Annotations to the CHAIR Ethical Policy on AI

Olle Häggström

June 2020

The CHAIR Ethical Policy, taken on May 9, 2019, is meant to give ethical guidance for our AI research and other activities at CHAIR. The policy document is, however, far from sufficient for this purpose, because AI ethics is a much too far-ranging topic to be adequately summarized in such a short document. What follows is my attempt at ameliorating this situation by offering some additions, annotations and suggestions for further reading. The body of the text below consists of the official policy, whereas the footnotes and the reference list at the end are all my additions. While I do try to be reasonably even-handed[1] about the various views on and approaches to AI ethics that can be found out there, please note that I am solely responsible for all contents of the footnotes, and that none of it can be taken as official CHAIR policy.

Ethical Policy for CHAIR[2]

This is the Ethical Policy for Chalmers AI Research Centre that was prepared by the AI Ethics Committee and approved by the centre's Steering Group on May 9, 2019.

Artificial intelligence (AI) technologies already have profound impact on the lives of most humans and on society as a whole, and this tendency is likely to increase dramatically in coming years.[3] The impact can be for better or for worse, and it is therefore crucial that ethical considerations are taken heavily into account in the development, implementation, dissemination and application of AI systems. Important ethical considerations that come up in various AI settings include but are not

[1] Yet it is inevitable that I have opinions not universally shared among scholars of AI ethics, and that these opinions influence what I write. For readers who, in order to be able to calibrate for this, look for a text where my particular political and other biases shine through more unabashedly, Häggström (2020) may be a useful reference.

[2] This is obviously not the only AI ethics policy document formulated in recent years. EU's *Ethics Guidelines for Trustworthy AI* (EU Commission, 2019) is the most well-known and perhaps also most influential example, but other policy documents abound; see Jobin et al (2019) and Hagedorff (2020) for overviews. There is no reason to believe that the present policy document would be superior to all the others and that the latter could thus be ignored, so a better approach than focusing solely on the present document is to also study the *Ethics Guidelines for Trustworthy AI* along with other policies and guidelines, in order to encounter a wider range of perspectives. (What one should obviously *not* do, however, is to scan this extensive range of documents in a kind of race to the bottom, searching for the one that has the most lenient requirements and recommendations as regards the ethical aspects most salient in one's own pet project.)

[3] While nobody knows quite what the future has in store for us (partly because we are in a position to influence it) historian Yuval Noah Harari (2018) and computer scientist Stuart Russell (2019) offer insightful, engaging and only partly overlapping ideas about what kinds of scenarios we need to prepare for in the coming century or so, in their respective books *21 Lessons for the 21st Century* and *Human Compatible: Artificial Intelligence and the Problem of Control*. A more near-term and conservative perspective is offered by Floridi (2019), and readers interested in even shorter time scales and tolerant of a bit of corporate jargon may consult the *Forbes* article by Marr (2020). A substantial fraction of the coming decade's growth in global GDP is expected to come out of innovation in AI; see, e.g., PwC (2019). On the other hand, there are serious concerns (as well as much uncertainty) about how its effects on the labor market and economic inequality may turn out. The number one reference here is still Brynjolfsson and McAfee (2014), while for relevant later works see, e.g., Roine (2016), Grace et al (2017), Campa (2018), Danaher (2019) and the above books by Harari and Russell, plus Frey (2019) for an ambitious historical perspective, and Scholl and Hanson (2020) for arguments skeptical about cataclysmic effects on the labor market being just around the corner.

limited to issues of fairness vs bias[4], transparency vs opacity[5], accountability[6], human autonomy[7], privacy[8] and integrity, democratic participation[9], safety[10] and sustainability[11].[12]

With this in mind, the ethics perspective should permeate all research projects and all other activities at Chalmers AI Research Centre (CHAIR), including those that are carried out in collaboration with an external partner. It should do so from the very start of a project, including the planning, proposal and funding application stages. All calls for project proposals within CHAIR will include instructions to explicitly address ethical considerations.[13]

