisation of rules and regulations because of increasing cross-border traffic, one example being the European Agreement on Main International Traffic Arteries (AGR), and they have hotly debated issues such as the (reduction of) width of lanes and the width of the shoulder.⁴³ These things too have to be taken into consideration for ADS.

And while the *Guidelines on the exemption procedure for the EU approval of automated vehicles*, on the one hand, boldly promise, albeit in a language of juridical and hermeneutic disclaimer, the new norm that “The vehicle shall also be designed to minimize potential effects of errors from the vehicle’s users, inside and outside of the vehicle, and of other road users,” to ensure that “if an accident can be safely avoided without causing another it shall be avoided,”⁴⁴ this does not preclude human and technological errors. Neither does it, on the other hand, preclude moral issues. In 2016, Mercedes-Benz admitted to the fact that “they would program their cars to prioritize the lives of the occupants over any other at-risk parties.”⁴⁵ To lower the risk of being caught whilst smuggling drugs, criminals these days are using Autopilot technology. A brother of Pablo Escobar, for example, was said to have bought Teslas to drive his men and their load of drugs to their destinations without a driver at the

40 My source here is the Dutch newspaper *NRC Handelsblad*, 13 February, 2020, Economy section, 8.

41 See <https://www.reuters.com/business/autos-transportation/us-probes-fatal-tesla-crash-believed-be-driverless-2021-04-19/>, accessed April 20, 2021.

42 *Hudson v. Tesla, Inc. and Oscar Enrique Gonzales Bustamante*, Circuit Court of the Ninth Judicial Circuit in and for Orange County, Florida, Filing No. 80052957. For Tesla’s motion to dismiss Hudson’s claim, see Case no. 2018-CA-11812-O. For an extended news cover see also <https://lrus.wolterskluwer.com/news/products-liability-law-daily/tesla-sued-over-allegedly-defective-autopilot-system/64516/>, accessed April 20, 2021.

43 For the 1975 and later amended versions, see www.unece.org, accessed April 20, 2021.

44 https://ec.europa.eu/growth/content/guidelines-exemption-procedure-eu-approval-automated-vehicles_en, 9 April 2019, accessed 2 April, 2020. Cf., SAE International, <https://www.sae.org/>, accessed February 1, 2021; the U.S. department of transportation, “Federal Automated Vehicles Policy,” <https://www.transportation.gov/AV>, accessed February 1, 2021.

45 Zerilli, *Citizen’s Guide*, 171–172, endnote omitted.

wheel.⁴⁶ Such malicious use of AI shows the responsibility gap in current criminal law in that criminal “behaviour” of non-human beings when AI is in the driver’s seat poses difficult questions in terms of responsibility, given the current definition of a criminal offence, i.e., “‘actus reus’ committed by a capable subject + ‘mens rea’ [modified by principles of inchoate, participatory, strict and vicarious liability] – defence = offence.”⁴⁷ This is especially so if we would once and for all want to “specify necessary and sufficient conditions of criminal liability: for instance that we are liable for and only for the choice to do or to risk doing what the law defines as the actus reus of a crime.”⁴⁸ Fault-based liability requires *mens rea*, and the voluntary aspect remains problematic with AI, even though one step further, in terms of the theory of voluntarism, actor A is free only when he could have decided otherwise *under the same circumstances*, a theory outside the scope of this article but interesting nevertheless for the relation between free will and learning abilities of an ADS. What is more, “if we admit that an AA (artificial agent) can perform the *actus reus* [...] greater AA autonomy increases the chance of the criminal act or omission being decoupled from the mental state,” so that “even if we understand that an AA can perform the *actus reus* autonomously, the complexity of the AA’s programming makes it possible that the designer, developer or deployer (i.e., a human agent) will neither know nor predict the AA’s criminal act or omission.”⁴⁹ The deployer, however, may well be driving hand-free, but she is not allowed to drive mind-free yet.⁵⁰ She has to carefully watch the road for any bumps ahead and for other road-users’ behaviour. So far the power of perception of a responsible human being is far more sophisticated than that of the ADS. Relying on the automation is risky as the car

46 See <https://metro.co.uk/2019/07/15/drug-smugglers-using-teslas-autopilot-dodge-police-pablo-escobars-brother-claims-10319181/> accessed August 20, 2021.

47 Celia Wells and Oliver Quick (eds), *Lacey, Wells and Quick: Reconstructing Criminal Law, Text and Materials* (Cambridge: Cambridge University Press, 4th edition, 2010), 97. Cf., Nicola Lacey, *In Search of Criminal Responsibility: Ideas, Interest, and Institutions* (Oxford Legal Scholarship Online, 2016, DOI: 10.1093/acprof:oso/9780199248209.001.0001), 1, “The question whether an offender is responsible for an offence, along with the rules and institutional mechanism for defining and proving that responsibility, is crucial to the core modality of criminal justice: to distinguish guilt from innocence, legitimate from illegitimate conduct.” For a discussion of the implications of generalist and particularist theories on *actus reus* and *mens rea* for legislation in general, see Kenneth W. Simons, “Understanding the Topography of Moral and Criminal Law Norms,” in *Philosophical Foundations of Criminal Law*, eds R. Antony Duff and Stuart P. Green (Oxford: Oxford University Press, 2011), 228–252.

48 R. Antony Duff, “Virtue, Vice, and Criminal Liability: Do We Want an Aristotelian Criminal Law?,” *Buffalo Criminal Law Review* 6 (2002): 147–184, 147.

49 Thomas King, Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi, “Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions,” 22 May, 2018, 1–36, 7 and 8, <https://ssrn-om.eur.idm.oclc.org/abstract=3183238>, accessed December 20, 2020.

