Attack of the Algorithm: The Demand for a new Hermeneutics Towards Autonomous Systems

Aren Tyr

Siam Technology College

Author Note

Submission for module 901-121 Ethics and Life (October 15th-19th 2018)

BA TESOL degree.

Submission date: 30th October 2018.

Abstract

This paper presents a brief survey of some of the contemporary ethical issues surrounding the ever more pervasive presence of autonomous systems and the algorithms running them, using self-driving vehicles and autonomous weapon systems as key studies. The issue of codifying ‘morality’ is considered, concluding with a phenomenological aside on the affect of such technology drawing on Heidegger.

Attack of the Algorithm: The Demand for a new Hermeneutics Towards Autonomous Systems

The incursion of algorithms into daily life accelerates: this paper is an attempt to provide a very brief survey of the numerous severe ethical dilemmas that the increasing autonomy of computational systems, artificial intelligence, and advances in machine learning and robotics bring. To simplify matters, I shall here focus on the ethical issues that the underlying algorithmic 'intelligence' that provides the operating logic of such systems presents; any such concerns over the precise nature of implementation and any specifics related to hardware/software architecture shall be bypassed. Instead, I am simply interested in some of the potential ethical concerns that such technologies will bring, and indeed have already brought. After a brief working definition of 'algorithm', I shall then introduce some of the ethically negative affects such automation brings. Next, I shall consider two key 'real world' technologies, self-driving vehicles and autonomous weapon systems, both of which serve to illustrate the serious ethical issues involved in concrete life-and-death scenarios. Finally, I shall briefly consider some of the enormous difficulties involved in attempting to 'codify' morality, before concluding with short phenomenological reflection drawing on the work of Heidegger concerning his observations on technology in general.

The exact definition of an *algorithm* is itself contentious. For the purposes of this paper we shall satisfy ourselves with the following very general definition: a finite procedure, based on a bounded sequence of instructions following an underlying mathematical/logical construct, implemented using software running on whatever contingent hardware is required, for the explicit purpose of solving *a particular task*. A complex machine/information system typically implements a very large number of algorithms to accomplish its objectives, and it is with the ethical implications of these resultant composite 'intelligent systems' (in whatever guise) that I am principally concerned (Kraemer, van Overveld, & Peterson, 2011; Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016; Rapaport, 2005).

If 'intelligent' or 'autonomous' systems are becoming increasingly ubiquitous, what problems do they introduce? After all, predictive or machine learning systems can help us

to shop better or see what matters to us on social media[1], and automate otherwise tedious menial tasks (Daffodil Software, 2017). The old adage of 'with power comes responsibility' is applicable here; as much as algorithms increasingly streamline our lives and offer cognitive 'outsourcing' (Danaher, 2018), so too are they vulnerable to human bias and prejudice. In the case of machine learning, the underlying neural networks are dependent on the quality of the dataset presented to facilitate their training; if this initial dataset is prejudicial, it follows that the resultant trained system is likely to encapsulate and reify those particular dispositions (Cramer, 2016; Mittelstadt et al., 2016). Furthermore, when 'profiling' algorithms are applied in the real world, "the individual is comprehended based on connections with others identified by the algorithm, rather than *actual* behaviour"[2] (Mittelstadt et al., 2016). The use of such profiling and associative algorithms is already prevalent and widespread, and is only certain to become increasingly extensive due to their profitability and perceived effectiveness (Brynjolfsson & McAfee, 2016; Buranyi, 2017; Ford, 2016; O'Neill, 2018).

It is perhaps tempting to ascribe such algorithmic discrimination and their resultant ethical difficulties as a relatively abstract conceptual problem, still somewhat ephemeral, incidental in nature, and largely pertaining to 'first world concerns'. All such notions immediately dissolve when we consider the distinctly real life-and-death implications of self-driving autonomous vehicles, and more starkly, autonomous weapons of war.

Here the demand for a workable implementation of machine ethics is not merely desirable and adventitious, but fundamental. Self-driving cars[3] have been suggested as a real-world case of the famous 'trolley problem', notwithstanding the limitations of this philosophical thought experiment when applied to the decision-theoretic structures that autonomous vehicle design demands (Anonymous, 2018; Broks, 2014; Nyholm & Smids, 2016). A vehicle travelling at speed is a potential lethal instrument to passengers,

[1] Witness the success of Amazon and Facebook, for example.

[2] This is unfortunately exemplified *par excellence* in the discriminatory criminal recidivism model used as an adjunct in the US justice system, together with related questionnaires that are processed through the penal system; see (O'Neill, 2018).

