Martijn van Otterlo
Department of Cognitive Science and Artificial Intelligence
Tilburg University
m.vanotterlo@uvt.nl

Ethics and the value(s) of Artificial Intelligence

In this article Martijn van Otterlo provides an introduction to some current issues in the ethics of intelligent algorithms.

> "Solving the value-loading problem is a research challenge worthy of some of the next generation's best mathematical talent" [6, p. 229]

> "Supercharged with a larger cerebral cortex, faster learning, and a longer time horizon, is it possible that we solve complex problems in mathematics the same way that monkeys find optimal paths?" [29, p. 20152]

Artificial Intelligence (AI) [25] is booming. Although it has existed for more than six decades, recently it is literally everywhere. Phones have 'AI inside', computer games utilize 'AI opponents' and internet services let AI analyze posts and photos, personalize news feeds and find who or what you need. Each day news articles about AI appear, increasingly so since 2010 [11]. These advances have often profound consequences for our daily lives. For example, Google search fundamentally changed the way we obtain information [41] and Facebook's face recognition changed our personal privacy forever. Much of the current progress comes from the subfield of *machine learning* (ML) [19] and more specifically from a technique called *deep learning* (DL) [16] which has its roots in earlier *neural networks*. ML makes use of *data* to *inductively generate predictive models*. For example, based on a huge dataset of images, computers can nowadays be *trained* to recognize various items on pictures, and companies such as Facebook are heavily investing in such technology.

Anticipations of AI: good and bad

It is hard to predict where things are heading with all this progress in AI. Terry Sejnowski wrote in 2010 that *reinforcement learning* (RL) [46], a special kind of ML, had just beaten a 5*th dan* human professional in the board game *Go*, using similar techniques used to learn *backgammon* much earlier [35]. AI experts predicted that beating the human overall Go champion would take a decade at least when suddenly in 2016 the DL *AlphaGo* program [31] did just that after first beating humans at *Atari* games. But whereas AlphaGo learned from lots of human moves, its successor *AlphaZero* [32] returned to ideas that solved backgammon, and taught *itself* by just playing against incrementally better versions of itself and beat everyone, including AlphaGo, 100 against 0. Interestingly, the same ML method also dominates (human and machine) in *chess* after only 4 hours of training, from scratch, by itself, only 20 years after the heavily engineered IBM's *DeepBlue* algorithm beat the human champion Gary Kasparov.

Predicting developments is hard [14], but it is important to anticipate what can happen if AI achieves general *human-level intelligence*. Half of the experts in a 2013 panel predicted it to happen between 2040–2050, and most experts predicted *superintelligence* [6] would follow 30 years later, but 30 percent did not see this as positive for humanity [24]. Another study predicted human-level skills such as *translation* (by 2024) and *surgery* (by 2053) [17]. Negative consequences of AI, ranging from loss of jobs, to *malicious* use [48], and to superintelligence dominating humanity, have appeared in mainstream media too [11]. Until recently, criticism of AI (and especially ML) technology often came from the social sciences and legal scholars and was mainly about *data-related* concerns such as *privacy*, *surveillance* and *discrimination* [40]. More recently the broader implications of intelligent algorithms on society are studied [23, 44] and more importantly, by scholars from AI and ML itself. Recent reports explicitly take into account the *ethical* dimensions of AI for society [28, 34]. Recent books join: Tegmark's [37] (near and far AI future), Walsh's [45] on the status of AI, and Shanahan's overview [30] of the *singularity*.