---

[4] When an AI makes or informs decisions about individuals, we obviously want it to be fair. There are many cases where this seems to have failed, including one involving gender discrimination in AI-assisted recruitment at Amazon (Dastin, 2018), and another involving racial discrimination in the software COMPAS for judging defendants' risk of recidivism in criminal courts (Larson et al, 2016; Cossins, 2018). A major problem when attempting to avoid such automated discrimination is that there are several more-or-less plausible but mutually contradictory definitions of algorithmic fairness, and that it is far from obvious which one ought to be applied in a given setting; see, e.g., Verma and Rubin (2018), Barocas et al (2019), and the slides by Johansson (2019) from his talk at the CHAIR AI Ethics Seminar.

[5] There is a large technical literature on transparency vs opacity (along with the related concept of explainability) of AI systems; see, e.g., Samek et al (2019). Zerilli et al (2019) offer an interesting comparison of transparency vs opacity concerns in automated vs human decision-making, while de Fine Licht and de Fine Licht (2020) emphasize the role of transparency in AI-assisted public decision-making.

[6] See, e.g., Cranefield et al (2018) and Nyholm (2020).

[7] Chapter 3 of Harari (2018) offers a beautiful non-technical treatment of how human autonomy may erode in a world with ever-more-capable AI decision-making tools. Related is the idea of *nudging,* which was received with some enthusiasm when it was popularized by Thaler and Sunstein (2008), but when the nudging is carried out by advanced AI the situation may be more problematic; see, e.g., Susser (2019).

[8] The books by Schneier (2015) and Zuboff (2019) both offer powerful treatments of societal aspects of the kind of mass surveillance that is enabled by AI technology. On privacy and mass surveillance, the Chinese experience (Kobie, 2019) is important to keep in mind, as is the question of whether we, in a possible future era with increasingly powerful and widespread biotechnology having potentially destructive capabilities, can do without far-reaching mass surveillance (Torres, 2019). In the more technical AI literature, differential privacy is an influential concept; see, e.g., Abadi et al (2016).

[9] An important aspect of this is how to salvage human participation in an increasingly algorithmic governance; see, e.g., Danaher (2016) and Danaher et al (2017), as well as Harari (2018). See also the aforementioned paper by de Fine Licht and de Fine Licht (2020). On the role of automated news recommenders, see, e.g., Helberger (2019), as well as Chessen (2017) who paints some rather alarming scenarios extrapolating just a few years ahead from today's filter bubbles, deceitful chatbots and info wars.

[10] The literature on safety of AI systems ranges from down-to-earth applications such as autonomous vehicles (e.g., Nascimento et al, 2019) to theoretical work pertaining to a future breakthrough in artificial general intelligence (e.g., Russell, 2019; Shah, 2020). It is sometimes complained that there is a huge disconnect between these two extremes, and while there is some truth in this, Amodei et al (2016) offer a significant attempt to bridge the gap; see also Cave and ÓhÉigeartaigh (2019).

[11] Vinusea et al (2020) survey the AI literature as it pertains to the United Nation's various Sustainable Development Goals. Among the many important aspects at the interface between AI and sustainability is the rapidly increasing amounts of electricity consumed in the training of large machine learning systems; see, e.g., Amodei and Hernandez (2018) and Knight (2020).

[12] This list of topics within AI ethics is not exhaustive, and partly for this reason the reader is strongly advised to consult some broader general treatment of the subject – or preferably more than one, because perspectives vary, as can be seen from the rather different emphases in the book by Dignum (2019) and the encyclopedia entry by Müller (2020), respectively. See also Floridi and Cowls (2019) for an attempt at integrating some of these topics.

[13] We began this routine in 2019, but the results were rather mixed (Häggström, 2019b), as many AI researchers at Chalmers were clearly unprepared for the request to address ethical concerns. I hope the present document will contribute to improving this situation.

