AI Systems and Continuous Social Impact: The Case for Forethought and Active Maintenance

Emily LaRosa[1] & Heather Douglas [2]

[1] Department of Philosophy, University of California San Diego [2] Department of Philosophy, Michigan State University

Abstract

As Autonomous Systems become more pervasive in our daily lives, and the deployment of such systems ubiquitous within many fields as the 'best' tech to use, developers are being called upon by deployers to envision societal impacts of their systems pre-deployment. However, the issues encountered by society are not always evident a priori, in part because such systems are deployed in complex contexts we call Sociotechnical Biomes. We use the modern threat Autonomous Vehicle Systems pose to pedestrian-way-dependent neighborhoods and said neighborhoods' ability to withstand increased traffic and stresses. We then analyze a case study of a Child Abuse Prediction Model, underlining the issues deployers face with unintended consequences of well-intentioned systems. We articulate a systemic change in thinking about what it means to perform active maintenance of a system.

Autonomous Systems and Their Place in Modern Society

Autonomous systems are becoming both more commonplace and taking on more important functions in contemporary society. If the full benefits are to be realized from such systems, they must be both built and deployed in a trustworthy way. In this paper, we will argue that in order for this to happen, more than public understanding of AI systems and avoidance of bias in such systems is necessary. Trustworthiness will require appropriate governance structures for both the development and deployment of AI systems. We will describe two cases where AI has been deployed that have not gone well, creating trust deficits, and diagnose where weaknesses lay in those cases. Crucial lessons for AI developers will be drawn.

Central to assessing the cases and the trust dynamics in them is understanding AI systems as part of our Sociotechnical Biomes (STBs), best defined as an "integrated ecology of multiple interacting and overlaid sociotechnical systems." (LaRosa and Osoba 2019) An STB, broadly defined, includes the ecological and environmental placements and interactions. Because AI systems operate within STBs, we argue we need both to assess their plausible impact on these systems before deployment (paying attention to both intended and unintended foreseeable consequences) and to actively monitor such systems to assess ongoing (and often unforeseeable) impacts. For this paper, impacts of AI on social systems is our main focus. Governance of the development and deployment of AI systems is crucial to meeting these requirements, and such governance must be part of the sociotechnical biome around AI.

Currently, autonomous systems elicit response in the media and from the public as either a form of novelty or a potential source of danger, as AI interacts with society in a way which the public finds ill-suited to the intended function. This is an inadequate range of responses and a one-dimensional raising of concerns. The public framing of AI needs to move beyond whether or not AI systems function as intended, to consider unintended effects, both those that can be predicted beforehand and those that emerge only in the midst of deployment. Such broader concerns (in the sense of extending beyond the "quality, lack of bias, and traceability of data" that is noted as the basis for trustworthiness in the official G7 Science Academies statement of March 2019) will be central to AI trust. Multiple scholars have proposed different systems to understanding trustworthiness of systems; we will address these scholars, such as Danks and London, as they become relevant to our argument. However, our framing of trust in the case of continuously deployed autonomous systems depends on the meeting of full responsibilities for scientists (Douglas 2003) and on assessing the place of AI within the

STB in which it functions. Trust in a dynamic system depends on communication between key agents within the system, both human and technical; agents must be capable of adaptability; and STBs must be understood as complex, dynamic systems (LaRosa and Osoba 2019).

Impacts on Societal Structure: AVS and Child Abuse Prediction Models

Autonomous Vehicle Systems Deployment: Social Implications

Looking at technological shift as an evolution of our sociotechnical biome, Autonomous Vehicle Systems (AVS) are the most visible indicator species of autonomous system public adoption success. This is because the public will directly interact with AVS and depend upon the system for safety provision. AVS are the most prominent of AI systems engaged publically in society: they both embody the black-box issue (making assessment difficult), and seem to be coming to a street near you regardless of individual desire. Indeed, there have been many social protests regarding these systems' deployments across the country. In addition, full deployment of AVS may require serious infrastructure changes. Though the city-nation of Singapore has the finances in place to restructure their roadways around the technology, the majority of our globe does not (Fischer 2019).[1]