Ethics of intelligent algorithms

Powerful ML algorithms can have profound influence in our digital society. Consider a case where a social network would approach users with a request to 'go nice' on a particular friend because it has statistically predicted that there is a higher than average probability that this friend has suicidal tendencies. This may sound *creepy* [38], but it is one of Facebook's recent plans: to predict potential suicides [50]. In a related effort, Google wants to detect depression [51]. Such predictions are technically interesting if they are possible, and possibly an opportunity to 'do good', but at the same time they *create* ethical issues: can a social network disclose or use such predictions without asking, and would that change people's behavior (of that friend, and towards that friend)? Even more ethically challenging, similar predictive capabilities can also be used to target insecure and troubled teenagers for *marketing purposes* [52]. Most people would not find that ethically correct, but some may see this as normal market practices. Other situations where powerful ML algorithms are used with ethical consequences are *search engines*, *personalized news feeds*, and *curation* on the internet. Examples include Facebook fighting terrorism [53], Google battling fake news [54] and Twitter's moderation [55]. All these technologies solve a problem: *information overload*. Without filtering and curating, humans could not handle the enormous amount of information. However, the ethical issue here is that these algorithms *select*, *hide* (e.g. the removal of the iconic 'napalm girl' photo due to Facebook's anti-nudity policy [56]), and *prioritize* sources for us. They basically decide what we get to see, and what not [41], which can be 'for good' again, but may equally well be considered *censorship*.

The *algorithmization* of society brings us many novel ethical issues. The study of societal consequences of AI beyond simple privacy and surveillance has become known as *ethics of algorithms* [23,42,44]. Sometimes *law* can be the answer, but so often it is far too slow to adapt [38] and we need to consider general ethical analysis. Algorithms basically transform *data* into *evidence*, which then is used to compute *actions*. Evidence can be inconclusive, inscrutable or misguided and this can cause many ethical consequences of actions, relating to *fairness*, *opacity*, *unjustified actions*, and *discrimination*. In the Facebook suicide case, evidence based on only Facebook data may be inconclusive and actions to act upon suspected suicidal tendencies may be unjustified. Overall, algorithms have an impact on *privacy* and can have *transformative effects* on *autonomy*. For example, the simple fact that Google selects our news may change how we think about particular subjects, and affect our autonomy to make decisions and trap us into *filter bubbles* [20]. It is useful to distinguish several core *classes* of algorithms I identified elsewhere [44]:

1. inference algorithms (*descriptive*),
2. learning algorithms (*predictive*),
3. optimization algorithms (*prescriptive*),
4. *superintelligence* [6,30,37].

A fifth class consists of *'physical'* algorithms such as *internet-of-things* and *robots*. Physicality will create another level of ethical problems [33], such as social backlash with surveillance robots [67] in public spaces or privacy issues with 'connected toys' [68]. Finding the right metaphor for how robots relate to non-physical *algorithms* may also help [27].

As said earlier, AI as a field is becoming aware of the issues itself. Novel ethics research centers arise [57], companies coordinate efforts [58], and scientists [59] and employees [60] speak out. In addition, education becomes aware [7,47]. AI sub-communities have started to openly discuss the issues and identify *challenges* and *design principles*. The engineering community started the IEEE *Ethically Aligned Design* initiative [64], aimed at developing a *vision for prioritizing human well-being with autonomous and intelligent systems*, to capture concepts like *transparency* and *responsibility*, and to develop industry standards for *ethics*. The robotics community came with their own EPSRC *design principles* for robotics systems [66]. In January 2017, the AI community came up with their own *Asilomar principles* [65], a *code of ethics* which explicitly states 23 ethical principles for AI developers. The list includes general goals such as that AI should not target the creation of *undirected* intelligence, but instead should develop *beneficial* intelligence (principle 1). It also contains statements about *judicial* (principle 8) and *failure* (principle 7) transparency: an AI system should be able to *explain* its decisions.

Important *values* desirable in complex AI systems include *transparency* and *explainability*, to avoid *opacity* in decision making. Equally important are *responsibility*, *liability* and *accountability* [9] ("Who is to blame if an algorithm does harm?"). This should lead to *safe* and *trustworthy* AI systems. All these efforts should lead to professional *codes of ethics* for the development and employment of AI [5]. As I illustrated elsewhere [43], there is a strong parallel between these values (and intentions) and those behind *human* codes of ethics for various professions: they are used to openly and transparently communicate to the outside world what are the norms and values in a particular profession, and by doing that to earn trust and acceptance from outside. For increasingly intelligent AI, it is vital that not just the humans comply with the code, but the AIs too. For the latter, one literally obtains a *code* of ethics, embedded in the AI's program.


