Technological Opacity of Machine Learning in Healthcare

Herzog, Christian

Erstveröffentlichung / Primary Publication
Konferenzbeitrag / conference paper

Empfohlene Zitierung / Suggested Citation:
Herzog, C. (2019). Technological Opacity of Machine Learning in Healthcare. In *Proceedings of the Weizenbaum Conference 2019 "Challenges of Digital Inequality - Digital Education, Digital Work, Digital Life"* (pp. 1-9). Berlin https://doi.org/10.34669/wi.cp/2.7

PROCEEDINGS OF THE WEIZENBAUM CONFERENCE 2019
CHALLENGES OF DIGITAL INEQUALITY
DIGITAL EDUCATION | DIGITAL WORK | DIGITAL LIFE

TECHNOLOGICAL OPACITY OF MACHINE LEARNING IN HEALTHCARE

Christian Herzog
University of Lübeck, Institute for Electrical Engineering in Medicine
Lübeck, Germany
Christian.Herzog@uni-luebeck.de

ABSTRACT

Recently, a host of propositions for guidelines for the ethical development and use of artificial intelligence (AI) has been published. This body of work contains timely contributions for sensitizing developers to the ethical and societal implications of their work. However, a sustained embedding of ethics in largely algorithm-based technology development, research and studies requires a precise framing of the origins of the new vulnerabilities created. Recently, scholars have been referring to ethics associated with technology that is in some way "opaque" to at least part of its associated stakeholders. This "opacity" can take several forms which will be discussed in this paper. There are various ways in which such an opacity can create vulnerabilities and, hence, relevant ethical, societal, epistemic and regulatory challenges. This paper provides a non-exhaustive list of examples in healthcare that call for educational resources and consideration in development processes that try to reveal and counter these opacities.

KEYWORDS

Artificial Intelligence; Machine Learning; Ethical and Societal Implications; Technological Opacity; Healthcare

DOI: 10.34669/wi.cp/2.7

1 INTRODUCTION

The field of "control engineering"—a field that is in an increasingly close connection to the field of "machine learning"—has been famously dubbed a "hidden technology" by Karl J. Åström in 1999. Åström, a remarkable pioneer in his science, characterized the field as mostly "hidden", despite the fact that it solves novel problems humans are incapable of, is critical to a system's successful operation and almost omnipresent. Application examples range from "generation and transmission of energy, process control, manufacturing, communication, transportation and entertainment" (Åström, 1999). This perception has been reiterated as recently as on Sept. 21st, 2018, by Ian Craig in 2018, in his presentation "Automatic control: The hidden technology that modern society cannot live without". Being a control scientist by trade myself, I have to admit that this almost romantic notion is not without appeal. However, as has been noted, automatic control can "'camouflage' system failure by controlling against the variable changes, so that trends do not become apparent until they are beyond control" (Bainbridge, 1983). Thus "hiddenness" or "opacity" must be a prime consideration in holistic system designs. In case the presence of some control algorithm acting on the system is not apparent at all, every chance to counteract may be lost.

The field of "machine learning" constitutes a fundamental building brick of what is nowadays considered artificial intelligence (AI) and cannot be accused of being underreported even in mainstream media outlets these days, see, e.g., Brennen, Howard and Nielsen, 2018. In terms of popularization, politics and economics, the general existence of AI solutions is far from hidden, yet it may be, or bound to become, a "technology that modern society cannot live without".

As such, many use-cases of AI can be categorized as classification tasks and decision support systems, whereas this in turn can be viewed as the sensory input to human-in-the-loop control. In the future, some of these approaches may extend to fully automated decisions with a mere human supervision, e.g., in healthcare or autonomous driving (Topol, 2019). Since the range of applications of machine learning has been tremendously enriched as compared to classical control engineering due to the ability to operate on largely unstructured data, it is the task of this paper to investigate ways in which opacity in machine learning and AI applications may yield potentially undesirable ethical, legal and societal implications (ELSI).

