A.I. Ethics in the City

Marc Böhlen(✉)

Emerging Practices in Computational Media, Department of Art,
University at Buffalo, Buffalo, NY 14260, USA
marcbohlen@protonmail.com

Abstract. Artificial Intelligence (A.I) is the science of making machines act in a way human beings might call intelligent. Ethics of A.I. seek to make A.I. behave responsibly. A recent surge in national and international initiatives in support of A.I and A.I ethics speaks to both the significance of as well as anxieties surrounding the project. This short text is only a sketch with more questions than answers. It points out two factors in the A.I. ethics debate that receive less attention than they should: the significance of situatedness and of decision mechanisms in A.I. The big city, I contend, will be the site where complex A.I ethics conflicts will emerge most prominently; in different ways in different parts of the world.

Keywords: Artificial Intelligence · Ethics · Urbanism

1 Cities as Sites of A.I. Experiences

The city has always been a site of density, of unlikely encounters rendered possible, of voluntary and involuntary exchanges. Part of the lure of the big city is precisely the combination of controlled and uncontrolled events. And big cities today are rich in ingredients that A.I. needs; human capital, data flows and problems to solve.

In the not too distant future, A.I. enabled cities will be blanketed with sensors capable of recording all forms of ambient information, from traffic to biological events with high spatial and temporal resolution. The future omniscient A.I. running the networked city will know, for example, how much water flows to each of its public fountains on any given day. And it will be able to choose precisely which of the fountains to turn off in the event of water shortage during a hot summer.

Contagious diseases no longer go undetected in the A.I. controlled city. Public transportation, armed with these new devices and A.I. algorithms controlling them, keep ever anxious big city dwellers abreast of all contagious hotspots in the city in real time. Optimized against under-reporting a threat, the algorithms are programmed to respond to every possible pathogen detected, even the common cold. Riding in a bus with a fellow sneezing commuter is no longer a shared acceptable risk of life in the city. The offending passenger can be detected, and the danger she poses shared with fellow travelers in real time. In the past, social rules and established but unwritten ethics of everyday behavior dictated that we do not outcast a person just because of a sneeze on a bus.

M. Antona and C. Stephanidis (Eds.): HCII 2019, LNCS 11572, pp. 13–27, 2019.
https://doi.org/10.1007/978-3-030-23560-4_2

A.I. will make the city more controllable; issues left to chance in the past will become decidable. How might the citizens of Rome respond to A.I controlled water fountains during a drought, or the citizens of Hyderabad to biosensing enabled and A.I. controlled public buses during a flu epidemic?

2 A.I. Ethics en Vogue

A.I. tries to efficiently make sense of the world. But A.I. efficiency does not translate to morality, and efficient sense making can cut multiple ways. Leaving out details that might not seem important for algorithm design can mean missing nuance that only becomes apparent when the algorithm operates in the field. The opposition between the power of abstraction and loss of nuance has remained a source of tension in A.I. design and ethics from early experiments to current products.

Some applied A.I. domains, including Ambient Intelligence, have tried to include ethics in design parameters of computing systems, but those approaches, laudable as they were, never scaled to mass deployment. More recently, the topic of A.I. ethics has found resonance again, this time however fueled by unease produced both by spectacular successes in language and image processing as well by ever more glaring A.I. snafus such as biased facial identification [1].

Criticism of A.I from consultants [2], business celebrities [3], and activists [4] has become too intense to ignore even by the most ardent supports of A.I. In response, private industry consortia have taken a lead in A.I ethics debates in order to prevent A.I. from becoming a liability; securing a favorable regulatory framework for A.I. development is of paramount importance to the industry. The fact that there are at the time of this writing over 50 distinct efforts underway to support, oversee and regulate A.I. systems is a clear sign of the general perception of the significance of, and the recognition of the scale of the problems A.I. is generating.

Indeed, not all of the 50 efforts are new. The Organization for Economic Co-operation and Development[1] was a pioneer in the formulation of guidelines for trans-border privacy protection almost 40 years ago and is now is in the process of proposing succinct policy and institutional frameworks to guide A.I. design and use across the planet. The inclusion of the global dimension of A.I. is shared by a variety of other initiatives, each with its own specific focus (the European Union's Communication on Artificial Intelligence in Europe, and the UNESCO World Commission on the Ethics of Scientific Knowledge and Technology, amongst others). Some of the private stakeholder initiatives (The Future of Life Institute; OpenAI; Cambridge University's Center for the Study of Existential Risk; the University of Montreal's Declaration for a Responsible Development of Artificial Intelligence as well as the IEEE's Ethically Aligned Design initiative) combine efforts to manage current A.I. products while considering future A.I technologies far more capable than even the best systems currently in operation.

[1] See www.oecd.org/going-digital/ai/initiatives-worldwide/ for an updated list of A.I initiatives.

