Responsible Design of Technical Applications of AI

Wiggers, P.; Pauwels, Eric; Piersma, N.

Publication date
2020
Document Version
Final published version

Responsible Design of Technical Applications of AI

PASCAL WIGGERS, Amsterdam University of Applied Sciences, The Netherlands
ERIC PAUWELS, CWI, The Netherlands
NANDA PIERSMA, Amsterdam University of Applied Sciences, The Netherlands and CWI, The Netherlands

While the technical application domain seems to be to most established field for AI applications, the field is at the very beginning to identify and implement responsible and fair AI applications. Technical, non-user facing services indirectly model user behavior as a consequence of which unexpected issues of privacy, fairness and lack of autonomy may emerge. There is a need for design methods that take the potential impact of AI systems into account.

ACM Reference Format:
Pascal Wiggers, Eric Pauwels, and Nanda Piersma. 2020. Responsible Design of Technical Applications of AI. 1, 1 (September 2020), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Artificial Intelligence (AI) and machine learning powered services are rapidly becoming part of our everyday lives, for example in the form of news feeds on social media content recommendation and in smart devices such as speech-based home assistants and smart thermostats or fitness trackers.

In the socio-economical domain, decisions such as whether someone qualifies for a mortgage or fits a particular job description, are increasingly based on AI systems that are trained on historical data that may lead to biased decision making.

This large-scale application of AI to human-behavior data in the socio-economical domain and for personalized services has shown the need for more responsible and human-centered approach to AI, as undesirable ethical and social consequences of the 'unreasonable effectiveness of data' [3], such as bias, unfairness and the lack of transparency and loss of autonomy were identified [e.g. 5, 8, 9].

While the application of AI to human-behavior data and in user-facing systems is a relatively recent development, AI algorithms and machine learning have been applied to engineering, manufacturing and technology for decades. The mathematical notion of optimization of accuracy underlying AI algorithms that may cause fairness-issues when applied to behavioral data [4] is supposed to fit very well to the goals of for example production environments or logistic processes, where ethical dilemmas do not exist. It also aligns with the overarching goals of cost minimization and profit maximization of market capitalism. Typically, such technical environment are closed worlds in which it is

Authors' addresses: Pascal Wiggers, Amsterdam University of Applied Sciences, Responsible IT group, Amsterdam, The Netherlands, p.wiggers@hva.nl; Eric Pauwels, CWI, Department of Intelligent and Autonomous systems, Amsterdam, The Netherlands; Nanda Piersma, Amsterdam University of Applied Sciences, Urban Analytics group, Amsterdam, The Netherlands , CWI, Department of Intelligent and Autonomous systems, Amsterdam, The Netherlands.

Manuscript submitted to ACM

clear what can be measured and what the consequences of algorithmic decisions are. Therefore, in such environments, traditional error minimization, cost optimization as model fundamentals are well defined and adequate. Much of the active discussion around ethical issues of AI in the socio-economical domain finds its origin in the direct use of these same model fundamentals on social and behavioral data, showing unwanted effects that we need to correct for.

However, technical processes also grow in complexity and rely more and more on data obtained from clients or end-users, for instance with the growing emphasis on just-in-time production and personalized services. Looking through the ethical lens that has been sharpened on AI applications in the socio-economical domain, we argue that we should also examine the supposedly neutral applications of AI in industrial and engineering settings to check if we find indirect and unwanted effects. Once we commit to examining the supposedly neutral applications of AI, we identify increasing numbers of examples with unwanted effects. These include charge behavior for electric vehicles, energy consumption in households, water use, package delivery, and ambient lightning programs.

In this paper, we argue that we should readdress these original contexts of AI systems in order to check for direct and indirect biases in the application of AI in engineering and manufacturing applications. Also we discuss the way forward, using a designing perspective to resolve the issues found in the engineering environment.

2 EXAMPLES

As a more elaborate example, consider the charging behavior of electric vehicles. To optimize the energy network in urban environments smart algorithms are made that control vehicle to grid installations. These are installations where a car's battery is used as an energy source for households on time intervals with network shortage/network expensive time intervals or network non sustainable energy time intervals. In other time intervals, the battery is charged. For this, a (personalized) profile is needed of the car owner to make sure that the car is connected at the right time intervals, and the battery will not be empty when disconnected for use. Even in a network setting with no information about the individual car owner, a smart algorithm will find subgroups of charge stations or cars ID's that are save to charge or discharge. The algorithm is smart in the sense that it wants to optimize the energy use goals. As a result the car owner is profiled in time and location. For instance some car ID's or some neighborhoods are known to disconnect cars very early in the morning (blue collar workforce), other areas show irregular nightly car connection patterns (single households) or very long car connections intervals (expats, international travellers). Such patterns thus invade privacy and may lead to unwanted, and unexpected bias in the predictions of the smart charging algorithm.