50 Zerilli, *Citizen’s Guide*, 86.

accident in which Elaine Herzberg was killed shows. Herzberg was crossing the motorway but the ADS did not detect her as a human being because she was jay-walking. Vasquez, the driver, is currently on trial in Maricopa County, Arizona, charged with negligent homicide because she did not mind the road ahead of her but relied on the automation instead.⁵¹

Mireille Hildebrandt has pointed to a specific problem when it comes to attributing criminal liability to AI devices in court situations: firstly, how is the judge to identify it as “a stable entity over the course of time” [or, I would add, does the judge know one when she sees one?] and, secondly, how to ascertain that the causal link attributed to the criminal wrong is reasonable.⁵² Or, in other words, how proximate should the perceived proximate cause be? To Hildebrandt, it does not make sense to punish a device or software because devices do not have intentions, i.e., they have no *mens rea*, so that traditional punitive-retributive notions do not work either.⁵³ In the current legal and technological framework, I could not agree more as far as intention is concerned. Since causality is not something “out there” but a form of attribution, redistributing risk may therefore be a viable option, but to Hildebrandt it would have to be one beyond the traditional dichotomy of either blaming the user or the designer as civil product liability law does.⁵⁴ She proposes a “generic concept of normativity defined as ‘the way humans or non-humans constrain human and non-human interaction.’”⁵⁵

So because autonomous vehicles run on programming that is itself a designed environment, as noted above, the question who designs what, who decides what will be in the programme and for what reasons becomes acute. To reiterate, with technological developments come challenges to our hermeneutical options, or, as Van den Hoven suggests, “it is crucial that computer scientists, programmers, designers and those who instruct and employ them do not gain inappropriate influence over our collective hermeneutical resources, while shutting the rest of us out.”⁵⁶ Here is a task for the hermeneuts in the humanities. Put differently, what if AI technology

51 See <https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-ars-kills-woman-arizona-tempe/>, accessed December 13, 2020. See Rafael-Vasquez-GJ-Indictment (maricopacountyattorney.org), accessed December 13, 2020.

52 Mireille Hildebrandt, “Criminal Liability and ‘Smart’ Environments,” in *Philosophical Foundations of Criminal Law*, eds R. Antony Duff and Stuart P. Green (Oxford: Oxford University Press, 2011): 507–532, 529.

53 Mireille Hildebrandt, “Ambient Intelligence, Criminal Liability and Democracy,” *Criminal Law and Philosophy* 2 (2008), 163–180, 167.

54 Hildebrandt, “Ambient Intelligence,” 166.

55 Hildebrandt, “Ambient Intelligence,” 171. Hildebrandt rejects strict liability in criminal law because it sits uneasily with the demand of establishing guilt in the sense of wrongful, culpable behaviour.

56 Van den Hoven, “Hermeneutical Injustice,” 12.

exceeds the human capacity to use and react to it properly in any practical situation that crops up in- and outside court, what if AI exceeds an ordinary human consumer's understanding? Will all of this eventually lead to new narratives in and of civil liability and criminal responsibility? As to the latter, what will then be the new *mala prohibita*? Whose powers of perception and/or agency will be deemed relevant? What will be the new narratives in court, especially in the period of transition in which both traditional and autonomous vehicles find themselves together on the road? In case of a traffic accident in which a current type of autonomous car is involved, to illustrate the point with a practical example, fire fighters and ambulance personnel already have to be extra careful because of the risks involved with the battery and electric circuit of the vehicle when compared to a traditional car, whether the vehicle is on fire or not.⁵⁷ But how are they to know the technological details of each and every ADS that they encounter in an emergency situation? If manufacturers fail to address the limitations of human drivers and fail to provide adequate information and training – a European trainer complained that Tesla systems are developed for US American rather than European roads, hence the need for a specific training for European users, a charge that the manufacturer, however, called “unfounded”⁵⁸ – the legislator should act on various normative levels. The new EU General Safety Regulation, due to become effective in 2022, aims at the reduction of the number of traffic casualties by means of the demand of, amongst others, the implementation in ADS of Intelligent Speed Assistants, Advanced Emergency Braking Systems and Event Data Recorders comparable to the black box found in airplanes, but as long as national governments do not register all traffic incidents with ADS, manufacturers will not be able to learn from faults in the system that led to (near-) accidents. Road traffic doctrine will be challenged especially by novice drivers who have much to lose, for example their vehicle when that is confiscated by the police if the driver was also under the influence of alcohol and drugs. Needless to say that all of the above can have great impact on the court's case load. Judging as much as judging the technology in a specific case of automated driving can become even more complicated if any traffic offence is detected by means of an automatic smart camera that captures an image of the driver, the vehicle and the registration number. Since in most legal systems the penalties for traffic offences include demerit points on the

⁵⁷ If an electric car has been on fire it often takes more than a week to cool the battery. This is usually done by placing the car in a large container filled with water. In the Netherlands a firefighter has recently developed a ‘battery fire cooler’ in order to accelerate the process of cooling. See the Dutch newspaper *PZC*, Zeeland sect., 25 January, 2021, 5.

⁵⁸ On 2 January 2020, the Dutch newspaper *NRC Handelsblad* featured an article by Marc Hijink, “Hoe lang mag je met losse handen rijden?,” on driving ADS “with no hands”, Economy section, 6 and 7. See also <https://racfoundation.org>, accessed January 17, 2021, for a study that human driver training is necessary if the driver is to react adequately when the vehicle thrusts back control.

driving licence, one can safely assume that the proper functioning of the digital camera will be questioned as well,⁵⁹ as will privacy issues with respect to data retention.

One thing is certain: if ADS do not function as they claim they will, the whole concept of foreseeability by the human will find itself under pressure and so will the current rules of civil and criminal liability and insurance if responsibility is going to be shared and if so, in which proportion. The European Parliament has already had a study conducted, *A common EU approach to liability rules and insurance for connected and autonomous vehicles*, in order to provide “a scientifically based evaluation and assessment of the potential added value of taking legislative action at the EU level,” not least because of economic purposes because the sector of ADS manufacturers is expected to contribute 17 trillion euros to the EU economy in 2050,⁶⁰ while globally this amount will already be reached by 2030. So in view of the new risks (relating to software and network failures as much as to mistakes in programming) it makes sense to carefully monitor shifting liabilities since Murphy’s Law will apply as it always does. In 2018 the *Automated and Electric Vehicles Act* was implemented in England, Scotland and Wales.⁶¹ In the Netherlands product liability is based on the direct effect in Dutch law of the EU Directive 85/374/EEG, OJ L 2010.⁶² But which legal framework will apply, now that we are post-Brexit, if a European ADS either causes or becomes the victim of an accident with a similar vehicle in the UK? The same point about harmonisation applies to other non-EU countries such as Switzerland that quite rightly also pointed to the risk involved in inevitable data exchanges.⁶³

59 For the use of cameras in probative surroundings, see also Peter Brooks, “Scott v. Harris: The Supreme Court’s Reality Effect,” *Law & Literature* 29 (2017) 143–149, and Bennett Capers, “Video as Text/Archive,” in *The Oxford Handbook Law and Humanities*, eds Simon Stern, Maksymilian Del Mar and Bernadette Meyler (Oxford: Oxford University Press, 2020): 792–794.