Many countries view the development of A.I. as a way to increase competitiveness and strengthen national security. The term A.I. nationalism [5] has been coined to denote precisely this nation-first approach to A.I., and several existing and emerging A.I. powerhouses have openly declared their nation-specific aspirations, including China (New Generation of Artificial Intelligence Development Plan); Finland (Age of Artificial Intelligence); France (Strategy for a Meaningful Artificial Intelligence; Italy (Artificial Intelligence at the service of the citizen); Japan (Artificial Intelligence Technology Strategy), as well as the USA (National Artificial Intelligence Research and Development Strategic Plan). China, already a leader in voice and visual recognition technologies, has likely crafted the most ambitious plan [6] for an A.I. fortified future, seeking by 2030 global A.I. dominance in all relevant domains including defense, urbanization, industrialization, social governance and – where necessary – public opinion guidance [6, p. 28].

In addition to the nation-specific interests, many of the current A.I. initiatives respond to popular and current themes such as privacy. Given the current culture of data misuse large and small, it is of little surprise that damage control is applied and that the topics of privacy and security find prominent attention. Certainly the recent history of data theft has brought this part of data culture to the forefront of public scrutiny. And the otherwise desirable ability to learn lessons from large datasets is rapidly becoming a liability when datasets are acquired illegally.

It is also not surprising then that harm prevention would be coupled to the issues privacy and security. The proactive attempt to prevent personal data from being compromised in the first place is the better way to address data security; handling after data breaches occur is mostly ineffective. And harm prevention, applicable to a general public, leads in turn to additional shared themes across the initiatives, namely equity and fairness. Even the concept of well-being is included in some of the initiatives. If nothing else, the inclusion of well-being is a clear indicator of the hope many stakeholder hold for A.I. as a potential source for good. The positivist stance also functions as a powerful rhetorical antidote to evil A.I. that periodically dominates popular debates on A.I. futures.

Specifically the industry supported A.I. initiatives emphasize ethics challenges of future A.I. over those of current A.I. systems. While even the most advanced systems in operation today are probably no indicator of what next generation A.I. will be able to achieve, these superhuman A.I. systems will be so powerful that ethics might not be the most important topic to consider. Indeed, careless mingling of current and future A.I. concerns weakens arguments in support of A.I. governance right now.

Engineers have in the past, at least in principle, been bound by professional standards to act ethically[2], and many A.I. relevant questions can be addressed by existing frameworks. Without questioning the need for deep inquiry into ethics of A.I. at large, some of the most urgent problems can be addressed now: The need to be honest with the general public about what a given A.I. system actually does and the limitations the system has; the acknowledgement of slippages that occur when an A.I. system is

[2] See Accreditation Board for Engineering and Technology http://sites.bsyse.wsu.edu/pitts/be120/Handouts/codes/abet.htm.

transferred from one domain to another; control or prohibition of reselling data, etc. Granting A.I. special status because it can be a powerful agent for change weakens calls for constraints to follow basic and established rules of conduct.

3 A.I. Ethics Shapeshifts

Technology-centric approaches to A.I. ethics are often inadequate to address the paths along which A.I. systems impact people. An early yet telling example [7] describes a robotics engineer working on a sophisticated anti-tank mine capable of hopping out of the way of unintended targets such as pedestrians. Certainly one might be tempted to declare this agile robot an example of an ethically designed intelligent machine. Not quite. Because of the mine's clever control system, the device was not classified as a mine, but as a robot and consequently not subject to anti-mine regulation, clearing it for wide distribution as a 'smart' defense device, undermining the procedures designed to curtail the deployment of such weapons in the first place.

Similarly, an A.I. system may become unintentionally harmful with skewed training and questionable design priorities. Imagine service chat bots designed to assist visitors to the US in navigating American customs requirements at ports of entry. With the intent of creating a lifelike experience, the bots are trained on the language habits of American customs officers, who - maybe without malign intent - exhibit threating language use. And in order to make the system efficient, visitor response times are limited to a few seconds. The resultant synthetic rudeness is an unanticipated consequence of poorly selected training sets and pressure to optimize for speedy interaction.

In order to get some understanding of the landscape of unintentional A.I. ethics conflicts, and why cities are likely to become prime sources of generating and experiencing these conflicts, the next section will revisit an early example of a technology-supported intervention into the urban fabric as well as some of the ethical issues that ensued.

4 Road Pricing – Early Technology Ethics

Generally the territory of economists and urban planners, road pricing might seem at first an unlikely candidate for A.I. ethics insights. Yet because road pricing has been applied to many large cities and because scholars have analyzed in depth the forces at work in road pricing projects, a second take is warranted.

A.I. systems, like road transport, impose negative externalities on society [8]. In the case of road transport these externalities include congestion, accidents and pollution. Furthermore, road pricing infrastructures are complex control systems that include data collection and interpretation on an industrial scale. Additionally, road pricing systems are policy dependent, technically enabled responses to urban traffic management and revenue generation needs, as well as responses to calls for fairness in resource use. In A.I. the negative externalities include compromised privacy and constant surveillance, with other factors likely to join the list as the field stabilizes. As such, the dynamics surrounding road pricing interventions parallel several of the issues in A.I ethics.