[3] Tesla have just introduced another software update that further increases the autonomous driving capabilities of their cars, which right now are present on the road (Wong, 2018).

pedestrians, and other vehicles, so by necessity, the anticipatory collision avoidance systems imply an acute level of ethical oversight. Should the vehicle deliberately swerve, thereby killing a single pedestrian, if so doing would avoid a fatal head-on collision resulting in the deaths of multiple passengers? The programming requires that a decision-structure specification exists for evaluating collision avoidance actions, and crucially, the particular implementation of that algorithm by the developer *unavoidably implies a set of value-laden choices, whether implicit or explicit* (Cramer, 2016; Kraemer et al., 2011). Absent of specifically assigned moral 'weightings', a trained machine learning system still *ultimately reflects whatever biases and values are implicitly present in the training dataset used, the input used for rule formation, or the particular preferential heuristics implemented* (Mittelstadt et al., 2016; Nadin, 2018). In short then, ethical accountability is inseparable from the development of the system, irrespective of whether it is a 'second order' function.

It is clear that the ethical ramifications are exponentially amplified when we consider the case of autonomous weapon systems (AWS) and unmanned aerial vehicles (UAVs). AWS, UAVs, and other automated military hardware are already an existent reality, rather than a hypothetical construct, though undoubtedly their true technological capabilities remain largely classified from the public eye. Nevertheless, the development rate in terms of their destructive utility arguably far outpaces our ability to formulate and delegate appropriate ethical frameworks for such technologies (Borenstein, 2008; Sparrow, 2007). For systems whose express purpose is the destruction of human life, the necessary ethical demands become increasingly critical. Specifically, such AWS are not merely 'remote controlled' systems utilising a human operator; instead such systems are able to operate with 'battlefield intelligence' independent of human input. Given the critical advantage time savings even in the milliseconds can provide (viz. financial trading; see (Ford, 2016)), it seems predestined that AWS will inexorably develop toward increasing autonomous operational capability in order to survive/eliminate enemy AWS during combat; human reaction speed would become an unworkable performance limitation (Sparrow, 2007). Such systems therefore demand the implementation of vastly complex ethical directives, not only in order to avoid grievous war crimes, but to be able to differentiate friend from foe and evaluate situations where the information is inordinately

'fuzzy' and circumspect – a situation that may still prove too complex for AWS to be fully viable (Borenstein, 2008).

Potentially discriminative predictive algorithms; potentially deleterious preference shaping algorithms; autonomous weapon systems and self-driving vehicles: all represent real world 'litmus' cases where a demand for astute ethical programming is tantamount. Yet just how achievable and plausible is this goal? On the one hand, the notion of explicitly programming a system of 'meta-ethical' reasoning (i.e. able to self-evaluate between differing moral choices) or even a more limited system of explicit moral 'rules' seems inordinately difficult, given that it requires a precise formal specification of whatever moral methodology you intend to implement (Bello & Bringsjord, 2013; Lokhorst, 2011; Purves, Jenkins, & Strawser, 2015). Accomplishing such a codification is philosophically distant, if not unattainable: "As philosophers, we clearly lack widely accepted solutions to issues regarding the existence of free will, the nature of persons and firm conditions on moral agency/patienthood; all of which are indispensable concepts to be deployed by any machine able to make moral judgements." (Bello & Bringsjord, 2013). Certainly no concrete formal system in the vein of a set of Kantian 'categorical imperatives' exists (Ess & Thorseth, 2008; Tallman, 2016). If the anti-codifiability thesis is sustained, it precludes explicitly articulating any such 'moral system' due to its impossibility *a priori* (irrespective of the corresponding insurmountable complexity[4]) — a direct blow to 'machine theologists' (Nadin, 2018).

A potentially more feasible option is to use a 'deep learning' approach as exemplified by systems such as Alpha Go and IBM Watson, thereby allowing the system to inductively generate a hierarchical rule set accumulated by induction given sufficiently expansive training data (DeepMind Technologies Ltd., n.d.; Engadget, 2011; IBM, 2017; Silver et al., 2017). The system then progressively 'self-learns' over time, augmented by appropriate remedial input by expert human developers. This system has proven effective given the extraordinary difficulty in deriving in advance a static decision-theoretic explicit

[4] Meanwhile, a limited set of apparently simple high-level directives soon leads to extraordinarily complex results, as Asimov successfully explored in his robot fiction with his 'three laws' ('Three Laws of Robotics', 2018), or as the beautiful images that can result from fractal mathematics neatly illustrates ('Fractal', 2018).

set of rules that result in successful outcomes for any particular dynamic task/game situation. Even the comparative simplicity of a game like Go illustrates this – let alone implementing 'morality' – and real world life could certainly be viewed as a type of hugely complex 'game' ('Game theory', 2018; Mittelstadt et al., 2016). More pragmatically, a hybrid approach could be adopted, combining limited rule-directed behaviour — instantiated deductively using moral 'exemplars' and probabilistic proscriptions/prescriptions — coupled with deep-learning induction (Bello & Bringsjord, 2013).