While project managers and participants always bear full responsibility for ethical concerns pertaining to the project, CHAIR leadership is nevertheless responsible for fostering an environment that cultivates informed discussion on ethical issues, as well as for supporting only projects that adequately address relevant ethical concerns.[14]

An overarching principle is that AI systems whose risk of causing harm is not clearly outweighed by their beneficial effects should not be built or disseminated. When estimating benefit vs harm, it is not always sufficient to consider the problem from the viewpoints of developers, owners and users of an AI system; in many cases, there is a need to consider also further stakeholders including third parties affected by the use, as well as effects on the environment. This fundamental principle should never be allowed to be overridden by commercial, military or other considerations.[15]

It should be recognized that an action is not automatically ethically justified or ethically permissible just because it is legal. Furthermore, it does not suffice to focus solely on the direct effects of an AI system: it is also necessary to consider its possible indirect or longer-term consequences, such as the risk of contributing to an AI arms race[16] or of being integrated into a system that can spiral out of control[17].

Finally, it should be noted that ethical AI is not only about avoiding unethical actions or harmful consequences, but even more[18] about developing and using AI to bring about good consequences,

---

[14] It is important to understand in this context that responsibility is not a zero-sum game; see, e.g., Verweij and Dawson (2019). The head of a lab can be morally fully responsible for making sure no scientific misconduct is carried out at the lab, while each individual researcher maintains full responsibility for not committing any such misconduct. Relatedly, for technology developers wishing to escape difficult ethical considerations, it is easy and all too common to be seduced (like, e.g., Burrus, 2012) by the idea that there are no good or evil technologies, only good and evil uses of them. It is true that if I develop and publicize some easy-to-scale weaponized autonomous drone technology, and if this technology is used by terrorists to kill people, then these terrorists carry full responsibility for these killings, but it is also true that I then have moral responsibility for having created a situation where terrorists get easy access to weapons of mass destruction. For more in malicious use of AI, see Brundage et al (2018).

[15] The requirement to think through and weigh benefits vs risks of an AI project is extremely demanding, but it pretty much follows if we accept that humans are morally responsible for the consequences of their actions, and that AI researchers are not exempt from this; see Häggström (2018). The task borders on the impossible if taken literally, yet it is not impossible to take meaningful steps in the right direction; see Ashurst et al (2020) for a helpful guideline on how to do this in practice.

[16] See, e.g., Russell (2015) and McDonald (2019), but also Geist (2016). Note furthermore that the phenomenon of an AI race can be dangerous even if there is no military aspect to it; see Cave and ÓhÉigeartaigh (2018). On the geopolitics of AI, see, e.g., Alexandre and Miailhe (2017) and Lee (2018), while on the need for cross-cultural cooperation in AI ethics, see ÓhÉigeartaigh et al (2020).

[17] The most famous example of such an out-of-control spiral is the (still hypothetical) idea of an artificial general intelligence entering a recursive self-improvement cycle which, if the dynamic is sufficiently accelerating, may lead to what has been called a singularity or an intelligence explosion. But that is a pretty big "if"; see Yudkowsky (2013) for an ambitious treatment of this issue. Bostrom (2014), Piper (2018) and Russell (2019) argue that the possibility is a cause for concern, while LeCun (interviewed by Wakefield, 2015) and Floridi (2016) argue that it is not.

[18] This phrase "even more" is the one passage of this policy document that I am not quite happy with, because it can easily be taken as a taken as an endorsement of a claim along the lines of "on the whole, AI's benefits outweigh its risks" – to quote the draft version from 2018 of the EU Commission's aforementioned *Ethics Guidelines for Trustworthy AI*. That draft was made public in order to give stakeholders and citizens an opportunity to react before the final versions of the Guidelines were released. I took that opportunity and protested the "AI's benefits outweigh its risks" passage: "This is unfounded. I'm not saying the situation is symmetric or that the balance goes the other way. I'm saying we are far from knowing which way the balance goes. There simply isn't any serious study that systematically goes through the various potential benefits and

such as equity, accessibility for the disabled, humanitarian action, environmental protection, human flourishing and a sustainable society.[19]