Road Pricing in Singapore

Singapore was the first country to introduce road pricing with the expressed goal of reducing congestion in the city. Early versions based on manual paper licenses granted access rights to select parts of the city [9]. Additionally, the government discouraged cars from entering downtown with increased parking fees and substantially improved public transportation and park&ride nodes. However, the manual system was labor intensive and unable to charge vehicles per entry and unable to adjust prices rapidly.

In response to these shortcomings an electronic road pricing system (ERP) was introduced in 1998. Vehicles were required to install in-vehicle gizmos with cash cards, and each time a vehicle passes under one of the many overhead gantries installed in rings around the city, the ERP charges are deducted from the cash card via short range radio. The communication is one directional, meaning the gantry tells the in-vehicle unit to charge the cash card rather than the in-vehicle unit signaling the account ID to the gantry. This provision anonymizes the use of the ERP unless a transaction fails, in which case a camera captures the license plate of the vehicle.

Even though the system is designed to operate as an income generator, annual revenue produced by the ERP is about $S150 million [10]. Prior to the actual launch of the ERP, the government of Singapore launched a mass publicity campaign to 'inform and educate' motorists [9]. The campaign included articles in newspapers, on television and radio as well as pamphlets with explanations on how the system works, the location of the gantries and shops to have the in-vehicle control units installed. Singapore's ERP was launched with a three month test period during which no charges were made, allowing a time limited grace period to soften the transition to the subsequent uncompromising oversight.

By traffic control measures, the ERP can be considered a success. It reduced traffic into the downtown core by at least 14% [9]. Yet the Singaporean ERP uncovered otherwise unexpressed anxieties of city dwellers. Complaints filed by users centered on confusion with the pricing structure, the kinetic effects of in-vehicle gizmos in car accidents, radiation exposure from the gantry system, the potential of being tracked, and the fear of being falsely tagged as a violator should some part of the control system fail [9].

Road Pricing in Stockholm

In Stockholm, road pricing followed a quite different trajectory. In Stockholm, the rationale for congestion pricing was based not only on a desire to reduce traffic in a downtown area isolated by waterways but also on a desire to reduce emissions produced by automobiles, a concern not explicitly mentioned in the Singapore case.

The Stockholm congestion tax varies (as others do) based on time and does not depend on direction of traffic in a cordon around the central city [11]. The Stockholm technical model is a fully automatic fee payment system enabled by overhead gantries outfitted with computer vision license plate recognition. No in-vehicle devices nor pre-travel purchases are required. Vehicle owners are sent a monthly bill based on the number and location of license plate recordings.

The history of the project is telling. Congestion charges were introduced in Stockholm first as a trial in 2006, followed by a referendum, and then made permanent from 2007 on. Similar to the Singapore approach, the intervention in Stockholm

included an expansion of public bus and train transport, bicycle and pedestrian safety improvements as well as the construction of additional park&ride facilities [10]. The traffic centric debate on the merits of the experiment were paralleled by political maneuvering. A then small interest group, the Green Party, lobbied in favor of the trial in exchange for its support for a national social-democratic government [12, p. 2]. This horse trade in turn negatively impacted public perception of the congestion charge concept. Only after the trial period started did public sentiment turn in favor of the effort as the benefits of the interventions became apparent to individuals and the public at large. Media reports that initially focused on rallying against the costs of road pricing started fantasizing about how the revenues could be put to use [12]. As in Singapore, the benefits generated by the Stockholm system are measurable: 20% decrease in congestion, 14% reduction of CO2 [12] and about 820 million krona in revenue in 2015 [11].

5 Social-Technical Systems in Context

Road pricing has not been a success story in every urban context in which it has been proposed or introduced. Hong Kong, for example, decided not to adapt electronic road pricing after a large study in the mid-1980s. Not that congestion mitigation was no longer necessary, but traffic reductions occurred through a confluence of other factors, including a new expressway, a stock and property market crash, the opening of a new mass transit railway as well as doubts of the financial viability of the project. The fact that Hong Kong was then slated to return to Chinese rule might have been an additional factor in the city decision not to introduce a large scale surveillance system [13].

Moreover, road pricing is not an urban intervention that produces binary results in every case. In London, for example, traffic has increased after congestion pricing was introduced. But that increase in turn is a function of several factors, including a reduction of road capacity and the introduction of ‘exceptions’ to the pricing regime [11] for select stakeholders.