The term "AI" is quickly changing and adapts to the current state of the art. From a scientific perspective, however, there is a need to identify sources of ELSI reliably in terms of a commonly understood und persisting taxonomy. One such potential class of sources has recently been discussed by means of the term "technological opacity". As a contribution to the discussion, this paper aims at further refining the taxonomy of forms of technological opacity. An application to the healthcare sector, a field in which adverse effects are often among the most dramatic, is aimed at showcasing the applicability of the proposed refinement.

The remainder is structured as follows: Section 2 reviews literature on the forms of opacity found in algorithmic technology development. Section 3 proposes an extended, but concise list of forms of technological and techno-social opacities, while section 4 provides examples in healthcare and section 5 outlines proposals for remedies. Section 6 concludes the paper.

2 TECHNOLOGICAL OPACITY—A BRIEF REVIEW

Technological opacity is a rather recent framework to denote that long- and short-term effects, ways of interaction, interdependencies, the inner workings, an influence on one self's or other's actions or the very existence of some technology and its application remains hidden to some stakeholders. While this is an attempt at an all-encompassing denomination, the literature does

not seem to have converged to a common definition so far.

For instance, Surden and Williams, 2016, define technological opacity in the following way:

> *"[…]'technological opacity' applies any time a technological system engages in behaviors that, while appropriate, may be hard to understand or predict, from the perspective of human users."*

In their paper, the authors argue that autonomous vehicles will typically not conform to the average driver's mental model about how other participants in traffic react. While they do assume that autonomous driving will yield reduced rates of traffic accidents, they essentially call for technology developers and regulators to standardize the behavior of autonomous vehicles. Considering a future, where autonomous vehicles abound, it might be meaningful not to restrict the afore-mentioned form of technological opacity to concern only "human users", as Surden and Williams do. If autonomous vehicles remain interconnected in ways that prevent a faultless prediction of future trajectories, this could also be denoted a technological opacity.

In a completely different context, Endo, 2018, identifies technological opacity in "predictive coding", which is machine learning used to assess or predict the relevance of documents in law-suits. Predictive coding is advertised as a technological fix to make the assessment of large amounts of documents economically viable, even in small-value claims and for parties with few resources. However, Endo argues that, currently, a lack of understanding of the technology and the cost of hiring technological experts actually prevents parties with few financial resources to make use of the technology.

Pasquale, 2015, focusses on opaque algorithms used in finance and largely attributes its existence to corporate secrecy, or rather the fear to give away economic advantages and attempts to consolidate power. Topol, 2019, in turn, associates the intransparency of trained deep neural networks in healthcare—or more specifically, a lack of means to interpret how actual outputs are determined from input data—with the controversy about "black box algorithms". Topol remarks that the recognition of the existence of such algorithmic opacity has led to the incorporation of transparency requirements in the European Union's General Data Protection Regulation (GDPR) as a prerequisite for practical deployment.

Vallor, 2016, introduces the term "techno-social opacity" which can be broadly referred to as the lack of understanding about the societal implications of specific technologies. While this review remains non-exhaustive and limited to literature that actually uses the term "opacity", the afore-mentioned references largely refer each to a single and specific form of opacity. In what follows, an attempt is made to distinguish between several forms with the aim to allow a systematic and holistic analysis of technological opacities ranging from opacity that is intentional, opacity based on a lack of understanding or complexity, opacity based on a lack of perception of societal effects over opacity from transdisciplinarity to procedural opacity.

3 FORMS OF TECHNOLOGICAL OPACITY

Burrell, 2016, provides a nuanced view on technological opacity with respect to algorithms and identifies three forms: (i) Opacity as intentional secrecy (attributed to Pasquale, 2015), (ii) opacity as technical illiteracy and (iii) opacity from complexity as the mismatch between mathematical machine learning outputs and human ways of interpretation.