Along similar lines, the water or energy use of households is closely related to household rhythms. It is known that machine learning algorithms will identify households with for instance alcoholism, medical conditions or dementia and psychological problems from observed daily household energy use or water use.

In the context of charge infrastructure a smart algorithm that aims to maximize the amount of charge volumes on solar energy seems to have a sustainability goal that everyone can agree on. However the system is optimized by changing the speed of current in the charge stations over time. As a result, charging cars may experience variable charging power over time and less expensive electric vehicles are technically incapable of dealing with these changes (Kia versus Tesla). As a consequence the interaction between the the smart algorithm and the hardware of different cars has an unwanted effect of favoring technically advanced, but more expensive electric vehicles.

These are some examples where privacy, profiling and fairness issue do occur in supposedly neutral AI systems. We discuss the consequences and the way forward in the next section.

3 DISCUSSION

What sets the applications of AI discussed above apart from for example recommendation systems or a fitness tracker, is that user behavior is modeled indirectly, as a by-product of a technical optimization process in order to achieve another goal, such as sustainability in case of smart charging, and that there is no intention to model users or collected this data per se. The resulting system behavior not only is a black box for the end users, but also for the service providers.

Part of the problem is that algorithms and technology are traditionally designed in a lab or R&D department, then deployed and implemented, where potential issues only show up after a period of deployment or may go unnoticed altogether.

We argue that if systems get sufficiently complex and make use of more and more data sources, unanticipated system behavior and unwanted side effects will appear with near certainty. Complexity and unexpected behavior may also arise in the interaction between different systems and finally, the embedding of an AI system in a particular context may alter that context. Predictive policing, where sending more police into areas perceived as high-crime might actually result in more crime being reported, creating a positive feedback loop [5], is a well know example of the latter.

For systems indirectly measuring human behavior such emergent behavior may lead to privacy issues, unfairness and a reduced sense of autonomy for the users of the technology.

From a design perspective, this means that we should not only focus on solving the problem at hand, e.g. creating a dynamic charging algorithm aimed at sustainable energy use, but also take the potential impact of the algorithm on different stakeholders into account. Methodologies such as Value Sensitive Design [2] and Reflective Design [7] that point at the dynamics between values embedded in technology and user practice can provide insight how the assumptions underlying a design may be challenged. But how do we design for the unexpected? Or to put it less dramatic, how can we design starting from the assumption that unwanted, yet unknown side-effects may occur? Side-effects that may lead to unfairness or otherwise unethical outcomes, and that may occur only for a subgroup of users as a consequence of the interaction between multiple smart systems. This requires new design methods that aim for *resilience*, ensuring that even in the face of unwanted local effects or interactions, the overall outcome of a service will be guaranteed to provide (a certain level of) fairness.

It might therefore be useful — or even necessary — for designers to adopt methodologies from research in multi-agent systems and game theory. These disciplines study the individual interactions of self-interested agents and the types of emergent behaviour that result from that. Of special interest for the problems at hand would be *mechanism design* that studies ways of setting up games in such a way that even if all agents act purely selfish, specific globally relevant results will emerge. The various flavours of fairness could be some of these emergent properties that are engineered into the system [6].

Another example might be the emergence of socially acceptable behaviour (among AIs or in their interaction with humans) based on partner selection. In [1] it was shown how the possibility of partner selection in the repeated prisoner's dilemma engenders the spontaneous emergence of large sub-populations that act cooperatively and fair, even if their only concern is to maximize their own pay-off. Since it is well-known that prisoners' dilemma occurs in many guises in rational agents, we can expect these theoretical considerations to have practical design consequences.

Taking this standpoint also means that in designing AI systems we cannot stop the design process once the system is deployed and that we should make an effort to understand the potential impact of a system before deployment. This

is not unlike the situation in medicine, where new types of medication are extensively tested on controlled groups and every drug comes with the warning that side-effects may occur and may differ per person. In the same way, experimentation and technology audits in practice are needed for AI based systems. In the case of self-learning AI systems that may change their behavior while in use, this should be a continuous process and this requires automated tools.

REFERENCES

[1] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. 2020. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 7047–7054. https://aaai.org/ojs/index.php/AAAI/article/view/6190

[2] Batya Friedman and Alan Borning. 2008. Value sensitive design and information systems. *The handbook of information and computer ethics* (2008), 69–101.

[3] A. Halevy, P. Norvig, and F. Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[4] Michael Kearns and Aaron Roth. 2020. *The ethical algorithm: the science of socially aware algorithm design.* Oxford University Press.

[5] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy* (first edition ed.). Crown, New York.

[6] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* Penguin Random House.

[7] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*. 49–58.

[8] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. AI Now Report 2018.

[9] Shoshana Zuboff. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (1st ed.).