Ethical choices for an autonomous car in case of an upcoming accident

Towards building ethical AI systems

These *ethical* dimensions require AI designers to act *responsibly*. Some hope for a 'big red button' [49] to shut down 'rogue' AI systems, but that seems naive [4]. A better idea is to incorporate ethical thinking in the design of AI systems, possibly with the use of AI technology itself. For example, if a profiling algorithm is discriminatory, one can modify it such that inherent biases are changed, or if autonomous cars can crash they should learn how to do it 'least harmfully'. Building ethical values into AI requires two things [6]:

1. a capacity to *acquire* ethical values, possibly from humans,
2. knowledge of *which human values* are important.

The field of AI is heavily invested lately in working on these two main issues. Some employ so-called *ethics bots* [10] to assist humans, while others focus on obtaining *human* ethical values through massive experiments [71]. In the following paragraphs, I briefly mention four current research directions.

Fairness in machine learning
Many successful AI systems are based on ML, of which predictions can be biased by the data, parameters and application processes. A well-studied case concerns *recidivism risk score predictions* in the American judicial system [69], which sparked a lot of debate on what fairness actually is. In addition, it also showed that inherent biases in such decision making systems can have profound legal consequences. In relation to that, the notion of fairness has many connections to other concepts such as *diversity* and *discrimination*, which also makes it a highly interdisciplinary topic. Fairness is a *multi-objective problem* and intuitively, but also formally, algorithms that are fair on all possible accounts are impossible. Many different interpretations of *fairness* are being studied and much effort is being put into *fairness enhancing* techniques, to remove some of the unwanted biases in predictive algorithms [12]. A strong community has risen under the name of *fairness, accountability, and transparency in ML* (FAT) [62].

Explaining black box systems
On top of fairness and bias-related issues comes the fact that most powerful decision-making AI systems have become too complex to be *understandable* by humans, even though many of their decisions affect people's lives in various ways. Especially for deep learning systems there is a strong trade-off between *accuracy*, which contributes to the success of those models, and *interpretability*, which is hindered much by their complex, *black box* nature [8]. A current trend in ML is to investigate *interpretability* as a concept, and models that are (more) *interpretable* [70], sometimes capable of explaining the decisions of classifiers in human-understandable terms [26]. Much work is also being done on *extracting* (interpretable) knowledge *from* trained deep networks, for example by *distilling* [13], or other techniques that increase *transparency* and *explanation* [18]. Such approaches fit in a revived interest in so-called *explainable* AI [61].


AI and the good, the bad, science and the law

Value-based AI
Much of the previous waves of attention to ethical issues dealt with a more limited set of aspects such as *privacy* and *surveillance* consequences of ML [40]. Currently, the focus is on the much broader notion of *value alignment* (VA): autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation. Obviously, VA is a *multi-objective optimization* problem too [39], and does view AI as an autonomous *agent* which takes *actions* and optimizes its behavior according to a *utility function*. Such settings are typical for the AI subfield of *reinforcement learning* (RL) [46]. RL is an ideal model for computational ethics [1], and many current VA issues studied in AI come from this area [2, 36]. Examples are

1. *scalable oversight* of ML systems by humans,
2. *mild optimization*, or 'not optimizing too hard',
3. *learning from humans*,
4. *safe exploration*.

RL is often used in combination with deep learning [21] (and because of that connects to all previously mentioned issues) but it can also be seen as an ideal suite of algorithms just because it is aimed at *value-based* optimization.

Ethical reasoning
The fourth direction towards ethical AI involves *reasoning*. As said, much of the success of current AI comes from *learning* approaches, but now (and in the past) these are being criticized [22] for their lack of explainability, and their incapability to insert and extract domain knowledge. Domain theories, models and formal logic, typically computer science tools, have been shown to be effective in the validation and verification of systems, and could be used to ensure their proper functioning [28]. It is only natural to consider logical models in the context of value alignment too, since they directly support *declarative*, *explainable* and *verifiable* reasoning of systems. AI has a rich tradition of such models, including those targeting *reasoning about ethical aspects* [3]. In recent work I introduced a novel approach combining formal logic, *decision-theoretic optimization*, and *supervised machine learning* for transparent (declarative), ethical reasoning in the face of uncertainty [43]. It is foreseeable that more such systems will follow to combine learning and reasoning about ethics in an explicit, and transparent way.