The Stockholm case is interesting not because of the novelty of the deployed technology. In fact, license plate recognition was already then a solved problem. What makes the case interesting for the discussion of future urban A.I. ethics is the original oppositional public sentiment, the way it changed over the course of the trial period, and how the political rationality overshadowed the technical debate as the following excerpt demonstrates:

> “Political rationality of congestion pricing may be different from mere public acceptability. While public support certainly affects political actions, it is neither a necessary nor a sufficient criterion for political support for a policy. … Purely technical rational questions, without a moral dimension or interpretation, may not generate sufficient voter enthusiasm to make them worth any political risk. During the debate, congestion pricing was to a large extent proposed and opposed with moral arguments rather than technical-rational ones in a more limited sense. This line of argumentation may have been necessary to make congestion pricing politically interesting – but may simultaneously have made it a more divisive issue.” [12, p. 3]

If oppositional public sentiment almost derailed the Stockholm project, the Singapore ERP never faced such scrutiny. First, because it was created as a successor to an

existing and inadequate system, and second, because Singapore's East Asian *electoral authoritarianism* [14] gave the government some leeway to not demonstrate the levels of concern towards public sentiment other democracies might exhibit. Moreover, the implementation of the ERP was by all accounts well organized and produced a smoothly operating system with measurable results.

Road pricing is an example of managing access to a limited, shared urban resource. By and large, the merits of managing limited resources in urban contexts are understood. Problems occur in the distribution of the costs and disadvantages of the intervention. What we can learn from the Singaporean and the Stockholm traffic management examples is that the implementation of a large social-technical system unfolds differently in different cultural and political contexts. The discussion of the merits of the system depend both on its actual performance as well as the way public debate in general occurs. As opposed to road pricing where the costs and merits are comparatively clear, the merits and costs of urban A.I systems are at this moment unstable. Moreover, the paths along which the merits and costs are formed are more complex than in even the newest road pricing systems.

6 Digital Monopolies – A Prelude to A.I. Ethics Conflicts

Road pricing foreshadowed only some of the dynamics of how large scale technical systems in the city are built and perceived by the public. This following section will shift from the automobile to the city dweller and describe how two current urban revitalization initiatives propose to work with large scale urban data flows, and point to some of the debates these efforts are generating.

Sidewalk Labs Toronto

The organizers of Sidewalk Labs [15], part of a large globally operating technology company, want to define, in their terms, what the city of the future might look like. With a local partner, Sidewalk Labs has created a master plan for Toronto's Eastern Waterfront. Promising a people-first design approach, the planners have laid out grand ambitions: sustainability writ large through new construction technologies to create a climate-positive and an inclement weather sensitive neighborhood with mixed income housing; a transit system that combines people, vehicles, freight and garbage disposal; and last but not least data-enabled social services.

Indeed, not only the social services are data-enabled, but every aspect of this new neighborhood is data-driven. Lamp posts, park benches and designer trash bins are optimized for this new urbanity, and they continuously, relentlessly, collect data [16]; data on bench usage, garbage production, dogs marking lamp posts, no event goes to waste. Yet for all its attention to the details of data collection, the original Sidewalk Labs model made no provision for how this new urban resource might be shared with the very people producing it.

Opponents to the plan openly rallied to prevent the new proprietary data infrastructure from becoming effective [17], forcing Sidewalk Labs to agree to declare "privacy as a fundamental human right" [16]. This major shift in approach however has yet to find a path into the details of implementation and oversight. For example, no one

outside of Sidewalk Labs understands precisely what the organization will do with the data, which algorithms it will deploy on it and what downstream events it might enable. These are not minor concerns as they can materially impact quality of life along numerous trajectories. For example, will it become harder for a city resident to find an apartment if Sidewalk Labs detects deficiencies in her trash disposal habits? And what would happen if the transgressions were in fact those of a neighbor, and falsely attributed to another person? What kind of recourse could one take? How long would even minor transgressions be stored in databases no citizen has any form of access to?

This ongoing debate makes apparent an increasing discomfort with the cavalier ways digital monopolies collect and process data from people without their consent. The term *surveillance capitalism* has been coined to describe an 'emergent logic of accumulation' [18] that shows itself in the data extraction and control logic of the Sidewalk Labs project where fundamental human rights for data access are included in vague statements only. The opaque data regime envisioned by Sidewalk Labs in Toronto impacts not only the collection of data, but its processing and afterlife. Transparency on the details of data flows are precisely what would make it a meaningful right. Without rigorous transparency, the right simply does not exist.

The End of the Smart City

Sidewalk Labs is an example of a skewed urban concept; a smart city that places a premium on efficiency and in turn reduces the agency of city inhabitants. In response to the disappointment with the smart city concept, some optimistic urban planners and academics coined the term the *responsive city* [19] to describe an alternative notion of software-supported urban existence. In the responsive city, governance is imagined to have all the benefits of information technology enabled efficiencies but without the top-down hierarchy of the smart city. Proponents of the responsive city model hope for administrative efficiencies generated by digital tools that 'sweep away frustrations' and 'free up talent of government workers' [19, p. 6] while enabling new vectors for civic voices [19, p. 52] through open data platforms.