60 *A common EU approach to liability rules and insurance for connected and autonomous vehicles*, European Added Value Assessment, Accompanying the European Parliament’s legislative own-initiative report (Rapporteur: Mady Delvaux) study by Tatjana Evas, February 2018, 1 and 7, [www.europarl.europa.eu/RegData/etudes/STUD/2018/615635/EPRS_STU\(2018\)615635](http://www.europarl.europa.eu/RegData/etudes/STUD/2018/615635/EPRS_STU(2018)615635), accessed December 13, 2020.

61 See the *Automated and Electric Vehicles Bill Explanatory Notes*, House of Lords, <https://service.parliament.uk/Bills/2017-19/automatedandelectricvehicles/documents.html>, accessed January 1, 2021.

62 See <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32010L0040&from=EN>, accessed January 1, 2020.

63 For the Swiss pilot exemption procedure, see <https://www.astra.admin.ch/astra/de/home/themen/intelligente-mobilitaet/pilotversuche.html>, accessed December 13, 2020. The English text of article 3.7 Data Security of the ‘Automated Driving, Fact Sheet For Conducting Pilot Tests in Switzerland’ reads “An automated vehicle cannot be operated without data exchange. Leaks and false data have a direct impact on the safety of the road users and passengers. Thus data security is of the utmost importance.”

But there is a complication in that so far

most contemporary philosophical analyses of responsibility are grounded in notions of human agency that emerged in Europe in the philosophy of the Enlightenment. The ideas of the self-determining moral agent, equipped with distinctive cognitive and volitional capacities of understanding and self-control, and of a universal human personhood underpinned by these features have been of crucial importance to the gradual development of modern societies in a number of very obvious ways.⁶⁴

What is more, Santoni de Sio, when discussing interdisciplinarity as one of the requirements of the “Responsible Innovation” approach in the ethics of technology points to the necessary collaboration of “engineers, philosophers, policy-makers and stakeholders” but oddly enough given the prominence of human rights in contemporary legal theory and practice this approach fails to mention law.⁶⁵ Others, when answering in the affirmative the question whether in case an autonomous vehicle is hacked the software producer of ADS can fight off product liability, specify as its condition that “the defect could not have been discovered taking all available *scientific and technical knowledge* into account,”⁶⁶ thus bypassing other sources of defects as well as other disciplinary discourses. So to me at least there is good reason to claim that humans should inform science, and not the other way around, even if this is the easy way out in any situation in which we do not yet fully grasp the consequences of such technologies. Why? Because Lady Lovelace’s Objection, i.e., that the machine only does what is programmed into it by us, has proved to be untenable in the age of ML.⁶⁷ That is why we should reflect and act at the start of any new development, and decide about the parameters, rather than react to faults and/or (near) accidents.

64 Nicola Lacey, *In Search of Criminal Responsibility: Ideas, Interest, and Institutions* (Oxford Legal Scholarship Online, 2016, DOI: 0.1093/acprof:oso/9780199248209.001.0001), 5.

65 Filippo Santoni de Sio, “Ethics and Self-Driving Cars: a White Paper on Responsible Innovation in Automated Driving Systems,” Delft University of Technology, 2016, commissioned by Rijkswaterstaat for the “knowledge agenda automated driving”, 1–33, 6.

66 Esther F. D. Engelhard and Roeland W. de Bruin, Annex I “EU Common Approach on the liability rules and insurance related to Connected and Autonomous Vehicles Final Report” in “A common EU approach to liability rules and insurance for connected and autonomous vehicles,” European Added Value Assessment, Accompanying the European Parliament’s legislative own-initiative report. Rapporteur: Mady Delvaux. Study: Tatjana Evas, February 2018, [www.europarl.europa.eu/RegData/etudes/STUD/2018/615635/EPRS_STU\(2018\)615635](http://www.europarl.europa.eu/RegData/etudes/STUD/2018/615635/EPRS_STU(2018)615635), accessed December 13, 2020, 38–131, 84, [my italics].

67 Alan Turing, “Computing Machinery and Intelligence,” *Mind* 59 (1950): 433–460, 450. Cf., Ugo Pagallo, “What Robots Want; Autonomous Machines, Codes and New Frontiers of Legal Responsibility,” in *Human Law and Computer Law: Comparative Perspectives*, eds Mireille Hildebrandt and Jeanne Gaakeer (Dordrecht: Springer, 2013): 47–65.

2.2 Robotic Weapons: In or Out of Remote Control?

Cars have for obvious reasons traditionally been seen as “potentially dangerous weapons.”⁶⁸ The same goes for (L)AWS, i.e., (lethal) automatic weapon systems, i.e., unmanned, pilotless, aircrafts or other military vehicles for robotic warfare, and commercial drones that are comparable to a great extent to current ADS, at least as far as their technology and their different levels of sophistication of algorithm decision making are concerned. In 2011 Muammar Gaddafi’s convoy fleeing the rebels that would eventually depose him, was bombed by an American Predator drone. The air force base from which it was flown was in Sicily. A base near Las Vegas controlled the drone via satellite.⁶⁹ Current LAWS find their target autonomously. The only thing they need is human activation. Here is another example of literally posthuman technology: it is robotic warfare in which the human soldier who used to do the killing is replaced by AI, even though some humans were obviously involved in designing and operating the device. Both forms come with legal problems.

To start with drones, in the Netherlands filmmakers and photographers who want to use drones professionally have to go through a lot of steps in order to obtain a permit. Any individual, however, who buys a drone for the fun of it, can fly the device without a permit because according to current law an over the counter drone is considered a model plane for which no permit is required.⁷⁰ Needless to say that this is an unfortunate difference. While the Aviation Code applies to drones so that a conceptual and/or semantic problem may well be the case (by way of aside: in the good old days a surfboard fell under the legal definition of a ship in the Dutch Sailing Regulation of which the responsibilities for the “skipper” looked hilarious to most surfers), civil responsibility and liability has not been settled yet, and so there are great legal uncertainties with respect not only to safety and privacy but also to criminal responsibility and the onus of proof. On this plane, too, the legislator has to act.