This text, but also the current state of the art, has only scratched the surface of what is needed to build truly ethical AI. The quotes at the beginning of this text show that this requires (mathematical!) advances to *build* such intelligent systems, but also to obtain the means to endow such systems with *our* (human) values.

References

1 D. Abel, J. MacGlashan and M.L. Littman, Reinforcement Learning as a Framework for Ethical Decision Making, in *AAAI Workshop: AI, Ethics, and Society*, 2016.

2 D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, Concrete problems in AI safety, *CoRR*, abs/1606.06565, 2016.

3 M. Anderson and S.L. Anderson, Machine ethics: Creating an ethical intelligent agent, *AI Magazine* 28 (2007), 15–26.

4 T. Arnold and M. Scheutz, The 'big red button' is too late: an alternative model for the ethical evaluation of AI systems, *Ethics and Information Technology* 20(1) (2018), 59–69.

5 P. Boddington, *Towards a Code of Ethics for AI*, Springer, 2017.

6 N. Bostrom, *Superintelligence*, Oxford University Press, 2014.

7 E. Burton, J. Goldsmith, S. Koenig, B. Kuipers, N. Mattei and T. Walsh, Ethical considerations in AI courses, arXiv:1701.07769, 2017.

8 D. Castelvecchi, Can we open the black box of AI? *Nature News* 538(7623) (2016), 20.

9 N. Diakopoulos, Accountability in algorithmic decision making, *Communications of the ACM* 59(2) (2016), 56–62.

10 A. Etzioni and O. Etzioni, Designing AI systems that obey our laws and values, *Communications of the ACM* 59(9) (2016), 29–31.

11 E. Fast and E. Horvitz, Long-term trends in the public perception of AI, in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.

12 S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E.P. Hamilton and D. Roth, A comparative study of fairness-enhancing interventions in machine learning, arXiv:1802.04422, 2018.

13 N. Frosst and G.E. Hinton, Distilling a Neural Network Into a Soft Decision Tree, *CoRR*, abs/1711.09784, 2017.

14 M. Gini, N. Agmon, F. Giunchiglia, S. Koenig and K. Leyton-Brown, AI in 2027, *AI Matters* 4(1) (2018), 10–20.

15 N. Goodall, Ethical decision making during automated vehicle crashes, *Transportation Research Record: Journal of the Transportation Research Board* 2424 (2014), 58–65.

16 I. Goodfellow, Y. Bengio and A. Courville *Deep learning*, Vol. 1, MIT Press, 2016.

17 K. Grace, J. Salvatier, A. Dafoe, B. Zhang and O. Evans, When will AI exceed human performance? Evidence from AI experts, arXiv:1705.08807, 2017.

18 R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar and K. Sycara, Transparency and explanation in deep reinforcement learning neural networks, in *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

19 M.I. Jordan and T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science* 349(6245) (2015), 255–260.

20 D. Lazer, The rise of the social algorithm, *Science* 348(6239) (2015), 1090.

21 Y. Li, Deep reinforcement learning: An overview, arXiv:1701.07274, 2017.

22 G. Marcus, Deep learning: A critical appraisal, arXiv:1801.00631, 2018.

23 B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter and L. Floridi, The ethics of algorithms: Mapping the debate, *Big Data & Society* 3(2), 2016.

24 V.C. Müller and N. Bostrom, Future progress in AI: A survey of expert opinion, in V.C. Müller, ed., *Fundamental Issues of AI*, Springer, 2016, pp. 555–572.

25 N.J. Nilsson, *The Quest for Artificial Intelligence*, CUP, 2009.

26 M.T. Ribeiro, S. Singh and C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in *Pr. ACM SIGKDD*, 2016, pp. 1135–1144.

27 N.M. Richards and W.D. Smart, How should the law think about robots?, in R. Calo, A.M. Froomkin and I. Kerr, eds., *Robot Law*, Edward Elgar Publishing, 2016, pp. 3–22.

