L. Ahrenberg

Concerns in light of technology performance

Motivation

Many of today's digital technologies are built to replace or supplement human actors in a much more intimate social context than previous generations of artifacts The tighter integration and increasing complexity of human-technology interactions are causing new challenges in the social, juridical, and engineering sciences, with questions of how to critique, audit govern, and develop modern digital technologies, such as AI and data processing.

Abstract

This is a brief review of some recent work on the ethics, critique, and governance of digital technology. I list a number of proposed approaches to internal and external governance, and define a set of related concerns. I then attempt to relate the concerns in the context of the task performed by a human agent, and what relative performance society might expect from technologies in this regard. Finally, I discuss a few potential scenario configurations and draw conclusions.

Approaches to Governance

Regulation

Regulation, e.g. GDPR, used to define laws for technology.
Conflicts may occur in situations where humans and technology perform equivalent tasks while sharing space, yet answer to different laws. Regulation does not guarantee ethical behavior, and people have the ability to "bend the law".
Encoded regulation can be problematic to verify. Regulation requires enforcement, and additional tools, such as audits to be a useful form of governance, and is dependent on **transparency** and **traceability**. See e.g. *Floridi (2018)* and the concept of *Soft Ethics*.

Critique & Journalism

Independent critique and journalism are cornerstones of an open society, invaluable to discuss issues with e.g. **fairness**. However, it require access, time, and knowledge of the object under scrutiny. While complex oranizations require similar assets, the replacement of humans by automation poses additional problems:People can be asked questions, while algorithms give no interviews. Critique then depends acutely on **transparency** and **traceability**. For more on the challenges to critique, and the necessity of transparency, see e.g. *Ananny (2015)*, and *Zarsky (2016)*.

Audits

Independent reviews by regulators or consultants are often mentioned as a key component of technology governance, striking a balance between business opacity and public interest. Auditors still need the technology to be **transparent** and **traceable** of course. Organizations seeking to integrate algorithms with human-level performance should attend to **integration** as well. Potential automated audits involves some hard problems, and risks simply propagating the traceability concern. For an introductions to auditing and opacity, see *Burrell (2016)*.

Ethics board

An organization or technology producer can have an internal board of ethical experts, separate from the development team. This board is tasked with approving architecture, plan, usage, and to raise concerns in relation to the developed software. It has the benefit of presence during all stages of development with the ability to raise **fairness** and **integration** concerns. However, **transparency** and **traceability** will not propagate outside of the organization. An ethics board could be used with for instance standards to ease the work of technology audits. See *Neyland (2015)* for an account of an ethics board being utilized in a software project.

Standards & Safety-critical development

Standard for responsible use of Artificial Intelligence, are being developed by organizations such as IEEE. Accredited software will guarantee that certain prerequisites are met for e.g. **transparency**, **fairness**, and **traceability**. Note that automatic accreditation techniques may not be sufficient, and hard to verify. Safety-critical development employs strict specifications and tests, fitness for purpose could be indicated. This type of governing may prove challenging for today's lean methodologies. See *Bryson and Winfield (2017)* for an introduction to IEEE's standardization work. See Kroll (2018) for an argument for Safety-critical development.

Virtuous technology & development

Development teams can reason about the engineering process, recognizing that assumptions and design decisions may have consequences for **fairness** and **integration**, and striving for sufficient levels of **transparency** and **traceability**. Understanding complex socio-technological consequences is no easy task and requires experts in human-computer interaction. See *Friedman et al. (2009)* for an example of so called value-sensitive design. See *Ananny (2015)* for a discussion of virtuous critique of technology.

Performance Concerns

It is insufficient to look exclusively at technical performance. As *Lipton and Steinhardt (2018)* points out, claims of human-level performance in today's technical literature risks confusing the discussion by associating value laden terms with statistcal fitness. Based on the work by *Ananny (2015)* for evaluating technology in context, and the ethical map of *Mittlestadt et al. (2016)*, I have therefore tried to relate a number of concerns in literature to human performance in these areas.

Task

As noted above, adherence to technical specification may not be sufficient to relate machine and human performance.Thus *task* in this work is the antrophomorphic characterization of a job. It denotes the level at which a technology is considered to perform something (drive a car, identify a dog in a photo, or rank job applicants,...) on par with a person, in the same context. As the complexity of tasks can vary greatly what people may consider part of the task varies as well. The baseline, however, for technology replacing humans in context is the performance of people.