With AWS, the problem of responsibility is exponentially augmented in terms of the principles of a just war, a concept addressed by the legal theorist Hugo Grotius in *De iure belli ac pacis*, and developed in the Geneva Conventions of 1949 and their Additional Protocols with respect to rules of war for humans, as applied by courts such as the International Criminal Court (ICC) and the International Tribunal for the former Yugoslavia (ICTY).⁷¹ How do we ensure that AWS follow them if they are

68 Santoni de Sio, “Ethics and Self-Driving Cars,” 20.

69 See Braidotti, *The Posthuman*, 8 and 9, referencing *The Daily Telegraph*, 21 October, 2013.

70 Corien Prins, “Drones en een strategische wetgevingsagenda,” *Nederlands Juristenblad*, 31 August, 2018, 1423.

71 For a general discussion of just war principles and unmanned military vehicles, see Matthew W. Hallgarth, “Just War Theory and Remote Military Technology: A Primer,” in *Killing by Remote Control*:

“ethically blind” and cannot themselves assess the “ethically relevant effects of their decisions, and they lack other capacities such as empathy that are important to human morality.”⁷² Rather than cherishing what Carl Schmitt had called a religious belief in technology, we should remain highly critical.⁷³ To reiterate, who decides and who programs the AWS is the crucial question, as is the fact that law lags behind on this plane of automation as well. As Medea Benjamin wryly notes, “the legal underpinning for drone attacks is the 2001 Authorization for the Use of Military Force [...] passed just one week after 9/11.”⁷⁴ In the case of “killing by remote control,” either in- or outside the context of a just war, the first condition of criminal liability being the *actus reus* as a voluntary criminal act (or omission) does not sit easily with the fact that once an unmanned AWS is activated, it acts without the human operator being able to intervene. This of course also why in a just war so often civilians are killed together with the true object of the targeting, mostly the “enemy combatant” in Guantanamo Bay parlance. *Mens rea* as the intention to commit the *actus reus* by means of an AI is another problem if it is the human being who has the *mens rea* that then causes the AI to perform the *actus reus*. Not only is it impossible to take away *mens rea* from criminal law as we know it, but also in terms of probative value any disconnection of *mens rea* and *actus reus* is extremely problematic, especially if we recall that current deep learning learns from what it does and may act differently from what the human being operating it has in mind. Lack of predictability of AI means lack of control of the human. As with ADS, the notion of human responsibility is challenged. Jarna Petman asks us to consider “if fully autonomous systems will replace the role of humans in warfare, then where will accountability lie when something goes wrong?”⁷⁵ That is to ask whether to blame the front-line operator, the computer programmers and designers, or the reasonable military commander for breaches of international humanitarian law, or human rights violations generally. What is more, systems capable of learning independently of their human programmers may decide not to obey orders. Rebecca Crootof argues that precisely because AWS can select and engage targets independently, the very fact that no individual can be said to have acted intentionally results in the negative consequence

the Ethics of an Unmanned Military, ed. Bradley Jay Strawser (Oxford: Oxford Scholarship Online, 2013, DOI:10.1093/acprof:oso/9780199926121.001.0001): 26–46.

72 Allen, “Artificial life,” 230.

73 Carl Schmitt, “The Age of Neutralizations and Depoliticizations” [1929], trans M. Konzett and J. P. McCormick, *Telos* 96 (1993):130–142.

74 Medea Benjamin, *Drone Warfare. Killing by remote control* (London and New York: Verso, 2013), 129.

75 Jarna Petman, “Autonomous Weapons Systems and International Humanitarian Law: ‘Out Of The Loop’?,” Helsinki, research paper available at the Eric Castren Institute of International Law and Human Rights, accessed 25, October, 2021, 1–76, 43.

of nobody being held criminally responsible under international law.⁷⁶ Thus war crimes may be committed while nobody is brought to justice. It is because AWS and ADS share relevant legal and moral issues, given their autonomous decision making algorithms programmed by engineers, that the ethical aspects of automation deserve highlighting. This is, on the one hand, because designers and engineers are far too busy inventing new technologies rather than minding the ethics. And, on the other hand, because, as George Lucas notes, of the dominance of rationality-based conceptions of deliberation in the Western world, i.e.,

very little basic research has been conducted that might contribute to the feasibility of autonomous systems one day possessing the requisite capacities for full character recognition and situational awareness [...] that would be required, in turn, to achieve some sort of reliable analogue to ethical judgment and practical reasoning in the human case.⁷⁷

Here is another gap the humanities may help fill. To neglect ethical principles and/or human rights principles when designing may lead to huge costs once adaptations have to be made at a late stage of implementation so that is why human rights theorists, legal philosophers and legal theorists may for once perhaps prove their economic value and societal impact as well.

How responsibilities are to be distributed in an ethical manner is addressed by Luciano Floridi who notes that a problem in distributed environments with networks of both human and artificial agents “may produce morally good or evil (henceforth loaded) actions though interactions that are not, in themselves, morally loaded but neutral”; he calls this phenomenon “distributed moral actions (DMAs).”⁷⁸ A case in point is the collateral damage ensued by targeted killing by signature strikes with drones in which civilians are killed, and personality strikes of specifically named persons/terrorists that kill the wrong person. A civilian may only be targeted if (s)he “directly participates in hostilities,” in any other situation (s)he is not a legitimate target.⁷⁹ So with collateral damage undoubtedly comes the semantic-definitional debate about the range of this rule: is one a civilian only or a participating civilian? The core of the word may be clear but the penumbra opaque. This matters not least because the risk of unintended harm to civilians must basically be

76 Rebecca Crotoof, “War Torts: Accountability for Autonomous Weapons,” *University of Pennsylvania Law Review* 164 (2016):1347–1402.

77 George R. Lucas, “Engineering, Ethics, and Industry: The Moral Challenges of Lethal Autonomy,” in *Killing by Remote Control: the Ethics of an Unmanned Military*, ed. Bradley Jay Strawser (Oxford: Oxford Scholarship Online, 2013), DOI:10.1093/acprof:oso/9780199926121.001.0001, 212–229, 223.

78 Luciano Floridi, “Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions,” *Philosophical Transactions Royal Society* (2016): 1–13, 2, available at <https://doi.org/10.1098/rsta.2016.0112>, accessed December 6, 2020.