The above forms of opacity address different stakeholders: While the first form is largely associated with active decisions on the part of the developing enterprise or intentional regulatory omissions, the second form is passive in the sense that it is associated with the fact that, e.g., the majority of the population is not capable to understand the intricacies of technological development and its products, e.g., programming. Burrell disambiguates this form from the third,

which also applies to proficient technology developers and elaborates that at the heart of the third form of opacity are (self-)learning algorithms that may change the decision logic and operate on vast amounts of data. Such algorithms hence incur complexity that results in algorithmic outputs whose underlying rationale may be difficult to comprehend. Similarly, Matthias, 2004, observes that machine learning can yield algorithms, derived, e.g., by supervised learning, for which *"the human trainer himself is unable to provide an algorithmic representation."* Opacity arising from technological complexity is also the apparent focus of Stahl and Coeckelbergh, 2016, who identify "ubiquity and pervasiveness", "speed of innovation", the "distributed and networked nature" and "logical malleability" (unforeseeable other use-cases) as novel challenges in information and communication technologies (ICT). Royakkers *et al.*, 2018, add that (self-)learning algorithms also challenge the mental models of users, which may fail to have a working anticipative notion of what a self-learning algorithm may do.

Technology related to 'Big Data' are solutions that help analyze amounts of data that humans cannot handle (Mittelstadt and Floridi, 2016). With increasing computational resources, in some years' time, ever larger amounts of data may not be difficult to analyze technologically, but whether the technology that was used to analyze it is still humanly manageable is an issue of technological opacity resulting from complexity. Unforeseeable use-cases and effects are also covered by Vallor, 2016, who has further introduced the notion of *techno-social opacity*, meaning the inadequacy of abilities to either predict adverse effects or to work towards desired societal effects by means of technological innovation (or refraining from it). In the following, I propose to append the list of forms of opacity by a fifth and sixth notion, which, in the sequel, will be argued to be distinct from the other notions by means of illustrating their relevance in the healthcare context. The fifth form denotes opacity as a result of the transdisciplinary nature of applications and the complexity this incurs. This form is not specific to machine learning algorithms, while at the same time, I acknowledge that possibly the primary examples for such applications today stem from this field. To a large degree this is a result of the complexity incurred when the design of algorithms requires both in-depth mathematical and technical knowledge as well as application knowledge, e.g., proficiency in the medical sciences. This form is distinct from techno-social opacity, in that the development team may boast a clear vision about what societal impact is desired, but, e.g., fails to identify the relevant modalities, such as prevailing hospital workflows, lack of practitioner training, etc., that hinders their development in realizing this impact.

A sixth and—for the purposes of this paper—final form of opacity can be found in *procedural opacity*. While one might reasonably choose to consider this as part of broader, fourth and fifth forms, I argue that it is more appropriate to reserve yet another form of technological opacity associated with effects on behalf of a processes end-point, or "customer". While this can be argued to not be strictly a form of *technological* opacity, I further argue that the achievements in machine learning and its endeavors in, e.g., administrative automation, intertwine both processes and technology in novel ways. For instance, the above-mentioned example of predictive coding may encompass this form of opacity if other parties in some law-suit are not even aware if, on which data sets and to which extent the technology has been applied (Endo, 2018). Procedural opacity in the technological sense may also always exist in situations where end-customers, patients or any form of recipient relying on some service has to trust that within the underlying technology-supported process (possibly unknown to her or him due to intentional secrecy) the responsible personnel did not suffer from too much technical illiteracy and the development team had enough transdisciplinary competence and kept technological complexity at a manageable level. In this sense, opacity

resulting from faults in technology-driven processes usually concerns the interaction of humans and technology, e.g., algorithmic decision support systems. It can be dismissed as being a meta-form aggregating other occurrences of technological opacity—however, it stimulates a holistic perspective on the involved processes and is hence useful for this reason.

To summarize, it is proposed to distinguish the following forms of technological opacity:

i. Opacity from intentional secrecy
ii. Opacity from technical illiteracy
iii. Opacity from technological complexity
iv. Opacity from techno-social interdependence
v. Opacity from application transdisciplinarity
vi. Opacity from technology-driven processes

This list may not be complete, but it will hopefully act as a structuring taxonomy for future discussions and will provide a frame for examples of technological opacity in healthcare.