References

Abadi, M., Chu, A., Goodfellow, I., Brendan McMahan, H., Mironov, I., Talwar, K. and Zhang, L. (2016) Deep learning with differential privacy, *CCS ´16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308-318, https://dl.acm.org/doi/abs/10.1145/2976749.2978318

Alexandre, L. and Miailhe, N. (2017) The geopolitics of AI and robotics, *Field Actions Science Reports* **17**, 84-87, https://journals.openedition.org/factsreports/4507

Amodei, D. and Hernandez, D. (2018) AI and compute, OpenAI, May 16, https://openai.com/blog/ai-and-compute/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D. (2016) Concrete problems in AI safety, https://arxiv.org/pdf/1606.06565

Ashurst, C., Anderljung, M., Prunkl, C., Leike, J., Gal, Y., Shevlane, T. and Dafoe, A. (2020) A guide to writing the NeurIPS impact statement, *Medium*, May 13, https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Brundage, M. and 25 coauthors (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, https://maliciousaireport.com/

Brynjolfsson, E. and McAfee, A. (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, New York.

Burrus, D. (2012) Is Technology good or evil? *Huffington Post*, August 24, https://www.huffpost.com/entry/is-technology-good-or-evi_b_1826270

Campa, R. (2018) *Still Think Robots Can't Do Your Job? Essays on Automation and Technological Unemployment*, D Editore, Ladispoli, Italy.

Cave, S. and ÓhÉigeartaigh, S. (2018) An AI race for strategic advantage: rhetoric and risks, *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, s 36-40, https://dl.acm.org/doi/10.1145/3278721.3278780

Cave, S. and ÓhÉigeartaigh, S. (2019) Bridging near- and long-term concerns about AI, *Nature Machine Intelligence* **1**, 5-6, https://www.nature.com/articles/s42256-018-0003-2

Chessen, M. (2017) The MADCOM future, Atlantic Council, September 26, https://www.atlanticcouncil.org/in-depth-research-reports/report/the-madcom-future/

risks, in order to establish that 'AI's benefits outweigh its risks'" (Häggström, 2019a). My protest was successful in that the claim was removed in the final version of the Guidelines (EU Commission, 2019). Of course, the truth value of the statement, if at all meaningful, remains unresolved, but findings that can optimistically be seen as a baby step towards a resolution were made by Vinuesa et al (2020).

[19] Again, see Vinuesa et al (2020) for an overview of various AI work in these directions.

Cossins, D. (2018) Discriminating algorithms: 5 times AI showed prejudice, *New Scientist*, 12 april, https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/

Cranefield, S., Oren, N. and Vasconcelos, W. (2019) Accountability for practical reasoning agents, *International Conference on Agreement Technologies: AT 2018*, 33-48, https://link.springer.com/chapter/10.1007/978-3-030-17294-7_3

Danaher, J. (2016) The threat of algocracy: reality, resistance and accommodation, *Philosophy & Technology* **29**, 245-268, https://link.springer.com/article/10.1007/s13347-015-0211-1

Danaher, J. (2019) *Automation and Utopia: Human Flourishing in a World without Work*, Harvard University Press, Cambridge, MA.

Danaher, J., Hogan, M., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S., Morison, J., Murphy, M.H., O'Brolchain, N., Schafer, B. and Shankar, K. (2017) Algorithmic governance: Developing a research agenda through the power of collective intelligence, *Big Data & Society*, July-December 2017, 1-21, https://journals.sagepub.com/doi/pdf/10.1177/2053951717726554

Dignum, V. (2019) *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer, New York.