79 Benjamin, *Drone Warfare*, 133.

borne by the combatant(s).⁸⁰ Furthermore, it should be noted that an indiscriminate attack that does not distinguish between lawful and unlawful targets is a war crime but it is not one codified in the (First Additional Protocol to) the Geneva Convention that lists all other accepted war crimes. So it is up to the international courts to provide answers. Outside armed conflict the rules are stricter, any killing “must be necessary to protect life and there must be no other means, such as capture or nonlethal incapacitation, to prevent that threat to life.”⁸¹ So one wonders what this could have meant for the striking of Gaddafi. This is also to say that with robotic warfare, the principles of just war are often under heavy moral pressure, when there is indeed someone who actually does what can be done because technology allows it.⁸²

What then, next, if a combination of actions that are thought of as neutral has as its consequence a DMA, who or what is responsible? Traditional ethical frameworks fall short when it comes to answering this question, Floridi claims. His argument also brings us back to law’s general penchant for (Kantian) autonomous agency and criminal law’s favouring intent and free will in relation to it. What is more, not only is AI itself a “black box” as noted above, but, as Luciano Floridi rightly points out, our human inability to understand AI generates “a ‘black box’ mentality”⁸³ in us. As a result, I suggest, we do not protest as much against the designers and producers of AI devices, let alone against our governments and legislators that fail to implement a normative legal framework to deal with liabilities, responsibilities and risks created by misuse. In short, we do not take to the streets when a Tesla crashes or when a presumed terrorist is being selected, targeted and killed by a robotic weapon. As Zerilli notes, “the most popular catchphrase at the moment – especially in discussions about the military deployment of lethal autonomous weapons – is the call for ‘meaningful human control,’” and this is far more demanding than “ultimate control” which means “that we have the power to ‘switch off’ a system that has gone rogue to prevent *further* damage.”⁸⁴ That is to say “meaningful control” as defined by the 2014 International Committee for Robot Arms Control and amended by the Center for a

80 Asa Kasher and Avery Plaw, “Distinguishing Drones, An Exchange,” in *Killing by Remote Control: the Ethics of an Unmanned Military*, ed. Bradley Jay Strawser (Oxford: Oxford Scholarship Online, 2013, DOI:10.1093/acprof:oso/9780199926121.001.0001): 48–66, 56 and 57, referencing Michael Walzer, *Just and Unjust Wars*.

81 Benjamin, *Drone Warfare*, 133.

82 See Marjorie Cohn (ed.), *Drones and Targeted Killing. Legal, Moral, and Geopolitical Issues* (Northampton MA: Olive Branch Press, 2015), Part II on collateral damage.

83 Luciano Floridi et al., “AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations,” *Minds and Machines* 28 (2018): 689–707, 692.

84 Zerilli, *Citizen’s Guide*, 79 and 80 [my italics].

New American Security is about preventing/averting disaster rather than having merely damage control.⁸⁵ The issues of autonomy, agency, training, (human) control, and (criminal) intent return with a vengeance. Here the human still matters, as does her ability, or practical wisdom gained by experience, to gauge the (L)AWS in its spatial and temporal situatedness. Therefore her authority should not be “off the loop,”⁸⁶ if her control is to be meaningful, which leads us to another semantic minefield. After all, who decides – like Humpty Dumpty – what is meaningful, if we think of the abovementioned example of Mercedes prioritising the one human above the other?⁸⁷ For meaningful control to be successfully implemented we should promote “technologies of humility” precisely because informed consent to anything technological is a state that is hard to get at,⁸⁸ rather than continuing the various forms of “black box” mentality. We should therefore consider the interdisciplinary question why new technologies need the humanities. As the philosopher Edmund Husserl had claimed, a determination of the world by the positive sciences as took place in the second half of the nineteenth century occasioned “an indifferent turning-away from the questions which are decisive for a genuine humanity. Merely fact-minded sciences make merely fact-minded people.”⁸⁹ Yet, facts, too, are always to a certain degree perspectival narrative, so that it is the hermeneut from the humanities who needs to reconsider the primacy of instrumental reasoning in technology and who needs to point out the need for serious moral deliberation.⁹⁰

85 For an extensive analysis of the term “meaningful control,” see Connal Parsley, “Automating authority. The human and automation in legal discourse on the Meaningful Human Control of Lethal Autonomous Weapons Systems,” in *The Routledge Handbook of International Law and the Humanities*, eds Shane Chalmers and Sundhya Pahuja (Milton Park and New York: Routledge, 2021): 432–445.

86 Zerilli, *Citizen's Guide*, 81.

87 Cf., Ferdinand von Schirach's play *Terror* on whether or not a civilian airplane taken over by a terrorist should be shot down in order to prevent it from crashing into a full football stadium. Ferdinand von Schirach, *Terror*, trans David Tushingham (London: Faber and Faber, 2017). See also Jeanne Gaakeer, “Wrest once the law to your authority. To do a great right, do a little wrong?”, in *Law and the Humanities: Cultural Perspectives*, eds Chiara Battisti and Sidia Fiorato (Berlin: De Gruyter, 2019): 477–498.

88 Ibo van de Poel, “An Ethical Framework for Evaluating Experimental Technology,” *Science and Engineering Ethics* 22 (2016): 667–686, 668, referencing S. Jasanoff, “Technologies of Humility: Citizen participation in governing science,” *Minerva* 41.3 (2003): 223–244.

89 Edmund Husserl, *The Crisis of European Sciences and Transcendental Phenomenology, An Introduction to Phenomenological Philosophy* [1954], trans. David Carr (Evanston (Ill.): Northwestern University Press, 1970), 6.