Exorbitant ethical difficulties remain. In particular, concerns over accountability and transparency would need addressing. Systems that exist today are already exceptionally opaque in their mechanisms of operation, and this problem seems guaranteed to dramatically escalate as architectures that combine multiple such algorithms and entire networks of systems form irreducibly complex inference structures that exceed the capability of any human analyst to understand or meaningfully interpret them (Miller, 2018). Such indecipherability effectively delegates all responsibility and oversight to the machines themselves; this would be as controversial as it would be morally suspect.

Meanwhile, we have not thus far considered sentience, or self-consciousness, in order for the prospective system to have any capacity to act meaningfully as a *moral agent* (Nath & Sahu, 2017; Rapaport, 2005; Sparrow, 2007). A detailed discussion of this far exceeds the scope of this paper, but suffice to say that if Searle's 'Chinese Room' argument is sustained[5], then the *intentionality* of the system is a prerequisite for its possibility to act as an independent moral agent. It is still unclear how such a phenomenological basis or 'ground state' could be achieved by a machine (though there are suggestions; see (Bello & Bringsjord, 2013; Whobrey & Searle, 2001), or indeed even whether so called 'machine intelligence' even displays any actual 'intelligence' as opposed to mechanistic/deterministic brute searches and heuristic rule-reinforcement (Nadin, 2018).

[5] See (Boyles, 2012; 'Chinese room', 2018). Searle's thought experiment essentially concerns how an agent, provided with appropriate symbolic translation rules/tables, could translate Chinese into English without actually having any real innate or intrinsic understanding/intelligence concerning Chinese; i.e. the agent simply proceeds via a 'dumb' algorithmic method, rather like a human version of Google Translate.

Prudential and pragmatic reality supervenes upon such philosophical predicaments, however, since *current* existent autonomous systems and the algorithms they implement already exhibit startling real-world effects that have significant ethical implications for millions of individuals (Brynjolfsson & McAfee, 2016; Buranyi, 2017, 2017; Ford, 2016; Legg & Hutter, 2007; O'Neill, 2018). We are therefore obligated to make an *attempt* at implementing some type of moral decision processing, together with suitable regulatory frameworks (Umbrello, Torres, & Bellis, 2018), since the pace of technological change tends to causes irreversible systemic social effects (and indeed personal phenomenological effects) far before we are even fully cognisant of them. This phenomenon is associated with any paradigm shifting (indeed *zeitgeist* creating) technological development throughout human history ('Paradigm shift', 2018). We therefore need to try to anticipate future ethical dilemmas related to automation.

Critics may note that such technologies as already exist have so far generally avoided difficulty — take autonomous vehicles, for example. Despite the mileage Google's self-driving cars have accumulated (Bhuiyan, 2016), they have so far operated with human backup at the wheel, and they represent only a microscopic fraction of the total traffic on the road. Hitherto, they have therefore been insulated by statistical likelihood. If widespread adoption of the technology leads to it becoming commonplace, it seems inevitable that at some point the ethical issues discussed in the abstract here become real world legal battles in the courtroom, involving human lives, with distinct consequences. In short, we are only at the very *beginning* of this thorny path to be negotiated in the world of algorithms and ethical accountability.

Finishing on a poetic note, perhaps it could be suggested that the age of autonomous algorithms and intelligence heralds a great darkening of human experience as much as it proffers a magnificent technological utopia:

> This is precisely what Heidegger saw happening to 'meditative thought' and essential language through the influence of Western metaphysics, which by means of imposing a technical form of 'logic' and 'grammar', determined the ways in which we think and speak about Being. When

> ‘meditative thought’ is imprisoned within a technological framework, or interpretation, when a scientific standard of verification is applied to authentic philosophical thought, 'Being, as the [essential] element of thinking, is abandoned by the technical interpretation' (Magrini, 2012).

And what are autonomous machine agents, if not those whose very mode of ‘thinking’ or ‘intelligence’ is purely that which is calculable according to a *scientific standard of verification* — i.e. an algorithmically determined output? Contra *deus ex machina*, perhaps it is now the machine that shall inveigh upon the proverbial ghost, as we absolve ourselves of higher moral responsibility by outsourcing all our difficult decisions to autonomous systems based on a deterministic reduction to probabilistic rule-inferences, and signal the death to any Kantian deontological aspiration. Will our innate moral capacity atrophy, by slavish adherence to quantitative machine learning algorithms based on supposedly ‘objective’ data and ‘neutral’ values?

A useable hermeneutics of algorithms is vital. We are *in* the machine age, and an ethics of algorithms is *necessary*, even if it is not and indeed perhaps cannot ever be *sufficient*.