EU Commission (2018) *Draft Ethics Guidelines for Trustworthy AI*, https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

EU Commission (2019) *Ethics Guidelines for Trustworthy AI*, https://ec.europa.eu/futurium/en/ai-alliance-consultation

de Fine Licht, K. and de Fine Licht, J. (2020) Artificial intelligence, transparency, and public decision-making: why explanations are key when trying to produce perceived legitimacy, *AI & Society*, https://link.springer.com/content/pdf/10.1007/s00146-020-00960-w.pdf

Floridi, L. (2016) Should we be afraid of AI? *Aeon*, May 9, https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible

Floridi, L. (2019) What the near future of artificial intelligence could be, *Philosophy & Technology* **32**, 1-15, https://link.springer.com/article/10.1007%2Fs13347-019-00345-y

Floridi, L. and Cowls, J. (2019) A unified framework of five principles for AI in Society, *Harvard Data Science Review*, https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/6

Frey, C.B. (2019) *The Technology Trap: Capital, Labor, and Power in the Age of Automation*, Princeton University Press, Princeton.

Geist, E.M. (2016) It's already too late to stop the AI arms race – we must manage it instead, *Bulletin of the Atomic Scientists* **72**, 318-321, https://www.tandfonline.com/doi/full/10.1080/00963402.2016.1216672

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2017) When will AI exceed human performance? Evidence from AI experts, https://arxiv.org/abs/1705.08807

Hagedorff, T. (2020) The ethics of AI ethics: an evaluation of guidelines, *Minds and Machines* **30**, 99-120, https://link.springer.com/article/10.1007/s11023-020-09517-8

Häggström, O. (2018) Vetenskap på gott och ont, in the KVVS volume *Vetenskaplig redlighet och oredlighet*, http://www.math.chalmers.se/~olleh/EtikKVVS.pdf with an English translation Science for good and science for bad at http://www.math.chalmers.se/~olleh/EtikKVVSEnglish.pdf

Häggström, O. (2019a) Comments on European Commission draft: Ethics Guidelines for Trustworthy AI, January 4, http://www.math.chalmers.se/~olleh/EU-TrustworthyAI.pdf

Häggström, O. (2019b) What is AI ethics? *CHAIR AI Ethics Seminar*, October 29, https://www.chalmers.se/en/centres/chair/events/Documents/WhatIsAIEthics.pdf

Häggström, O. (2020) En grön AI-politik? https://chalmersuniversity.app.box.com/file/645969220518

Harari, Y.N. (2018) *21 Lessons for the 21st Century*, Jonathan Cape, London.

Helberger, N. (2019) On the democratic role of news recommenders, *Digital Journalism* **7**, https://www.tandfonline.com/doi/full/10.1080/21670811.2019.1623700

Jobin, A., Ineca, M. and Vayene, E. (2019) The global landscape of AI ethics guidelines, *Nature Machine Learning* **1**, 389-399, https://www.nature.com/articles/s42256-019-0088-2

Johansson, F. (2019) Algorithmic fairness & machine learning, *CHAIR AI Ethics Seminar*, October 1, https://www.chalmers.se/en/centres/chair/events/Documents/Algorithmic_Fairness_and_Machine_Learning.pdf

Knight, W. (2020) AI can do great things – if it doesn't burn the planet, *Wired*, January 21, https://www.wired.com/story/ai-great-things-burn-planet/

Kobie, N. (2019) The complicated truth about China's social credit system, *Wired*, 7 juni, https://www.wired.co.uk/article/china-social-credit-system-explained

Larson, J., Mattu, S., Kirchner, L. and Angwin, J. (2016) How we analyzed the COMPAS recidivism algorithm, *ProPublica*, 23 maj, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lee, K.-F. (2018) *AI Superpowers: China, Silicon Valley and the New World Order*, Houghton Mifflin Harcourt, New York.