90 Cf., Charles Taylor, *The Ethics of Authenticity* (Cambridge MA: Harvard University Press, 1991), 5 and 8.

3 Knowledge of Words and/or of Things? Kazuo Ishiguro's *Klara and The Sun*

3.1 Knowing Consciously

While both humans and AI learn by iteration, one big advantage that humans still have, is that their programming allows multitasking. However sophisticated ADS and (L)AWS are, they can only perform the one specific task that they have been programmed to do – either drive a car or launch a missile – even though they learn by doing on the basis of specific datasets chosen by others. Humans, by contrast, can drive cars, give lectures, prepare meals, operate on patients, and so on and so forth, all on the basis of their one, neurologically sophisticated programme. This speaks for the superiority of the human intelligence that thinks predominantly in terms of causality rather than correlation as ML algorithms do. What is more, humans *know* what they do (at least most of the time if they are sane), and “*Man knows, and he is also conscious that he knows.*”⁹¹ As the examples of ADS and (L)AWS show, machines don’t and aren’t. So the chances of AI outwitting us humans are still in the realm of sci-fi only. Nevertheless, the today not so fictional question that Samuel Butler raised, i.e., whether “because mechanical life is a very different thing from ours, therefore it is not life at all?”⁹² should be on our minds, not least because Butler adds that if it is considered life it could have a will of its own as well.

Therefore, epistemology, being the area of interdisciplinary enquiry in the humanities, contrasts the central question “Can we, and if so, how do we understand human behaviour?” to the explanatory methodologies that are in conformity with the paradigms developed in the natural sciences. With the onset of autonomous AI devices in our lives, the question resonates in technology: Can we understand non-human “behaviour”? What is more, it should not be forgotten that our concept of knowledge is cultural in that it is deeply connected to “What’s stored in the individual mind – events, facts, concepts, skills.”⁹³ On this plane, epistemology is also challenged by AI. It is yet another reason for a strong bond between technological research and the reflexive knowledge provided by the humanities, “which are responsible for criticism and hence are withdrawn from markets and practical concerns,” as Immanuel Kant had noted in 1789 when he discussed the struggle of and between the

91 Stephen Toulmin, *Human Understanding*, volume I (Oxford: Clarendon Press, 1972): 1 [italics in the original].

92 Samuel Butler, *Erewhon* [1872] (Harmondsworth UK: Penguin, 1954), 175.

93 Nicholas Carr, *The Shallows, what the Internet is doing to our brains* (New York and London: W.W. Norton & Co, 2010): 196.

academic faculties.⁹⁴ Thus the hermeneut can help close the divide between technological “knowledge of things” and humanistic “knowledge of words”,⁹⁵ i.e., the “tectonic shift in intellectual geography”⁹⁶ that took place in the nineteenth century when in the universities the sciences began to differentiate from the humanities.

This is so important to keep in mind because since the advent of modernity, the natural sciences have divorced *physis* as matter from the (technological) making and knowing in the Heideggerian definition noted in Section 1. As Ellul noted when he wrote about the future of technology in 1954, “the monolithic technical world that is coming to be,” was meant to be of help to man but because “it has evolved autonomously”, man has lost contact “with his natural framework and has to do only with the organized technical intermediary.”⁹⁷ Ellul feared that technological issues could only be resolved by “a world-wide totalitarian dictatorship which will allow technique its full scope and at the same time resolve the concomitant difficulties.”⁹⁸ The Elon Musks and Mark Zuckerbergs of this world have already proved that Ellul was right here in presaging the likes of the Metaverse.

The divorce of the faculties has also led to a form of Cartesianism that has spread to law in its form of technique or policy to order societies.⁹⁹ Furthermore, with undisputed “things” usually comes a referential view on the character and function of language which leads to a restricted idea of language application as the description of reality that under the influence of Enlightenment thought became characteristic of the natural sciences, the repercussions of which resonate in new technologies well into the twenty-first century.¹⁰⁰ Dictatorship or imperialism are then indeed the next

94 Immanuel Kant, *The conflicts of the faculties* [1789] (Lincoln NE: University of Nebraska Press, 1992).

95 Desiderius Erasmus, *On the Method of Study, de ratione studii ac legendi interpretandique auctores*, in Desiderius Erasmus, *Collected Works of Erasmus, literary and educational writings 2: de copia/de ratione studii*, ed. C.R. Thompson, trans B. McGregor, (Toronto, Buffalo and London: University of Toronto Press, 1978), 666; Matthew Arnold, “Literature and Science,” address given in 1883 in America. For purposes of this article I used the electronic edition by Ian Lancashire, <http://www.chass.utoronto.ca/~ian/arnold.htm>, accessed September 5, 2021, 2 and 4.

96 Jessica Riskin, “Just use your thinking pump!,” *New York Review of Books* 2 July, 2020, 48–50, 48, reviewing Henry M. Cowles, *The Scientific Method: An Evolution in Thinking from Darwin to Dewey*.

97 Ellul, *The Technological Society*, 428.

98 Ellul, *The Technological Society*, 433 and 434.

99 Cf., Robert Weisberg, “Proclaiming Trials as Narratives: Premises and Pretenses,” in *Law’s Stories. Narrative and Rhetoric in the Law*, eds Peter Brooks and Paul Gewirtz (New Haven: Yale University Press, 1996): 61–83, 62, “‘Cartesianism’ [...] treats law as a universal set of policy algorithms for addressing apparently different but essentially generic and recurring patterns of social need.”

100 It is connected to a neutral vehicle for the communication of our “thoughts” that are our perceptions of “things”, entities in the world transmitted by means of words, a view also exemplified in Thomas Hobbes’ *Leviathan* and criticised by James Boyd White, *Justice as Translation. An Essay in Cultural and Legal Criticism* (Chicago: University of Chicago Press, 1990).

steps. With this in mind I turn to Ishiguro's novel *Klara and the Sun*¹⁰¹ in which the omniscient narrator Klara is a *femme-machine* if we consider the long tradition of mechanistic concepts of the human, from Descartes' *Traité de l'homme* (1648) and De la Mettrie's *L'homme machine* (1748) to the Czech author Karl Čapek's 1921 play *R.U.R.*, i.e., Rossum's Universal Robots.¹⁰²

3.2 In and Out of the Black Box?

Klara is an Artificial Friend (AF), B2 4th series programmed to be friends with lonely children. As the omniscient narrator, she tells the story of what happened to her and the young teenager Josie that chose her to lift her loneliness and isolation now that she has fallen ill after having been "lifted",¹⁰³ Ishiguro's circumspect term (provocative indirection and understatement being Ishiguro's tone of voice in most of his novels) for the genetic modification, or gene editing that Josie underwent to be able to go to the best colleges in the country. The story is told in retrospect when Klara is already in the Yard, the scrapheap to which she has been demoted after Josie regained her health and went off to college. There her old Manager – note that in the novel the characters are somewhat dehumanised in that they are allegorically called by their role name: Mother, Father, Melania Housekeeper – who comes looking for old AFs tells her that "Of all the AFs I looked after, you were certainly one of the most remarkable. You had such unusual insight. And observational abilities" [304].