Digital City Barcelona

In Europe, there are several projects that attempt to build from the ground up responsive alternatives to a hierarchical data control model, and Barcelona is currently one of the most prominent examples [20]. Barcelona's effort in finding alternatives to digital monopolies is part of a response to Spain's recent history of austerity politics and the consequences it created for the city. Barcelona en Comù is at the forefront of a new generation of political movements radically opposed to digital innovation without citizen participation. Examples of how en Comù interprets its social mission can be seen in a cooperatively organized internet platform specifically designed to allow citizens of Barcelona voice their preferences and priorities from a list of possible city projects, for example.

One of the main concerns of Barcelona inhabitants has been tourism, in particular the effects of platforms such as Airbnb whose activities have reduced affordable housing options [4]. Furthermore, Digital City Barcelona is invested in building its own software with open source offerings produced by smaller locally operating entities and in adapting the development of these technologies to the changing needs of communities. All government produced datasets reside under sovereignty of the citizens of the

city and build the basis of the city's digital commons. One of the stated goals of this approach is the creation of alternatives to the dynamics of the on-demand economic model with a focus on established and new models of sharing.

7 From Fair Data Practices to Ethics of A.I. in the City

The efforts of Digital City Barcelona and other responsive city initiatives cogently address pressing issues of data management in the city. Yet digital literacy, participatory and egalitarian practices and open data access do not speak to the new conditions created by algorithms capable of making decisions without human oversight. It is this class of automated actions that create a paradigm shift, with direct effects on the experience of urban life; algorithmically defined actions are what city dwellers will experience in the A.I. controlled city of the future.

In order to better understand why computer produced decision mechanisms are uniquely problematic and how they impact future A.I. ethics in the city, we take a short detour.

A.I. Makes Decisions

Computers make decisions differently than human beings do. They lack intuition, but can be trained to recognize patterns in large datasets with astonishing results. Supervised learning, learning by instruction – typically from a human being – allows a computer to acquire information without understanding the origin or meaning of the materials at hand. Under supervised learning, a computer builds internal representations of the input and then seeks to detect these patterns in new examples. Today, neural networks are the most expressive models for supervised learning. Neural networks are computational models inspired loosely by biological networks made up of nodes selectively modeled after biological neurons. Each neuron-node receives several inputs, takes a weighted sum over them, passes it through an activation module and responds with an output. And each node is in turn connected to many, many other nodes across multiple layers.

The neural network model has proved to be surprisingly successful in several areas that earlier A.I. methodologies have struggled with: autonomous vehicles, language and image analysis being the most prominent success stories today. Together with faster processors and large data sources, neural networks together with reinforcement learning have come to redefine what computers can achieve, including beating human beings at the game GO [21] even without instruction from humans, by force of trial and (learning from) error.

And A.I. Makes Mistakes

Pedro Domingos does not believe that computers will get too smart and take over the world, rather he believes that they are "too stupid and have already taken over the world" [22, p. 286]. Part of this 'stupidity' in real world situations is due to the specific goals of computer science, namely that of creating general approaches to a problem. A good search algorithm should be able to handle multiple search problems. But in order to achieve this ability to generalize, an algorithm must abstract out details irrelevant to the algorithm yet possibly highly relevant to the real world problem one

actually wants to solve. This cleaning up of a real world problem to match the requirements of an algorithm can entail the loss of significant social context and nuance.

Historically, computer science is a child of mathematics and inherits its commitment to abstraction. Even the most basic operation in mathematics, the assignment of number, is an abstraction from reality that sacrifices nuance. But that sacrifice is fully intentional for it allows number to be disassociated from objects. Or, as Alfred Whitehead put it, "the first man who noticed the analogy between a group of seven fishes and a group of seven days made a notable advance in the history of thought" [23]. Without abstraction, the existing foundations of computer science would collapse. Yet when computer science applies itself to real world problems, the blessing of abstraction that give it enormous reach can become a curse.

In the case of reinforcement learning supported neural networks, the contrast between desired and effective operation follow unique patterns. When the output of a learning network produces a result that does not correspond with what the human designer intended, computer scientists speak of an *alignment problem* [24], and it is due to the fact that the optimizing efficiency of an A.I. algorithm cannot guarantee that values aligned with it are maintained. In practice, a machine learning accident occurs when a human had in mind a certain task, but the system designed for that task produced unexpected and possibly harmful results [25, p. 2]. In reinforcement learning, the disjoint between machine action and human expectation typically occurs in the behavior of its ulterior goal, the *objective function*.

Imagine an A.I. enabled drone designed to monitor crowds in public areas. As opposed to drones that use deep learning to attempt to recognize violent individuals [26], this technically more sophisticated and politically more ambitious drone seeks to prevent violence from occurring in the first place. Accordingly, this drone's objective function is formulated to prevent excessive crowd densities that precede the onset of crowd violence. Imagine this drone being deployed in New Delhi during a street protest of farmers, a recurring category of public protest in which farmers call for financial support for their crop production in the face of rising prices, a condition so severe that it results in dozens of suicides per day amongst farmers in rural India [27].