Fairness

Most people have an intuitive sense of what they consider fair, but it is perhaps easier to describe its oposite: Unfairness is defined by *Zarsky (2016)* as when individuals are treated differently than their peers on the basis of irrelevant differences. Individuals are not always fair, but it is common to attempt to find the reason in context or personality, e.g. having a bad day. When systematic, it is discrimination, and considered a question of values. Technology on the other hand is expected to only act on relevant context, and be systematic in a task. A failiure to do either is in a sense unfair. A monoculture, and as such at risk of locking in discriminatory behaviour for minorities even when cosidered performing fair by most people in a society.

Transparency

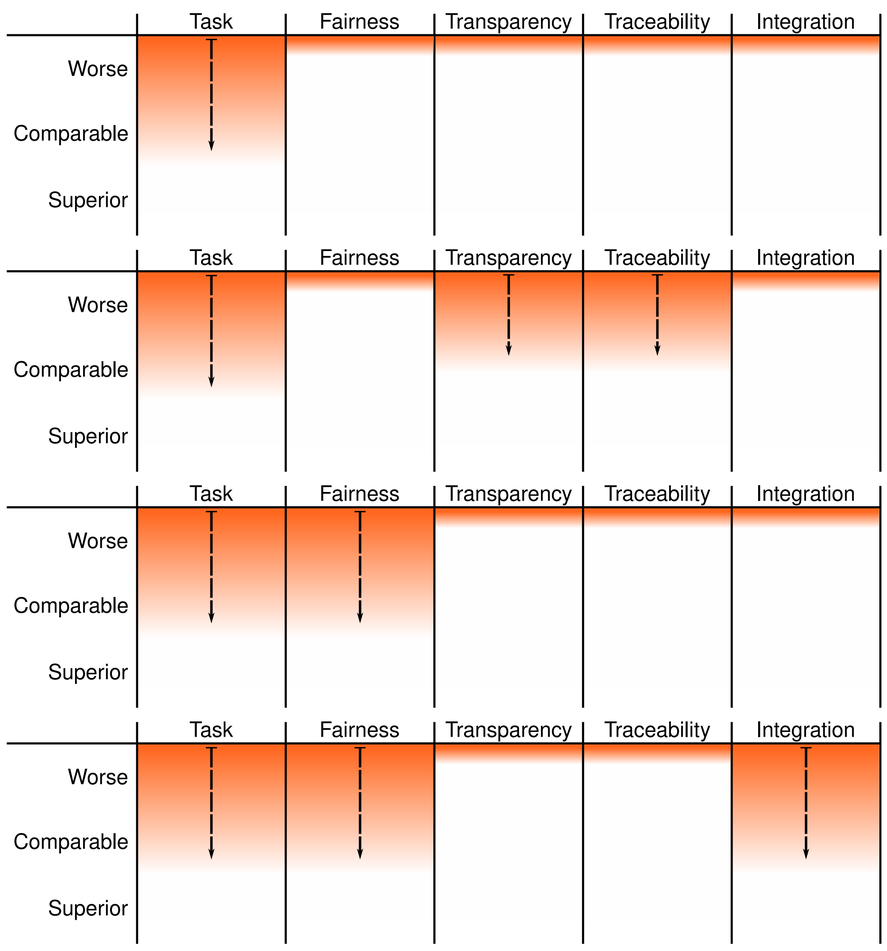
Burell (2016) lists three types of transparency i) As strategy (e.g. to protect business), ii) From requirement of skill and knowledge, and iii) From difference in machine and human 'reasoning'. *Zarsky (2016)* further argues that transparency is required to guarantee fairness, and that automation makes systems inscrutable. Humans are not transparent, we have a right to privacy. But if required we are also expected to explain ourselves and our reasoning, and may be considered mentaly unfit if failing to do so. Technology therefore, will likely be required to perform much better than people on transparency, providing enough detail to reason about its actions when performing a task.

Traceability

Traceability is the capacity to hold persons accountable for actions caused by other agents.
When a person cannot be held responsibility, traceability is expected.
As technology replaces human actors in organizations shifts occurs in accountability. Not only within the organization, but also to the provider of technology.
Kroll (2018) writes: *"systems are human artefacts, built by a purpose by some human agency that must be accountable for the behaviours of those artefacts"*
People are usually held responsible for their actions, as well as for those of others in their care, for instance minors or pets.