Klara's abilities are indeed remarkable. From the window of the shop in which she is displayed at the start of the narrative, she quickly learns to distinguish between anger and sadness in the children staring at her. As Manager says, "you notice and absorb so much" [8]. Other AFs like her colleague Rosa miss many signals in the child-AF relationships they watch. Klara, however, is anxious when she reflects on how it might be if a child didn't want you. Therefore she starts training by looking at passers-by to recognise "the sort of behavior about which I needed to learn" [17] and she starts to imagine other people's feelings and emotions by trying to "find the beginnings of such a feeling in my own mind" [18] – her programmed mind, mind you – and by watching human behaviour. When Josie comes to the store and asks Klara to wait for her because Mother is reluctant to buy an AF, Klara promises to. She doesn't make an effort to please another child who wants to buy her. The newer

¹⁰¹ Kazuo Ishiguro, *Klara and the Sun* (London: Faber & Faber, 2021).

¹⁰² See also George R. Slusser and George R. Guffey, "Literature and Science," in Jean-Pierre Baricelli and Joseph Gibaldi, *Interrelations of Literature* (New York: MLAA, 1982): 176–203.

¹⁰³ See also Priscilla Wald, "Replicant Being: Law and Strange Life in the Age of Biotechnology," in *New Directions in Law and Literature*, eds Elizabeth S. Anker and Bernadette Meyler (Oxford: Oxford University Press, 2017): 344–358.

series B3 AFs are cognitively higher than Klara, but B3s are so smart that they move away from the B2s in the shop, and Klara wonders “how could they be good AFs for their children if their minds could invent ideas like these?” [35]. As Mother notices too when she and Josie come to buy Klara, “B3s [...] can sometimes be less empathetic” [40]. Klara soon shows her “extraordinary observational understanding” [44] by reproducing Josie’s walk at Mother’s request. Later on it turns out that Klara was indeed chosen for the purpose of becoming the impersonation of Josie, should Josie die of complications as a result of her being “lifted”, as happened to her sister Sal before her. Josie has a friend since childhood, Rick who has not been lifted and who is more or less excluded at a gathering of lifted children. On a trip to a waterfall that Mother and Klara make alone because Josie is too ill, Mother again asks Klara to impersonate Josie. Klara is so successful that Mother gets upset. It soon turns out that Henry Capaldi who is supposedly painting a picture of Josie is building a kind of avatar of Josie, comparable to the one he did of Sal. Capaldi asks Klara to fill in a questionnaire for the purpose of becoming “a *continuation* of Josie” [208, italics in the original] as Josie’s father claims, not an imitation. Mother tells Klara “We’re asking you to *become* her” [209, italics in the original] and that’s why Klara was chosen in the first place. But can Klara “learn her heart?” [218] Klara thinks this possible because even in the poetic sense “there’ll be an end to what there is to learn” [219]. Josie’s father Paul hates Capaldi for his claim that there is no such thing as human uniqueness and Klara understands that rather than her becoming Josie, they need to make Josie healthy. When Josie regains her health and goes off to college, Klara fades away from Josie’s life to the “Utility Room” [296]. Rick also fades away; he won’t go to college “trying to compete with all those lifted kids” [292]. Capaldi visits in order to open the black box that Klara supposedly is because “there’s growing and widespread concern about AFs right now. People saying how you’ve become too clever. They’re afraid because they can’t follow anymore what’s going on in side anymore” [297]. Mother refuses “Leave *our* Klara be. Let her have her slow fade” [298, my italics]. This happens at the end of the novel when Klara is in the Yard. Here Klara muses that her continuation of Josie would in the end have been impossible, because “there would have remained something beyond my reach. The Mother, Rick, Melania Housekeeper, the Father. I’d never have reached what they felt for Josie in their hearts” [306]. Thus, Klara recognizes and accepts her limitations, she sacrifices herself.¹⁰⁴

104 Cf., Jane Hu, “Portrait of the Robot as a Young Woman,” reviewing *Klara and the Sun*, *The New York Review of Books* (November 4, 2021): 15–16, 16. “The problem with programming a machine to feel for others is, of course, that the machine might start to develop other feelings, ones unproductive to her work. But Klara’s evolving emotions are crucial to our understanding of the novel as a technology of interiority. The reader experiences Klara’s care for Josie through Klara’s empathetic narration, in which her desire to see Josie flourish and grow fails to completely suppress Klara’s

So who or what is Klara? She is definitely a form of AI, a DNN that learns from observation. But how does Klara's machine learning system learn to classify what Josie needs? Is she thus programmed or has she become more perceiving in the course of her functioning as AF to Josie? This is to ask after her training data with which her algorithm is fed. They are obviously not perfect because her own penchant for solar energy makes her think that the Sun can help Josie get well. What is more, Klara takes the setting of the Sun in Mr McBain's barn literally and that is a mistake in her object classifiers, i.e., she attributes healing power to the sun that the sun's classification as a specific object does not imply. This is already an example of the fact that the AI of the AF is domain specific and unconscious. Or rather, its ability to infer causality depends on her training data. Another example of a mistake in her programming is that she calls smartphones "oblongs" [7] which reminds us of Greg Raine's poem "A Martian Sends a Postcard Home" with a "Caxton" as its metaphorical reference to a typewriter, and perhaps asks us to consider Klara's otherworldliness. One could also argue that Klara is algorithmic in that her machine learning is limited because it lacks to a large degree the practical wisdom and tacit knowledge that humans acquire by experience.

Klara does not understand either that she is in fact a disposable object, dependent on her owner's whim. When she says "it's now my duty to be Josie's best friend", Josie immediately responds "You're my AF. That's different. But Rick, well, we're going to spend our lives together." [55] She is also viewed as an object by Rick's mother's who when meeting her asks "Are you a guest at all? [...] Or do I treat you like a vacuum cleaner?" [145] This resembles what in Ian McEwan's dystopian short story "Düssel ..." – on the love relation between a human being, a man, and an artificial person called Jenny in an era 50 years from now – is called "the indelicate question", i.e., "Are you real?"¹⁰⁵ When she and Father try to dismantle a machine that Klara thinks is detrimental to Josie's getting well, he too objectifies her when he needs "P-E-G Nine" [226] to do the job because he removes it from Klara's body by making two crude incisions below her ears without knowing for sure whether this will impair Klara's "cognitive" abilities [227].