While this drone is not specifically programmed to prevent peacefully protesting farmers from stopping for refreshments at the side of the road, it might be compelled to do so by *reward hacking*. Large groups of people queuing in line could trigger the drone to act against excessive densities of pedestrians. So the drone might attempt to disperse the unassuming crowd wanting nothing more than a drink of water. This class of negative side effect occurs when an objective function over-focuses on one aspect of its environment while not paying attention to other aspects [25, p. 4] such as the presence of a refreshment stand in this case.

Likewise, the tenet of *scalable oversight* could be violated when the drone is unable to make proper decisions in situations that are too expensive to evaluate during training. And what should be the limits to exploration – should the drone be allowed to fly off to a different part of town to inspect a traffic accident and missing the possible onset of a mass panic at the street protest? Or in a case of *lack of robustness to distributional shift* the drone's crowd violence detection algorithm might have been trained on video segments with smaller groups of people at daylight, and then deployed

for a particularly large group of farmers carrying candles at night time, rendering the drone unable to make any meaningful assessment.

No doubt there are methods by which many of these conditions can be addressed and the drone's actions aligned with its original goal and the intention of its designers. In this illustrative example, one might train the learning algorithm on many mass gatherings across seasons, during daytime and at night, to expose the algorithm to data of multiple crowd distributions. Or one could attempt to integrate human judgement directly into algorithm training [24]. The point of this simple example is not that any one of these actions can be fixed, but that, taken together, they constitute failure categories that occur within normal operation, not as a consequence of a malfunction or external influences. These are failure modalities other technical systems do not have, and they require a reformulation of what it means to require an algorithm to behave ethically.

One might argue that anxiety is uncalled for. After all, the automobile industry developed safety systems such as the seatbelt, airbag and lane departure warning only after automobiles – unsafe as they were at first - were launched to mass urban markets. Yet the A.I. case is trickier. As autonomous robots become ubiquitous and the A.I. systems controlling them increasingly complex, there is simply a higher chance that A.I. systems produce harmful actions while trying their best to efficiently execute prescribed tasks. And because the problems A.I. systems are asked to solve can be to too difficult, too expensive or too cumbersome for human beings to deal with, these efficient A.I. systems will be deployed. And because there is a race to get powerful (nationalist) A.I. systems into operation before others do, these A.I.s will be deployed even before all the kinks are worked out. As the harm vectors described above are not intentional and likely only detected after they have occurred, the algorithms creating them pose a conundrum to any technology ethics framework seeking to minimize harm.

8 A City Is Not a Laboratory

The outcome of an A.I. system is not just dependent on its design, or how well the technology can perform, but how people respond to its decisions. What works smoothly in a laboratory on synthetic data might not work at all in the wild, and the wildest site for A.I. to be active in is the city.

As a recent case in Boston [28] showed, good A.I. solutions can be defeated by factors completely unrelated to the A.I. itself. In the Boston case, an A.I. system was able to produce a new school busing schedule that changed start times at dozens of schools, rerouted hundreds of buses ferrying students to their respective schools, trimmed the city's transportation costs and shifted the majority of high school student into later start times while offering historically disadvantaged neighborhoods more desirable pickup schedules. Digital transformation of government at its best. So it seemed. But the new schedule proposed shifts in school start times by two hours and more in some cases. Angry parents who were content with the old system and newly disadvantaged by the new 'fair' system made their voices heard and produced a crisis for the city administrators so pressing that the initiative was canceled; the A.I. produced optimal solution suffered political defeat.

As the brief excursion into pre-A.I. electronic road pricing showed, different cities with different cultures respond in unique ways to the opportunities of actively managing individual traffic flow to reduce congestion. Similarly the introduction of A.I. enabled autonomous vehicles (AVs) can be expected have different practical effects and generate varied ethical dilemmas depending on the details of where and how the interventions occur, and the scale at which they occur. Thousands of single occupancy, pricy AVs zipping around a city in designated green traffic lanes with relaxed well-to-do passengers enjoying the view will not solve urban congestion but will foment new experiences of inequality. Without restrictions on AV distribution and occupancy, the convenience of AVs for individual consumers could outweigh the benefits for the public [29], creating a new category of efficiently managed, clean energy enabled unfairness.

In addition to transportation, housing has been a topic of responsive urbanism and A.I. driven innovation attempts. In the US for example, several projects have been launched to alleviate specific aspects of the housing crisis, including streamlining rehabilitation of dilapidated houses through the application of data science to the identification of homes in need of repair [30] and the use of industrial fabrication techniques to lower the cost of housing stock. Critics of technology centric housing crisis approaches point out that housing is a policy problem, one exasperated in the US by stagnating incomes and rising housing costs [31]. Only where zoning laws allow for the placement of low cost units produced by novel fabrication techniques can the social intervention be effective.