The public needs developers to respond to the needs, constraints, and realities of their lives, rather than developers depending on the public to adapt to their technologies regardless of impact. In particular, AVS must either be able to grapple with the full complexity of transportation actors or be constrained to operating only in predefined and controlled settings. Presently, the majority of governments have been avoiding binding measures of AVS in order to promote development (Taeihagh and Lim 2019). However, this attempt at openness to foster growth of the autonomous system may not always be to the benefit of all, even the autonomous system itself, within the STB. Consider the recent death of an Arizonan woman in the US from an AVS.

On March 22, 2018, an Uber vehicle equipped with AVS functionality was being run in 'self-driving mode'. The vehicle, which had a human driver behind the wheel, was following all safety and speed laws, going 38 mph in a 45 mph zone. A woman, Elaine Herzberg, was walking a bicycle along the roadway next to the vehicle and stepped suddenly into the path of the vehicle seemingly in an attempt to jaywalk. Though there was a human in the vehicle to prevent such tragedies, the human driver did not predict such a collision and did not take manual control of the vehicle. Subsequently, the AVS, which did not brake, struck and killed Ms. Herzberg. The details of this case raise substantial concern about AVS deployment. Because the AVS did not recognize the pedestrian walking a bike as something to be avoided, and the driver was too sanguine about trusting the technology, a rather surprising and fatal accident occurred. The dynamic system of the AVS and its self-driving mode was unable to handle what should have been ordinary complexity. As a result, Uber removed vehicles from the streets, suspending its self-driving testing, for months after the death. Other self-driving pioneers, including Argo-AI, are challenged by this; they are still trying to wrestle with the thorny issue of 'human behavior' and present the world with a system which is seen as societally trustworthy.

This case exemplifies how non-computerized modes of transportation stand to suffer in an environment where an AVS is deployed. If we deploy AVS which does not know how to interact with non-AVS agents and systems, the impact of AVS can be insidious, reaching beyond an uncanny fatality. The routes being planned and taken by these systems, meant to optimize route-time completion, often have a massive detrimental impact on both residential and indigent communities. In using routes which place these systems through communities high in pedestrian population, and making the streets busier than their 'natural' state, AVS are altering the social conditions of these areas. This makes it less safe for those who rely on walking to and from places in areas which lack sufficient infrastructure for pedestrians to begin with. While developers of AVS have often focused on the potential for massive unemployment with the rollout of AVS, systematic impacts on the level of our streets and transportation patterns has received less attention. However, it isn't simply route calculation which is affecting this issue; it's the future challenge posed by the use of AVS to perform 'urban tourism'. These systems will change when, where, and how tourists move (Cohen and Hopkins 2019). We foresee higher potential for tourists using AVSs to increase the traffic in areas ill or poorly equipped to cope with these additional stressors. The complex ethical issues of when and how urban tourism should be pursued remain an open and difficult question.

These instances of tourism and heavy trafficking in a previously isolated or removed neighborhood illustrates how the introduction of algorithmic artifacts (even simple ones) in STBs can lead to foreseeable impacts that require monitoring. Such monitoring should also be attuned to catch the unforeseeable impacts and provide impetus for developers and deployers to intervene on the system to mitigate deleterious impacts. Thus, we need more agility in assessing and responding to risk in our AI-equipped STBs as we may now be unable to pre-identify new harms. Providing for both careful thinking about unintended impacts and monitoring for unforeseen impacts as part of ongoing deployment is crucial for assuring system-level

[1] wiscav.org/singapore-sets-standards-for-autonomous-vehicles/

accountability and trustworthiness. AI or algorithmic artifacts pose a novel challenge for trustworthiness because they are often inscrutable and incapable of responding meaningfully to requests for explanation. Assuring accountability and trustworthiness as we integrate AI systems into our complex societies may not necessarily mean a focus on just human-centric design. Having a trustworthy system in our biome also means having a system which is accountable to its actions and is continuously updated and maintained as it interacts (and sometimes interferes with) our sociotechnical biome.