Marr, B. (2020) The top 10 artificial intelligence trends everyone should be watching in 2020, Forbes, Jan 6, https://www.forbes.com/sites/bernardmarr/2020/01/06/the-top-10-artificial-intelligence-trends-everyone-should-be-watching-in-2020/

McDonald, H. (2019) Ex-Google worker fears 'killer robots' could cause mass atrocities, *The Guardian*, September 15, https://www.theguardian.com/technology/2019/sep/15/ex-google-worker-fears-killer-robots-cause-mass-atrocities

Müller, V. (2020) Ethics of artificial intelligence and robotics, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/ethics-ai/

Nascimento, A.M., Vismari, L.F., Molina, C.B.S.T., Cugnasca, P.S., Camargo, J.B., de Almeida, J.R., Inam, R., Fersman, E., Marquezini, M.V. and Hata, A.Y. (2019) A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety, *IEEE Transactions on Intelligent Transport Systems*, https://ieeexplore.ieee.org/document/8892611

Nyholm, S. (2020) Humans and Robots: *Ethics, Agency, and Anthropomorphism, Rowman & Littlefield*, Lanham, MD.

ÓhÉigeartaigh, S., Whittlestone, J., Liu, Y., Zeng, Y. and Liu, Z. (2020) Overcoming barriers to cross-cultural cooperation in AI ethics and governance, *Philosophy & Technology*, https://link.springer.com/article/10.1007%2Fs13347-020-00402-x

Piper, K. (2018) The case for taking AI seriously as a threat to humanity, *Vox*, 8 maj, https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment

PwC (2019) *Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise?* https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

Roine, J. (2016) *Automatiseringens effekter på arbete och fördelning – en översikt av trender och mekanismer*, FORES, Stockholm, https://fores.se/wp-content/uploads/2016/05/AutomatiseringensEffekterRoine.pdf

Russell, S. (2015) Take a stand on AI weapons, *Nature* **521**, https://www.nature.com/news/robotics-ethics-of-artificial-intelligence-1.17611

Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, New York.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K. and Müller, K.-R. (2019) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, New York.

Schneier, B. (2015) *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*, W.W. Norton, New York.

Scholl, K. and Hanson, R. (2020) Testing the automation revolution hypothesis, *Economics Letters* **193**, 109287, https://www.sciencedirect.com/science/article/pii/S0165176520301919?dgcid=author

Shah, R. (2020) AI alignment 2018-19 review, *AI Alignment Forum*, January 28, https://www.alignmentforum.org/posts/dKxX76SCfCvceJXHv/ai-alignment-2018-19-review

Susser, D. (2019) Invisible influence: artificial intelligence and the ethics of adaptive choice architecture, *AIS '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society*, 403-440, https://dl.acm.org/doi/abs/10.1145/3306618.3314286

Thaler, R. and Sunstein, C. (2008) *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Yale University Press, New Haven.

Torres, P. (2019) Facing disaster: the great challenges framework, *Foresight* **21**, 4-34, https://www.emerald.com/insight/content/doi/10.1108/FS-04-2018-0040/full/html

Verma, S. and Rubin, J. (2018) Fairness definitions explained, *2018 ACM/IEEE International Workshop on Software Fairness*, http://fairware.cs.umass.edu/papers/Verma.pdf

Verweij, M. and Dawson, A. (2019) Sharing responsibility: responsibility for health is not a zero-sum game, *Public Health Ethics* **12**, 99-102.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S., Tegmark, M. and Nerini, F. (2020) The role of artificial intelligence in achieving the Sustainable

Development Goals, *Nature Communications* **11**, 233, https://www.nature.com/articles/s41467-019-14108-y

Wakefield, J. (2015) Intelligent machines: what does Facebook want with AI? *BBC News*, September 15, https://www.bbc.com/news/technology-34118481

Yudkowsky, E. (2013) Intelligence explosion microeconomics, Machine Intelligence Research Institute, Berkeley, CA, https://intelligence.org/files/IEM.pdf

Zerilli, J., Knott, A., Maclaurin, J. and Gavaghan, C. (2019) Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology* **32**, 661-683, https://link.springer.com/article/10.1007/s13347-018-0330-6

Zuboff, S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile Books, London.