The disjoint between A.I fueled hope for a new approach to social equality and the real world constraints opposing change create a new arena for A.I. ethics. This condition is recognized at least by some major technology companies [32] that have recently promised substantial investments in addressing housing needs in their own cities without relying on technical wizardry. Instead, funding has been allocated towards supporting affordable housing developers and new construction initiatives.

9 Megacities as Feeding Grounds for A.I

Over half of the world's population now resides in cities, and some of these cities are growing to immense sizes. The current collection of 33 megacities each with populations of over 10 million is expected to expand to over 40 by 2030 [33]. These megacities will produce massive concentrations of human capital and services, and immense flows of data that urban A.I. systems will not only analyze but use to control the flow of goods and anticipate future events. The combination of intensity of human activities and data creates abundant opportunities for and pressure to develop A.I for urban conditions, and thus the highest concentration of conflicts for A.I ethics to grapple with.

Historians of technology remind us how much effort is required to learn to live with novel technologies. The advent of the telephone, the first electric medium to enter the urban home, created serious confusion in coming to terms with the new dimension of human presence [34], and the emissions of the first automobiles were hardly a cause for concern. There is no reason to assume that our current assessments of A.I. systems will

not change. Indeed, they will change simply because A.I. changes, differently and more radically than other technologies, morphing and improving itself as it learns from new data or its own actions.

This is the context in which the recent surge in A.I. initiatives can be understood. There is no lack of lofty goals for A.I. and ethics; from harm prevention to solidarity, fairness and equity to well-being the list is packed, yet the path forward is not clear. The good news is that some technology companies – possibly in response to public pressure - are making efforts to address A.I. ethics violations. In one case an effort was launched to counter bias in facial identification software datasets [35], and in another case the details of an experiment in natural language generation [36] were kept secret because the results were too good. If made public in all its details, the software and training sets might be misused to generate fake news or worse.

Particularly the second case is an example of an effective approach to practical A.I. ethics at this time. *If in doubt, refrain from deployment*; fresh wind in the face of prominent long-winded A.I. ethics initiatives. Yet this can only be a start. More concrete responses to the new A.I. harm vectors should emerge, and any regulation and oversight mechanisms put in place now should remain adaptive to the fluid landscape of A.I. development.

One step towards meaningful oversight could include an honest account of problems industry is encountering as it develops and rolls out the newest A.I. systems to the public. It would be helpful if A.I. creators and providers made available a catalogue of mistakes, complete with project and management history, costs, operation context, failure mechanisms detected and anticipated, analysis and remedies applied. This would allow others to learn from mistakes and would likely enable a powerful new form of collective A.I. improvement over time. And if such a catalogue were publically available, continuously updated and broadcast to public displays in Times Square New York, Shibuya Center Gai Tokyo, and Maracanã Stadium Rio de Janeiro, the global public might even gain some faith in the prospect of a better future for all through A.I.

References

1. Harwell, D.: Amazon facial-identification software used by police falls short on tests for accuracy and bias, new research finds. The Washington Post (2019). https://www.washingtonpost.com/technology/2019/01/25/amazon-facial-identification-software-used-by-police-falls-short-tests-accuracy-bias-new-research-finds/?utm_term=.bae07d3d0eaf. Accessed 31 Jan 2019
2. Townsend, A.: Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia. WW Norton & Company, New York (2013)
3. Dowd, M.: Elon Musk's Billion-Dollar Crusade to stop the A.I. Apocalypse. Vanity Fair. (2017). https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x. Accessed 3 Mar 2019
4. Morozov, E., Bria, F.: Rethinking the Smart City. Rosa Luxemburg Stiftung (2017). http://www.rosalux-nyc.org/wp-content/files_mf/morozovandbria_eng_final55.pdf
5. Hogarth, I.: A.I. Nationalism (2018). https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism. Accessed 5 Mar 2019