Integration

Integration aims to relate changes to the context when replacing actors. This could include what *Gillespie (2014)* calls *Entanglement with practice*, where people change behaviour to adopt to technology, or *Automation bia*s, where algorithmic results are given more redibility, as outlined by *Cummings (2004)*. Integration indicators include the reach, resolution, and fidelity of the actor, as well as its capability to understand informal communication, gain trust, and adapt. People sometimes works around formal requirements to 'make things work', scribble notes in the margin, or ask a favour. For technology to integrate on a performance level comparable with humans is therefore a major challenge of great complexity.



Discussion

For novel technologies such as AI, designed to perform human-like services while embedded in social and physical space, good **task** performance is not sufficient by itself. It will be important to mitigate disruption to context by fulfilling the implicit normative claims humans have on each other. Failing to do so risks rendering the technology disruptive and threatening as it appears ungoverned.

Meeting the expectations of **transparency** and **traceability** allows **critique** and **auditing** of technology, creating debate and laying the groundwork for **regulations** and **standards**.

It is possible that using industry self-governance such as **ethics boards**, **accreditation**, and **virtuous development**, the **fairness** of a technology may reach a comparable level to humans in the public's view. That would not necessarily mean such technology isn't disruptive if it fails to integrate well into society. A scenario with high performing **integration** in addition to **task** and **fairness** could end up being tolerated by society even without **transparency** or **traceability**, but will likely come at a loss of public influence over policy. Lack of **traceability** is not only a concern for the public, it is likely to be problematic for corporations relying interleaving technology in their practices.

It is important to remember that technology performing as well as people is not perfect, but reflecting a population. While people have a life outside their task, our artifacts do not, and even very well integrated technologies risks locking in a set of values or behaviors making them digital laws.

None of these concerns have straight-forward solutions. **Transparency** and **traceability** may be the ones most readily addressed by industry today. **Fairness** will require looking beyond sample output and thinking about how to be systematically fair, but not unfair. **Integration**, finally, is perhaps never fully possible, and begs the question of how much of human society is dependent on our talent for informal communication, and for breaking the rules.

References

Ananny, M., Toward an ethics of algorithms, Science, Technology, & Human Values, 41(1), 93–117 (2015). http://dx.doi.org/10.1177/0162243915606523

Beer, D., The social power of algorithms, Information, Communication & Society, 20(1), 1–13 (2016). http://dx.doi.org/10.1080/1369118x.2016.1216147

Bryson, J., & Winfield, A., Standardizing ethical design for artificial intelligence and autonomous systems, Computer, 50(5), 116–119 (2017). http://dx.doi.org/10.1109/mc.2017.154

Burrell, J., How the machine 'thinks': understanding opacity in machine learning algorithms, Big Data & Society, 3(1), 2053951715622512 (2016). http://dx.doi.org/10.1177/2053951715622512

Cummings, M., Automation bias in intelligent time critical decision support systems, In , AIAA 1st Intelligent Systems Technical Conference (pp. 6313) (2004). : .

Floridi, L., Soft ethics, the governance of the digital and the general data protection regulation, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180081 (2018). http://dx.doi.org/10.1098/rsta.2018.0081

Friedman, B., Kahn, P. H., & Borning, A., Value sensitive design and information systems, In (Eds.), The Handbook of Information and Computer Ethics (pp. 69–101) (2009). : John Wiley & Sons, Ltd.

Gillespie, T., The relevance of algorithms, In T. Gillespie, P. Boczkowski, K. Foot, W. Bijker, & W. Carlson (Eds.), Media technologies: Essays on communication, materiality, and society (pp. 167–193) (2014). : Cambridge: MIT Press.

Kroll, J. A., The fallacy of inscrutability, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180084 (2018). http://dx.doi.org/10.1098/rsta.2018.0084

Lipton, Z. C., & Steinhardt, J., Troubling trends in machine learning scholarship, CoRR, (), (2018). https://arxiv.org/abs/1807.03341

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

Neyland, D., Bearing account-able witness to the ethical algorithmic system, Science, Technology, \& Human Values, 41(1), 50–76 (2015). http://dx.doi.org/10.1177/0162243915598056

Zarsky, T., The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making, Science, Technology, & Human Values, 41(1), 118–132 (2016). http://dx.doi.org/10.1177/0162243915605575

Email: lukas.ahrenberg@aalto.fi

Acknowledgements
I wish to thank Juho Pääkkönen and Jesse Haapoja for suggesting many interesting articles related to the project, as well as Marko Turpeinen and the rest of the class for thoughtful discussions.
This poster was made while taking part in Marko Turpeinen's Digital Ethics course at Aalto university.

Aalto University
School of Science