28 S. Russell, D. Dewey and M. Tegmark, Research priorities for robust and beneficial AI, *AI Magazine* 36(4) (2015), 105–114.

29 T.J. Sejnowski, Learning optimal strategies in complex environments, *Proceedings of the National Academy of Sciences*, 107(47) (2010), 20151–20152.

30 M. Shanahan, *The Technological Singularity*, MIT Press, 2015.

31 D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, *Nature* 529(7587) (2016), 484–489.

32 D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge. *Nature* 550(7676) (2017), 354.

33 S. Steinert, The five robots – a taxonomy for roboethics, *Int. J. of Social Robotics*, 6(2) (2014), 249–260.

34 P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, et al., *AI and Life in 2030. One Hundred Year Study on AI*, Report of the 2015–2016 Study Panel,Stanford University, http://ai100.stanford.edu/2016-report, 2016.

35 R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 2017, 2nd ed.

36 J. Taylor, E. Yudkowsky, P. LaVictoire and A. Critch, Alignment for advanced machine learning systems, MIRI, 2017 (unpublished) https://tinyurl.com/jemdx4b.

37 M. Tegmark, *Life 3.0. : Being Human in the Age of AI*, Allen Lane, 2017.

38 O. Tene and J. Polonetsky, A theory of creepy: Technology, privacy, and shifting social norms, *Yale Journal of Law and Technology* 16(1), 2014.

39 P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery, Human-aligned AI is a multiobjective problem. *Ethics and Information Technology* 20(1) (2018), 27–40.

40 M. van Otterlo, A machine learning perspective on profiling, in M. Hildebrandt and K. de Vries, eds., *Privacy, Due Process and the Computational Turn*, Routledge, 2013, Chapter 2, pp. 41–64.

41 M. van Otterlo, The libraryness of calculative devices, in L. Amoore and V. Piotukh, eds., *Algorithmic Life: Calculative Devices in the Age of Big Data*, Routledge, 2016, Chapter 2, pp. 35–54.

42 M. van Otterlo, From intended archivists to intentional algivists: Ethical codes for humans and machines in the archives, in F.P. Smit, A. Glaudemans and R. Jonker, eds., *Archives in Liquid Times*, Stichting Archiefpublicaties (S@P), 2017, Chapter 12, pp. 267–293.

43 M. van Otterlo, From algorithmic black boxes to adaptive white boxes: Declarative decision-theoretic ethical programs as codes of ethics, in *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

44 M. van Otterlo, Gatekeeping algorithms with human ethical bias: The ethics of algorithms in archives, libraries and society, arXiv:1801.01705, 2018.

45 T. Walsh, *Android Dreams: The Past, Present and Future of AI*, Hurst Publishers, 2017.

46 M.A. Wiering and M. van Otterlo, eds., *Reinforcement Learning: State-of-the-Art*, Springer, 2012.

47 See my 'Ethics of algorithms' course, https://tinyurl.com/ydanrvpz.

48 *The Malicious Use of AI*, https://maliciousaireport.com.

49 https://tinyurl.com/ya8fsatf.

50 https://tinyurl.com/mvh4y8s.

51 https://tinyurl.com/ya2v6lct.

52 https://tinyurl.com/le8wlyu.

53 https://tinyurl.com/y89s3ztw.

54 https://tinyurl.com/y89s3ztw.

55 https://tinyurl.com/ybjj4ske.

56 https://tinyurl.com/z7xn9sm.

57 https://tinyurl.com/y9bmnuum.

58 https://tinyurl.com/hyvmpwb.

59 https://tinyurl.com/h3mwshw.

60 https://tinyurl.com/y9b2ujhb.

61 https://tinyurl.com/yd4rwfow.

62 https://www.fatml.org.

63 http://www.aies-conference.com.

64 https://ethicsinaction.ieee.org.

65 https://futureoflife.org/ai-principles.

66 https://tinyurl.com/yb8a5w5c.

67 https://tinyurl.com/ybkhdjhm.

68 https://tinyurl.com/ybc5v798.

69 https://tinyurl.com/jzdpyz5.

70 https://tinyurl.com/y9lqksgf.

71 http://moralmachine.mit.edu.