COMMENT

ARMANDO L. SANCHEZ/CHICAGO TRIBUNE/TNS/GETTY

Chicago police use algorithmic systems to predict which people are most likely to be involved in a shooting, but they have proved largely ineffective.

There is a blind spot in AI research

Fears about the future impacts of artificial intelligence are distracting researchers from the real risks of deployed systems, argue **Kate Crawford** and **Ryan Calo**.

On 12 October, the White House published its report on the future of artificial intelligence (AI) — a product of four workshops held between May and July 2016 in Seattle, Pittsburgh, Washington DC and New York City (see go.nature.com/2dx8rv6).

During these events (which we helped to organize), many of the world's leading thinkers from diverse fields discussed how AI will change the way we live. Dozens of presentations revealed the promise of using progress in machine learning and other AI techniques to perform a range of complex tasks in everyday life. These ranged from the identification of skin alterations that are indicative of early-stage cancer to the reduction of energy costs for data centres.

The workshops also highlighted a major blind spot in thinking about AI. Autonomous systems are already deployed in our most crucial social institutions, from hospitals to courtrooms. Yet there are no agreed methods to assess the sustained effects of such applications on human populations.

Recent years have brought extraordinary advances in the technical domains of AI. Alongside such efforts, designers and researchers from a range of disciplines need to conduct what we call social-systems analyses of AI. They need to assess the impact of technologies on their social, cultural and political settings.

A social-systems approach could investigate, for instance, how the app AiCure — which tracks patients' adherence to taking prescribed medication and transmits records to physicians — is changing the doctor–patient relationship. Such an approach ▶

could also explore whether the use of historical data to predict where crimes will happen is driving overpolicing of marginalized communities. Or it could investigate why high-rolling investors are given the right to understand the financial decisions made on their behalf by humans and algorithms, whereas low-income loan seekers are often left to wonder why their requests have been rejected.

A SINGULAR PROBLEM

"People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world." This is how computer scientist Pedro Domingos sums up the issue in his 2015 book *The Master Algorithm*[1]. Even the many researchers who reject the prospect of a 'technological singularity' — saying the field is too young — support the introduction of relatively untested AI systems into social institutions.

In part thanks to the enthusiasm of AI researchers, such systems are already being used by physicians to guide diagnoses. They are also used by law firms to advise clients on the likelihood of their winning a case, by financial institutions to help decide who should receive loans, and by employers to guide whom to hire.

Analysts are expecting the uses of AI systems in these and other contexts to soar. Current market analyses put the economic value of AI applications in the billion-dollar range (see 'On the rise'), and IBM's chief executive Ginni Rometty has said that she sees a US$2-trillion opportunity in AI systems over the coming decade. Admittedly, estimates are difficult to make, in part because there is no consensus on what counts as AI.

AI will not necessarily be worse than human-operated systems at making predictions and guiding decisions. On the contrary, engineers are optimistic that AI can help to detect and reduce human bias and prejudice. But studies indicate that in some current contexts, the downsides of AI systems disproportionately affect groups that are already disadvantaged by factors such as race, gender and socio-economic background[2].

In a 2013 study, for example, Google searches of first names commonly used by black people were 25% more likely to flag up advertisements for a criminal-records search than those of 'white-identifying' names[3]. In another race-related finding, a ProPublica investigation in May 2016 found that the proprietary algorithms widely used by judges to help determine the risk of reoffending are almost twice as likely to mistakenly flag black defendants than white defendants (see go.nature.com/29aznyw).

THREE TOOLS

How can such effects be avoided? So far, there have been three dominant modes of responding to concerns about the social and ethical impacts of AI systems: compliance, 'values in design' and thought experiments. All three are valuable. None is individually or collectively sufficient.

Deploy and comply. Most commonly, companies and others take basic steps to adhere to a set of industry best practices or legal obligations, so as to avoid government, press or other scrutiny. This approach can produce short-term benefits. Google, for example, tweaked its image-recognition algorithm in 2015 after the system mislabelled an African American couple as gorillas. The company has also proposed introducing a 'red button' into its AI systems that researchers could press should the system seem to be getting out of control[4].

Similarly, Facebook made an exception to its rule of removing images of nude children from its site after the public backlash about its censorship of the Pulitzer-prizewinning photograph of a naked girl, Kim Phúc, fleeing a napalm attack in Vietnam. And just last month, several leading AI companies, including Microsoft, Amazon and IBM, formed the Partnership on AI to try to advance public understanding and develop some shared standards.

Yet the 'deploy and comply' approach can be ad hoc and reactive, and industry efforts can prove inadequate if they lack sufficient critical voices and independent contributors. The new AI partnership is inviting ethicists and civil-society organizations to participate. But the concern remains that corporations are relatively free to field test their AI systems on the public without sustained research on medium- or even near-term effects.