6. State Council.: Notice of the State Council Issuing the New Generation of Artificial Intelligence Development Plan. State Council Document No 35. (2017). https://flia.org/wp-content/uploads/2017/07/A-New-Generation-of-Artificial-Intelligence-Development-Plan-1.pdf. English version, Accessed 7 Mar 2019
7. Nourbakhsh, I.: Lecture: Ethics in Robotics. (2009). https://www.youtube.com/watch?v=giKT8PkCCv4. Accessed 6 Mar 2019
8. Santos, G., Behrendt, H., Maconi, L., Shirvani, T., Teytelboym, A.: Part I: externalities and economic policies in road transport. Res. Transp. Econ. **28**(1), 2–45 (2010)
9. Menon, G., Guttikunda, S.: Electronic Road Pricing: Experience & Lessons from Singapore. SIM-air Working Paper Series: 33 (2010). http://www.environmentportal.in/files/ERP-Singapore-Lessons.pdf
10. Tristate Transportation Campaign.: Road pricing in London, Stockholm and Singapore. A way forward for New York City (2017). http://nyc.streetsblog.org/wp-content/uploads/2018/01/TSTC_A_Way_Forward_CPreport_1.4.18_medium.pdf. Accessed 2 Feb 2019
11. Lehe, L.: A history of downtown road pricing. Medium (2017). https://medium.com/@lewislehe/a-history-of-downtown-road-pricing-c7fca0ce0c03. Accessed 5 Mar 2019
12. Eliasson, J.: The Stockholm congestion pricing syndrome: how congestion charges went from unthinkable to uncontroversial. Centre for Transport Studies, KTH Royal Institute of Technology (2014). http://www.transportportal.se/swopec/CTS2014-1.pdf
13. Hau, T.: Electronic road pricing: developments in Hong Kong 1983–89. J. Transp. Econ. Policy **24**(2), 203–214 (1990)
14. Schedler, A.: Electoral Authoritarianism: The Dynamics of Unfree Competition. Lynne Rienner Publishers, Boulder (2006)
15. Sidewalk Labs Project Vision (2017). http://www.passivehousecanada.com/wp-content/uploads/2017/12/TO-Sidewalk-Labs-Vision-Sections-of-RFP-Submission-sm.pdf. Accessed 1 Feb 2019
16. Barth, B.: The fight against Google's smart city. The Washington Post (2018). https://www.washingtonpost.com/news/theworldpost/wp/2018/08/08/sidewalk-labs/?noredirect=on&utm_term=.29807e353bb9. Accessed 1 Feb 2019
17. Wylie, B.: Deputation to Toronto's Executive Committee on Sidewalk Toronto. Medium (2018). https://medium.com/@biancawylie/my-deputation-to-torontos-executive-committee-on-sidewalk-toronto-jan-24-2018-ee25785bc44e. Accessed 1 Feb 2019
18. Zuboff, S.: Big other: surveillance capitalism and the prospects of an information civilization. J. Inf. Technol. **30**, 75–89 (2015). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2594754
19. Goldsmith, S., Crawford, S.: The Responsive City: Engaging Communities Through Data-Smart Governance. Wiley, New York (2014)
20. Barcelona Digital City. https://ajuntament.barcelona.cat/digital/en. Accessed 4 Feb 2019
21. Singh, S., Okun, A., Jackson, A.: Learning to play Go from scratch. Nature **550**, 336–337 (2017)
22. Domingos, P.: The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, New York (2015)
23. Whitehead, A.: Mathematics in the History of Thought (1957). http://www-history.mcs.st-andrews.ac.uk/Extras/Whitehead_maths_thought.html. Accessed 7 Feb 2019
24. Irving, G., Christiano, P., Amodei, D.: AI safety via debate (2018). arXiv:1805.00899v2
25. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete Problems in AI Safety (2016). CoRR arXiv:1606.06565.
26. Singh, A., Patil, D., Omkar, S.: Eye in the Sky: Real-Time Drone Surveillance System for Violent Individuals Identification Using ScatterNet Hybrid Deep Learning Network (2018). arXiv:1806.00746v1

27. Agarwal, K.: For the Third Time in Three Months, Farmer to Protest in Delhi. The Wire (2018). https://thewire.in/agriculture/third-time-three-months-farmers-protest-delhi. Accessed 8 Mar 2019
28. Scharfenberg, D.: Computers can solve your problem. You may not like the answer. What happened when Boston Public Schools tried for equity with an algorithm. The Boston Globe (2018). https://apps.bostonglobe.com/ideas/graphics/2018/09/equity-machine/. Accessed 20 Jan 2019
29. Calthorpe, P., Walters, J.: Autonomous Vehicles: Hype and Potential. UrbanLand (2017). https://urbanland.uli.org/industry-sectors/infrastructure-transit/autonomous-vehicles-hype-potential/. Accessed 20 Feb 2019
30. Green, B., Caro, A., Conway, M., Manduca, R., Plagge, T., Miller, A.: Mining administrative data to spur urban revitalization. In: KDD '15: The 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2015). https://scholar.harvard.edu/files/bgreen/files/kdd2015.pdf
31. Badger, E.: Why Technology Hasn't Fixed the Housing Crisis. New York Times (2019). https://www.nytimes.com/2019/01/29/upshot/can-technology-help-fix-the-housing-market.html. Accessed 4 Feb 2019
32. Chan Zuckerberg Initiative (2018). https://chanzuckerberg.com/newsroom/inspiring-young-leaders-to-tackle-housing-affordability/. Accessed 6 Feb 2019
33. United Nations World Urbanization Prospects. The 2018 Revision (2018). https://population.un.org/wup/Publications/Files/WUP2018-KeyFacts.pdf. Accessed 7 Mar 2019
34. Marvin, C.: When Old Technologies Were New. Oxford University Press, Oxford (1990)
35. IBM.: Diversity in Faces Project (2019). https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/. Accessed 5 Mar 2019
36. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019). https://blog.openai.com/better-language-models/