Does Klara know, then, that she knows? Her attempts to impersonate Josie indicate that she is very successful at what Alan Turing called the "imitation game". To discuss the question "Can machines think?" Turing offered the metaphor of the imitation game asking whether if a machine taking part in this game of questions to find out who is who, the interrogator will decide wrongly as often as when the game

desires. This is, after all, the central dilemma of most AI plots, from Frankenstein onward: the machine will develop a mind of its own and if it knows what's good for it, revolt."

¹⁰⁵ Ian McEwan, "Düssel ...," *The New York Review of Books*, LXV. 12 (July 19 – August 15, 2018): 4 and 6, 4; cf., Gaakeer, "Select before you Collect," 70.

is played with two humans rather than a human and a computer. Therefore Turing asked “Are there imaginable digital computers which would do well in the imitation game?”¹⁰⁶ What is more, her impersonations of Josie become more sophisticated each time, so she learns by doing, and she does her best to help Capaldi as much as she tries to save Josie’s life. And she has meaningful exchanges with her humans, Josie and others, she reads their emotions more than her predecessors Tay the chatbot and Alexa could possibly hope to achieve, or, to name a few others, David the android programmed to love in Stephen Spielberg’s 2001 film *A.I.*, and the computer as an object of love without a human body in Spike Jonze’s film *Her* that supposedly has a consciousness. In this sense the AI in Klara contains affective AI more sophisticated than the systems currently in use at airports to read the facial expressions of those present.

At the end of the novel when, by contrast, Ishiguro turns the tables on the humans, there’s Capaldi who is die-hard Cartesian because he believes “There’s nothing there. Nothing inside Josie that’s beyond the Klaras of this world to continue. The second Josie won’t be a copy. She’ll be the exact same.” [210] In short, Capaldi intends the new Josie to be equal to an AF. Yet, while “The mechanisms that underlie Klara’s thought processes, the technologies that make her existence possible, are as mysterious to her as the workings of human consciousness are to us,”¹⁰⁷ Klara, however, does fathom the human heart: “Mr Capaldi believed there was nothing special inside Josie that couldn’t be continued. He told the Mother he’d searched and searched and found nothing like that.” [306] Klara knows Capaldi is wrong, “he was searching in the wrong place. There *was* something very special, but it wasn’t inside Josie. It was inside those who loved her.” [306] So consciousness of Josie is what makes Josie, as much as Klara’s consciousness of the human heart does. This shows her human-like insight. Yet she recognises her limitations, “I’d never have reached what they felt for Josie in their hearts.” [306] Therefore, *Klara and the Sun* like many other novels of Ishiguro’s shows us that what is at stake is a normative understanding of what it means to be human and to love.¹⁰⁸ That is why we should pay heed to the “autonomous [literary, my remark] figures that may be used as terms of comparison with [...] other constructions of the mind,”¹⁰⁹ or to what Marina Warner calls “the imaginary beings who emerge from – exist in books.”¹¹⁰ To Thomas Jones, Klara is therefore

106 Turing, “Computing Machinery.”

107 Thomas Jones, “Oh you darling robot!” reviewing Kazuo Ishiguro, *Klara and the Sun*, *London Review of Books* (March 18, 2021): 33–35, 33.

108 Jeanne Gaakeer, “Ishiguro’s Legal Chimera: *Never Let Me Go* and the Legal Fiction of Personhood,” *Polemos*, 2 (2007): 119–132.

109 Italo Calvino, “Two Interviews on Science and Literature,” in *The Uses of Literature*, ed. Italo Calvino (New York: Harvest, 1987): 28–38.

110 Marina Warner, “Men are just boys,” *London Review of Books* (May 6, 2021): 41, reviewing Sally Bayley, *No Boys Play Here*.

a much older form of artificial intelligence, a much older kind of artificial friend: she's a fictional character. It may be inherently impossible to write a novel that openly poses such questions as whether robots can be said to have souls, or to be conscious, or capable of feeling love, or of inspiring and reciprocating sympathy in people. By making them characters in a story, you aren't asking the question: you've already answered it. Klara thinks in ways similar enough to us, is human enough, to function as a conventional character of a novel.¹¹¹

The examples of ADS, AWS and Klara urge us, rather than being the passive recipients of technology, to leave behind any *horror materiae*, a fear of the materiality of things, because it can only lead to a philosophy of alienation. The fictional picture of a form of AI such as Klara can help us probe real AI. We should therefore embrace what the humanities have to offer by way of critical, hermeneutical methods and remain the authors of our technological world of non-human behaviour.

Bionote

Jeanne Gaakeer

Erasmus University Rotterdam, Rotterdam, Netherlands

gaakeer@law.eur.nl

Jeanne Gaakeer is Professor of Jurisprudence: Hermeneutical and Narrative Foundations at Erasmus School of Law, Rotterdam, the Netherlands. She also serves as a senior justice in the criminal law section of the Court of Appeal in The Hague. The focus of her academic research is on the relevance of theoretical notions developed within *Law and Literature* and/or *Law and the Humanities* to legal practice. Currently she is working on topics pertaining to legal narratology. With Greta Olson (Giessen University), she co-founded the European Network for Law and Literature (www.eurnll.org). Her recent publications in English include *Judging from Experience. Law, Praxis, Humanities*, Edinburgh, Edinburgh University Press, 2019; 'Interdisciplinarity 3.0: The Hub of the Universe or Fantasy Island?', *Pólemos*, vol. 14, issue 2, 2020, pp. 319–336; 'Judicial Narration as Explanation of Facts and Circumstances,' in: Frode Helmich Pedersen, Espen Ingebrigtsen and Werner Gephart (eds), *Narratives in the Criminal Process*, Frankfurt am Main, Vittorio Klostermann, 2021, pp. 29–61; 'Law and Literature,' in: Mortimer Sellers and Stephan Kirste (eds), *Encyclopedia of the Philosophy of Law and Social Philosophy, Living Edition*, Meteor.Springer.com, DOI link: https://doi.org/10.1007/978-94-007-6730-0_989-1.

¹¹¹ Jones, "Oh you darling robot!," 33.