4 Technological Opacity in Healthcare

In what follows, a non-exhaustive overview is given about potential issues in healthcare that may arise due technological and techno-social opacity. This overview is intended to act as a stimulator to discussions, more in-depth analyses and a means of sensitization for those unfamiliar with the context. In pointing out issues particular to novel machine learning algorithms applied in the healthcare context, the aim is to help development teams adequately realize the potential of machine learning solutions in healthcare, where it is an appropriate technological fix (Sarewitz and Nelson, 2008).

4.1 Informational Power and Medical Data

Information is power and producing information from data that would otherwise not be informative is an instrument of power. Such an instrument requires regulation as it can lead to data processors having superior power over the subjects of the data (Mittelstadt et al., 2016). Especially in healthcare, data-driven medical technology, e.g., automated diagnosis tools, may potentially shift the balance in terms of the authority w.r.t. medical expertise from physicians to medical technology providers. In the extreme, the aggregation of diverse types and large amounts of information can drive humans out of the loop (Royakkers *et al.*, 2018). This has also epistemic implications, elaborated on in section 4.3. Without regulatory intervention, medical technology providers can consolidate their informational power by applying algorithm secrecy as a means to intentionally generate opacity. This may be largely considered an objectionable transition as educational resources should be freely available and may amount to increased commercialization of medical knowledge, and hence a driver of increased inequality.

Artificial intelligence tools might further provide the possibility for self-diagnosis. Here, it is highly probable that medical technology providers could try to prevent accidental diagnosis, e.g., by blocking access to the raw image material in image-guided medical diagnosis systems, for fear of being held morally obliged to provide the technological means to diagnose everything. What if the patient performing self-diagnosis actually observes something to diagnose from the raw image material, but the device does not recognize it? This would result in unwanted bad publicity and (a justified) loss in trust. Blocking access to raw image material hence incurs opacity both from, possibly unwarrantable, intentional secrecy and from the process itself.

4.2 Technological Illiteracy

Medical practice is multifaceted, subject to time pressure, and social and psychological nuances are highly relevant in making diagnoses. If electrocardiographic pathology detection would follow a set of simple rules, devoid of experience and objective guidelines that are easy to formalize, deep learning would not be needed to automate it (Hannun *et al.*, 2019). With the current

load of knowledge that medical practitioners always need to have at their disposal, it is questionable how much in-depth understanding of the inner workings, e.g., of algorithmic decision support systems, can be reasonably expected. However, automated diagnosis support is expected to curb costs, improve on diagnosis quality and potentially even transfer healthcare to home care (Derrington, 2017).

While patient compliance in medicine is an ongoing issue, it stands to reason that patients with varying degrees of technological illiteracy may show 'consent fatigue' (Royakkers *et al.*, 2018), when confronted with algorithmic suggestions. Further, it may not be meaningful to require disabled and impaired people (e.g., when suffering from dementia) to consent to technology that, e.g., monitors them (Royakkers *et al.*, 2018) for lack of an understanding about the implications. Home care tools following the rationale of "technological paternalism" as a concept denoting that "technology knows better what is good for us" may infringe upon personal autonomy, especially if performed without consent or knowledge of the user (Royakkers *et al.*, 2018). For patients with sufficient mental capabilities, this can be potentially addressed by forms of tolerant paternalism (Floridi, 2015), a framework that intents to increase the level of knowledge and to provoke a more informed decision.

4.3 EPISTEMIC ISSUES

Sometimes, AI evangelists propagate a vision of a utopian healthcare system, in which medical data can be securely shared under privacy regulations and can be effectively used for a globalized, automated inference-based diagnosis. However unlikely (Topol, 2019), in a transition towards this vision, physicians would have to work cooperatively with medical diagnosis devices and would need to understand the principles of the learning algorithms, be provided with transparent interfaces to be able to enter data with the appropriate quality and, hence, make use of automated diagnosis tools responsibly (Char, Shah and Magnus, 2018). These tools could rely on data that can be biased (most of the medical data is acquired in intensive care situations, which is often not representative), incomplete or out-of-date, facts that could not be apparent to the medical practitioner, the development team or service provider and, least of all, the patient.