Child Abuse Prediction Model: Best of Intentions, Worst of Outcomes

Even the deployment of a relatively simple algorithmic system (one that is not inherently opaque like a neural net system) can produce massive unintended societal consequences. Consider the case of the Allegheny Family Screening Tool child abuse prediction model, deployed in Pittsburgh, Pennsylvania.

In November 2016, a new predictive risk model was deployed in Allegheny County, Pennsylvania, with one goal in mind: to prevent recidivistic child abuse and protect children from abuse or neglect. The Allegheny Family Screening Tool (AFST) took in a number of key sources of information, including case notes, demographic information and program statistics of a family, and fed out information intended to help a caseworker screen and assess where a child's case fell along the "risk/severity continuum". The issues of neglect and abuse are thorny ones. Three-quarters of child welfare cases deal with neglect rather than abuse instances, and neglect is often characterized by struggles common to poorer families: not having enough food, having inadequate housing, lacking medical care, or being left alone for long periods (Eubanks 2018). In a state where abuse and neglect are legally narrowly defined, a predictive algorithm has very distinct confines in which to make its evaluations and predictions of abuse.

There were many unintended consequences of deploying the AFST as a child abuse prediction model. As Allegheny County is thankfully very low on deadly abuse, the system needed to use other datapoints (i.e., proxies) to assess risk. The AFST was using such proxy variables to assess child maltreatment as community re-referral (two calls from the community regarding a child who was initially screened out on the first call within two years) and child placement in foster care within two years. As noted above, the requirements for mandatory reporting by community members to call in to the child welfare system are often tied to poverty. The AFST, rather than modeling abuse, inadvertently modeled which families would be reported or reflagged, and thus increased the likelihood that poor children would be removed from their homes. There was a spike in false positives amongst families who were poor, rather than in cases where a child was likely to have been actually abused or neglected (Eubanks 2018). Due to the paucity of data of actual abuse cases, the AFST was not, as was intended by deployers, assessing and modeling child abuse or neglect. As a result of this system's deployment and overtrust, Allegheny County also saw increased community surveillance of and interference with impoverished families who were improperly flagged by this system, stressing these families while poorly allocating resources of an already over-extended Children, Youth and Families (CYF) Services. Thus, the system harmed poorer families—when these families are those being over-sampled and contributing to the accumulated 'predictive variables' that correlation to child maltreatment, it follows that we will see a recursive assessment. The system also failed to increase protection of children in Allegheny County as intended by deployers.

With the information we have, and the pressing issue at hand of needing to prevent child abuse before it occurs, we then need to assess: is it possible to alter or tweak the algorithm, in light of its failures, to meet the real needs of CYF? We have seen a lot of correlations between poverty and child abuse recidivism, but how much is poverty a causal factor in child abuse? Undoubtedly, the family which is placed under increased surveillance is therefore under much more stress and duress; how much of the system-advocated surveillance actually caused the abuse it was intended to prevent? These questions are troubling, but point to one concrete concept—active maintenance matters equally if not more so than a priori consequence assessment when a system is intended to have a societal impact.

Post-deployment: Passive v. Continual Active Maintenance

Autonomous or algorithmic systems deployed in a dynamic real-world environment present a potent challenge to developers. In addition to the intended purposes of such systems and foreseeable impacts, there must be built into the deployment the active monitoring of the system, in order to assess impacts unforeseen and unforeseeable. We must update our own modus operandi when new facts or concepts come into light; so too must developers update their systems. Developers must act as the external consciousness and conscience for their learning systems. These AI are often seen by the public to be robust enough to be held accountable for their actions—even if they are not yet so developed. Indeed, we are unsure what it would mean to hold such a system accountable. Instead the trustworthiness and accountability of the system must depend on the responsible development of the system by developers and the active and responsible maintenance of the system in deployment.