References

Anonymous. (2018). Trolley problem. In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Trolley_problem&oldid=865986340

Bello, P., & Bringsjord, S. (2013). On How to Build a Moral Machine. *Topoi*, *32*(2), 251–266. https://doi.org/10.1007/s11245-012-9129-8

Bhuiyan, J. (2016, October 5). After two million miles, Google's robot car now drives better than a 16-year-old. Retrieved 30 October 2018, from https://www.recode.net/2016/10/5/13167364/google-self-driving-cars-2-million-miles

Borenstein, J. (2008). The Ethics of Autonomous Military Robots. *Studies in Ethics, Law, and Technology*, *2*(1). https://doi.org/10.2202/1941-6008.1036

Broks, P. (2014, November 14). The Trolley Problem, Neuropsychologist Paul Broks on Morality and the Brain, A History of Ideas. Retrieved 28 October 2018, from https://www.bbc.co.uk/programmes/p02bx2hh

Brynjolfsson, E., & McAfee, A. (2016). *The second machine age: work, progress, and prosperity in a time of brilliant technologies.*

Buranyi, S. (2017, August 8). Rise of the racist robots – how AI is learning all our worst impulses. *The Guardian*. Retrieved from https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses

Cramer, F. (2016). Hermeneutics and analytics. Retrieved 23 October 2018, from http://cramer.pleintekst.nl/essays/crapularity_hermeneutics/

Daffodil Software. (2017, July 31). 9 Applications of Machine Learning from Day-to-Day Life. Retrieved 29 October 2018, from https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Danaher, J. (2018). Toward an Ethics of AI Assistants: an Initial Framework. *Philosophy & Technology*. https://doi.org/10.1007/s13347-018-0317-3

DeepMind Technologies Ltd. (n.d.). AlphaGo. Retrieved 23 October 2018, from https://deepmind.com/research/alphago/

Engadget. (2011, January 13). IBM's Watson Supercomputer Destroys Humans in Jeopardy | Engadget. Retrieved 29 October 2018, from https://www.youtube.com/watch?v=WFR3lOm_xhE

Ess, C., & Thorseth, M. (2008). Kant and information ethics. *Ethics and Information Technology*, *10*(4), 205–211. https://doi.org/10.1007/s10676-008-9158-6

Ford, M. (2016). *Rise of the robots: technology and the threat of a jobless future.*

Game theory. (2018). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Game_theory&oldid=865691698

IBM. (2017, October 15). IBM Watson. Retrieved 29 October 2018, from https://www.ibm.com/watson/

Kraemer, F., van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, *13*(3), 251–260. https://doi.org/10.1007/s10676-010-9233-7

Legg, S., & Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, *17*(4), 391–444. https://doi.org/10.1007/s11023-007-9079-x

Lokhorst, G.-J. C. (2011). Computational Meta-Ethics: Towards the Meta-Ethical Robot. *Minds and Machines*, *21*(2), 261–274. https://doi.org/10.1007/s11023-011-9229-z

Magrini, J. M. (2012). Worlds Apart in the Curriculum: Heidegger, technology, and the poietic attunement of literature. *Educational Philosophy and Theory*, *44*(5), 500–521. https://doi.org/10.1111/j.1469-5812.2010.00718.x

Miller, C. (2018, August 21). The terrifying, hidden reality of Ridiculously Complicated Algorithms. Retrieved 30 October 2018, from https://www.the-tls.co.uk/articles/public/ridiculously-complicated-algorithms/

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Nadin, M. (2018). Machine intelligence: a chimera. *AI & SOCIETY*. https://doi.org/10.1007/s00146-018-0842-8

Nath, R., & Sahu, V. (2017). The problem of machine ethics in artificial intelligence. *AI & SOCIETY*. https://doi.org/10.1007/s00146-017-0768-6

Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

O'Neill, C. (2018). *Weapons Of Math Destruction*. Penguin Books. Retrieved from http://www.worldcat.org/title/weapons-of-math-destruction-how-big-data-increases-inequality-and-threatens-democracy/oclc/1039545320&referer=brief_results

Paradigm shift. (2018). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Paradigm_shift&oldid=862989949

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethical Theory and Moral Practice*, *18*(4), 851–872. https://doi.org/10.1007/s10677-015-9563-y

Rapaport, W. J. (2005). Philosophy of Computer Science: An Introductory Course. *Teaching Philosophy*, *28*(4), 319–341. https://doi.org/10.5840/teachphil200528443

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Tallman, R. (2016, November 14). Kant & Categorical Imperatives: Crash Course Philosophy #35. Retrieved 27 October 2018, from https://www.youtube.com/watch?v=8bIys6JoEDw

Umbrello, S., Torres, P., & Bellis, A. F. D. (2018). The Future of War: Could Lethal Autonomous Weapons Make Conflict More Ethical? *Preprint under Review*, 14.

Whobrey, D., & Searle, J. (2001). Machine Mentality and the Nature of the Ground Relation, 40.