Furthermore, the apparent superiority of some AI systems in clinical trials (Hannun *et al.*, 2019; Topol, 2019) may lead doctors to refrain from questioning the validity of the computer models altogether and develop over-confidence in machine intelligence (Burrell, 2016). Practitioners might not be able to take responsible action, in particular, when the data presentation is overwhelming and poorly interpretable. The process of knowledge generation and preservation might become largely commercialized and potentially opaque to science or unavailable as educational resources. The distributed nature of medical data fusion and aggregation may further result in epistemic opacity.

Similarly, economic pressures could lead to the early deployment of opaque AI-based solutions at the risk of significantly reducing life-saving accidental diagnoses, because the system's focus may be too narrow and is promoted to work entirely without physician intervention. Data-driven systems that are end-to-end, i.e., that aim at directly drawing actionable conclusions from largely unstructured data, potentially only imply a modeling process (defined as deriving a semantic description of a system that is generalizable to some other system to at least some extent), however, one which neither practitioner nor developer has fully specified and hence been able to investigate its implications. If a technology's applicability to real-world scenarios is immediate, with little or no pre-processing of data and accessible tools, this can create the illusion of simplicity where this is actually not true. At the heart of this, there may be a conflation or confusion of causation and correlation (Lipton and Steinhardt, 2018). In complex applications, epistemic opacity at some level in the

dependence structure from methodical experts, application experts, economic stakeholders to application subjects will create vulnerabilities that propagate, eventually affecting the patients directly. Clearly, errors can never be completely avoided, but it appears as though popular algorithms in machine learning, such as deep learning, currently are susceptible to incorporating large epistemic gaps, i.e., a lack of scientific foundation in the modeling approach, that incurs opacity.

4.4 Techno-Social Interdependence

Access to both medical data and expertise for training AI systems will become (or rather already is) a commodity, novel stakeholders can acquire, utilize and trade. The low running costs of machine-support lead to a strong potential to realize greater epistemic equality via access to intelligent decision support systems (on par with the skill of medical experts) by the less privileged, e.g., for citizens of areas with lower density of healthcare professionals. But without proper considerations, AI-based decision support systems could be distributed unequally. There appears to be a high degree of uncertainty about whether either greater equality or inequality will come to pass using AI-based healthcare solutions (Topol, 2019). It becomes apparent, however, that an ever more wide-spread use of automated analysis of medical data could eventually force patients to consent to data sharing and opting-out of their rights to privacy, as otherwise they cannot be sure to receive the same quality of treatment (Char, Shah and Magnus, 2018). The fiduciary relationship between physician and patient may break entirely, potentially leaving patients without an adequate notion about the whereabouts of their data.

4.5 Application Transdisciplinarity

An explicit example that showcases the effects of inadequately addressing transdisciplinary issues in technology development can be observed in many current use cases of electronic health records (EHR) in the United States. Studies have shown that current EHR systems have, in fact, increased the workload and stress of physicians (Gardner *et al.*, 2019). An adequate transdisciplinary consideration of aspects of human-computer interaction, work psychology and knowledge aggregation can be a remedy.

A further example on the intricacies of transdisciplinary research is the automated drug delivery during anesthesia, which has largely relied on rather transparent dynamic pharmacokinetic model structures potentially augmented by inference algorithms for individualizing the model to a specific patient (Neckebroek, De Smet and Struys, 2013). However, automated drug delivery is not yet fully realized. It is illustrative to compare the model semantics in papers written for technologists and physicians, which gives a hint on the difficulty to express mathematical expressions in, e.g., prose. This friction in transdisciplinary research cannot be avoided, but it is important to be sensitive to the mutual opacities of the partners in a development team.