Developers are responsible for foreseeable unintended consequences, but not of the unforeseeable (Douglas 2003). Because of the concern over the systems' robustness, and of the unpredictable 'messiness' of the environments of the real world in which they are deployed, we must have protocols in place which ensure active maintenance post deployment of systems which interact with the public and have an impact on the societal structure or well-being. A closed-loop system without much additional input or interaction may not be beholden to such measures; however, as we have discussed, the majority of systems which we envision as 'best practice' systems in societal settings do in fact make decisions which impact the lives of those within the society in which they're deployed. The nature of these systems may end up producing negative consequences (Danks and London 2017).

Moving forward in seeking out potential issues in an active deployment scenario, developers and deployers alike must be cognizant of the trap coined well by Stegenga—"The Hollow Hunt for Harms." "Power is normally thought to be pertinent to detecting benefits of medical interventions. It is important, though, to distinguish between the ability of a trial to detect benefits and the ability of a trial to detect harms." (Stegenga 2016) Active maintenance and attention to the impacts of your system may look for what you're worried about initially, but if you don't keep your eyes open to the actual ongoing impacts, you won't catch the true problematic impact at play. For example, if developers are only looking for physical structural harm to the infrastructure on which AVS operate, or what form of property damage AVS may cause if they don't read road signage properly, developers may not realize they are upping traffic in residential neighborhoods and causing a change in the social fabric of the area by preventing children from playing outside, or keeping cyclists from using the roadways.

Conclusions

Developers must retrain their way of thinking about their dynamic systems. Continued Active Maintenance (CAM) is the only way to ensure that a system is responsibly deployed in its Sociotechnical Biome, and presently is the best measure at hand with which to cope and react to unforeseen social consequences of system deployment. We therefore maintain that:

1. Foreseeable unintended consequences must be considered pre-deployment

2. Unforeseeable unintended consequences cannot be considered a priori

3. Active maintenance ought to be pursued when developers deploy autonomous systems which (i) operate in any public space, or (ii) operate in ways which the public cannot opt out of, or (iii) operate in ways intended to have a substantial societal impact.

If the developers work under governance structures that encourage the robust meeting of these criteria, a central component of trustworthiness critical to justifiable AI deployment will be met: namely, there be active maintenance as well as attention to foreseeable impacts of the systems deployed in the Sociotechnical Biome.

References

Cohen, S., and Hopkins, D. 2019. Autonomous Vehicles and the Future of Urban Tourism. *Annals of Tourism Research* 74: 33–42. doi.org/10.1016/j.annals.2018.10.009.

Danks, D., and London, A. J. 2017. Regulating Autonomous Systems: Beyond Standards. *IEEE Intelligent Systems* 32(1): 88–91. doi.org/10.1109/MIS.2017.1

Douglas, H. 2003. The Moral Responsibilities of Scientists (Tensions Between Autonomy and Responsibility). *American Philosophical Quarterly* 40(1): 59–68. jstor.org/sable/20010097.

Eubanks, V. 2018. A Child Abuse Prediction Model Fails Poor Families. *Wired Magazine* 01.15.2018.

Fischer, R. 2019. Singapore Sets Standards for Autonomous Vehicles. *Wisconsin Automated Vehicle Proving Grounds*.

G7 Science Academies. 2019. Artificial Intelligence and Society. *Summit of the G7 Science Academies: Executive Summary and Recommendations*. French Academy of Sciences: G7 France.

Hurley, D. 2018. Can an Algorithm Tell When Kids Are in Danger. *The New York Times Magazine* 1.07.18: 30.

LaRosa, E., and Osoba, O. 2019. Trust in Sociotechnical Systems: Case Studies and Analyses. Paper presented at the 9th International Conference on Technology, Science and Society. Madrid, Spain, October 3–4.

Stegenga, J. 2016. Hollow Hunt for Harms. *Perspectives on Science* 24(5): 481–504. doi.org/10.1162/POSC_a_00220.

Taeihagh, A., and Lim, H. 2019. Governing Autonomous Vehicles: Emerging Responses for Safety, Liability, Privacy, Cybersecurity, and Industry Risks. *Transport Reviews* 39(1): 103–128. doi.org/10.1080/01441647.2018.1494640.