4.6 Procedural Opacity

At least initially, algorithmic decision support systems in healthcare, pathology detection algorithms or diagnosis tools will most likely be designed for a single or, at least, only a few multiple use-cases. Effectively incorporating these tools into the medical workflow will be challenging and their limits need to be clear. If some pathology detection tool, e.g., on electrocardiographs, is deployed only for specific detection tasks with the promise of curbing costs, it might, in fact, prevent the accidental diagnosis of other pathologies it is not designed for. On behalf of the patient, this incurs procedural opacity, because expectations might be to receive an all-encompassing treatment or diagnosis. Compartmentalization in the medical domain is already observable to thwart the full realization of this expectation. However, it is—for the most part—

sufficiently obvious for any patient to receive only specialized treatment. With automated and specialized diagnosis tools, further compartmentalization may yield further opacity. To counter this and at the same time provide an actual perspective in saving time for medical specialists, automated diagnosis technology should be designed as holistically as possible, which is more challenging to achieve than advertised.

Char, Shah and Magnus, 2018, further warn that machine learning designers could be tempted to optimize for reimbursement rather than quality of care—a vision which lies at the intersection of opacity in processes and from complexity.

5 POSSIBLE REMEDIES

Within the literature, there is a range of remedies proposed to mitigate the effects of technological opacity. For instance, Andras *et al.*, 2018, ask for natural language processing to provide explanations for opaque machine learning applications. Explainable AI is a current research topic that can range from image highlighting to automatically derive explanatory labels that shed light, e.g., on the rationale behind a classification task. Explainable AI can provide an inward look into trained models post-hoc, but the success of the training could remain trial and error and hence amounts to opacity from complexity on behalf of the developers.

To mitigate complexity in (supervised) learning-based algorithms and what is sometimes termed the "Reproducibility Crisis" of AI (Voosen, 2017; Hutson, 2018), leading researchers demand more rigor in neural network training (Sculley *et al.*, 2014; Rahimi and Recht, 2017). So-called "black-box" modeling approaches are quite common in control theory, where it is conservatively applied. "Conservatively", here refers to specialized data pre-processing, highly structured input-output data, reflection on modeling assumptions and structure as well as model verification and validation. In contrast, black-box neural network-based modeling processes are difficult to be validated and it appears to be both one of the largest advantages and (epistemic) weaknesses about the technology that it can be applied to very unstructured data.

Apart from technical remedies, non-technical solutions are required for all other technological opacities. Extending the application of "Responsible Research and Innovation" (RRI) (Grunwald, 2011) may be a solution (Stahl and Coeckelbergh, 2016), in which a wide range of stakeholders need to cooperate closely. Furthermore, post-hoc analysis of cases, in which (assumedly unintended) opacity yielded adverse effects, is necessary to answer essential questions, e.g., on the degree of necessary interdisciplinary education of engineers. Driving factors of specialization and the complexity of the technology will possibly set a limit, but an awareness for potential opacity should be a minimum goal. Consequently, there might be a need for a class of engineers trained in RRI.

6 CONCLUSION

The paper provided a non-exhaustive list of examples of technological opacities in healthcare, categorized into different forms and compiled with the purpose to illustrate the multi-faceted ways in which opacity can be generated by automation and machine learning. There exists no panacea that can act as a solution, but, it appears as though the variety of issues presented illustrate a need for transdisciplinary research teams to work on holistic approaches to machine learning and artificial intelligence solutions in healthcare that mix well with current workflows or improve them, circumvent a range of the above-mentioned adverse ethical, legal and societal implications and actually contribute to the improvement of the quality, equality and effectiveness in healthcare.

7 ACKNOWLEDGMENTS

The author would like to thank Vincent Müller for his feedback on the topic.

8 REFERENCES

1. Andras, P. et al. (2018) ‘Trusting Intelligent Machines Deepening trust within socio-technical systems’, IEEE Technology and Society Magazine. IEEE, 37(december), pp. 76–83.
2. Åström, K. J. (1999) ‘Automatic Control - The Hidden Technology’, in Frank, P. M. (ed.) Advances in Control, pp. 1–28.
3. Bainbridge, L. (1983) ‘Ironies of Automation’, Automatica, 19(6), pp. 775–779.
4. Brennen, A. J. S., Howard, P. N. and Nielsen, R. K. (2018) ‘An Industry-Led Debate: How UK Media Cover Artificial Intelligence’, Reuters Institute for the Study of Journalism Fact Sheet, (December), pp. 1–10.
5. Burrell, J. (2016) ‘How the Machine “Thinks:” Understanding Opacity in Machine Learning Algorithms’, Big Data & Society, 3(1), pp. 1–12.
6. Char, D. S., Shah, N. H. and Magnus, D. (2018) ‘Implementing Machine Learning in Health Care’, New England Journal of Medicine, 378(11), pp. 981–983.
7. Craig, I. (2018) ‘Automatic control: The hidden technology that modern society cannot live without’. University of the Witswatersrand, Johannesburg, South Africa.
8. Derrington, D. (2017) Artificial Intelligence for Health and Health Care, JASON Report. Available at: https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf.
9. Endo, S. K. (2018) ‘Technological Opacity & Procedural Injustice’, Boston College Law Review, 59 (forthcoming), pp. 821–876.
10. Floridi, L. (2015) ‘Tolerant Paternalism : Pro-ethical Design as a Resolution of the Dilemma of Toleration’, Science and Engineering Ethics. Springer Netherlands.
11. Gardner, R. L. et al. (2019) ‘Physician stress and burnout: the impact of health information technology’, Journal of the American Medical Informatics Association, 26(2), pp. 106–114.
12. Grunwald, A. (2011) ‘Responsible Innovation: Bringing together Technology Assessment, Applied Ethics, and STS Research’, Enterprise and Work Innovation Studies, 7, pp. 9–31.
13. Hannun, A. Y. et al. (2019) ‘Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network’, Nature Medicine, 25, pp. 65–69.
14. Hutson, M. (2018) ‘Artificial intelligence faces reproducibility crisis’, Science, 359(6377), pp. 725–726.
15. Lipton, Z. C. and Steinhardt, J. (2018) ‘Troubling Trends in Machine Learning Scholarship’, pp. 1–15.
16. Matthias, A. (2004) ‘The responsibility gap: Ascribing responsibility for the actions of learning automata’, Ethics and Information Technology, 6, pp. 175–183.
17. Neckebroek, M. M., De Smet, T. and Struys, M. M. R. F. (2013) ‘Automated Drug Delivery in Anesthesia’, Current Anesthesiology Reports, 3(1), pp. 18–26.
18. Pasquale, F. (2015) The Black Box Society - The Secret Algorithms That Control Money and Information. Cambridge, Massachusetts; London, England: Harvard University Press.
19. Rahimi, A. and Recht, B. (2017) Reflections on Random Kitchen Sinks Back When We Were Kids, arg min blog.
20. Royakkers, L. et al. (2018) ‘Societal and ethical issues of digitization’, Ethics and Information Technology. Springer Netherlands, 20(2), pp. 127–142.
21. Sarewitz, D. and Nelson, R. (2008) ‘Three rules for technological fixes’, Nature, 456(7224), pp. 871–872.
22. Sculley, D. et al. (2014) ‘Machine learning: The high-interest credit card of technical debt’, in Proceedings from SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop).
23. Stahl, B. C. and Coeckelbergh, M. (2016) ‘Ethics of healthcare robotics: Towards responsible research and innovation’, Robotics and Autonomous Systems. Elsevier B.V., 86, pp. 152–161.
24. Surden, H. and Williams, M.-A. (2016) ‘Technological Opacity, Predictability, and Self-Driving Cars’, Cardozo Law Review, 38(121), pp. 121–181.
25. Topol, E. J. (2019) ‘High-performance medicine: the convergence of human and artificial intelligence’, Nature Medicine. Springer US, 25(1), pp. 44–56.
26. Vallor, S. (2016) Technology and the Virtues. Oxford University Press.
27. Voosen, B. P. (2017) ‘The AI Detectives’, Science, 357(6346), pp. 22–27.