Values in design. Thanks to pioneers in the ethical design of technology, including the influential scholars Batya Friedman and Helen Nissenbaum, researchers and firms now deploy frameworks such as value sensitive design or 'responsible innovation' to help them to identify likely stakeholders and their values. Focus groups or other techniques are used to establish people's views about personal privacy, the environment and so on. The values of prospective users are then incorporated into the design of the technology, whether it is a phone app or a driverless car[5]. Developers of AI systems should draw on these important methods more.

Nevertheless, such tools often work on the assumption that the system will be built. They are less able to help designers, policymakers or society to decide whether a system should be built at all, or when a prototype is too preliminary or unreliable to be unleashed on infrastructure such as hospitals or courtrooms.

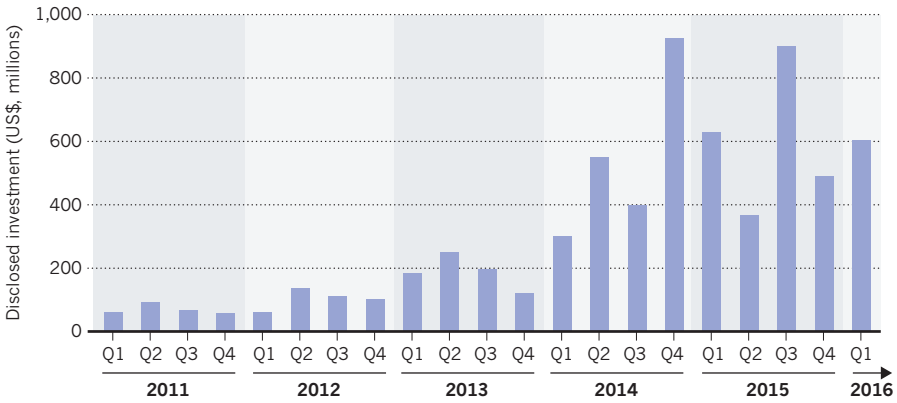
Thought experiments. In the past few years, hypothetical situations have dominated the public debate around the social impacts of AI.

The possibility that humans will create a highly intelligent system that will ultimately rule over us or even destroy us has been most discussed (see, for example, ref. 6). Also, one relevant thought experiment from 1967 — the trolley problem — has taken on new life. This scenario raises questions about responsibility and culpability. In it, a person can either let a runaway trolley car run along a track where five men are working, or pull a lever to redirect the trolley on to another track where only one person is at risk. Various commentators have applied this hypothetical scenario to self-driving cars, which they argue will have to make automated decisions that constitute ethical choices[7].

Yet as with the robot apocalypse, the possibility of a driverless car weighing up 'kill decisions' presents a narrow frame for moral reasoning. The trolley problem offers little guidance on the wider social issues at hand: the value of a massive

ON THE RISE

Investment in technologies that use artificial intelligence has climbed in recent years.



SOURCE: CB INSIGHTS

People with asthma were wrongly graded as low risk by an AI system designed to predict pneumonia.

CRAIG F. WALKER/*DENVER POST*/GETTY

investment in autonomous cars rather than in public transport; how safe a driverless car should be before it is allowed to navigate the world (and what tools should be used to determine this); and the potential effects of autonomous vehicles on congestion, the environment or employment.

SOCIAL-SYSTEMS ANALYSIS

We believe that a fourth approach is needed. A practical and broadly applicable social-systems analysis thinks through all the possible effects of AI systems on all parties. It also engages with social impacts at every stage — conception, design, deployment and regulation.

As a first step, researchers — across a range of disciplines, government departments and industry — need to start investigating how differences in communities' access to information, wealth and basic services shape the data that AI systems train on.

Take, for example, the algorithm-generated 'heat maps' used in Chicago, Illinois, to identify people who are most likely to be involved in a shooting. A study[8] published last month indicates that such maps are ineffective: they increase the likelihood that certain people will be targeted by the police, but do not reduce crime.

A social-systems approach would consider the social and political history of the data on which the heat maps are based. This might require consulting members of the community and weighing police data against this feedback, both positive and negative, about the neighbourhood policing. It could also mean factoring in findings by oversight committees and legal institutions. A social-systems analysis would also ask whether the risks and rewards of the system are being applied evenly — so in this case, whether the police are using similar techniques to identify which officers are likely to engage in misconduct, say, or violence.

As another example, a 2015 study[9] showed that a machine-learning technique used to predict which hospital patients would develop pneumonia complications worked well in most situations. But it made one serious error: it instructed doctors to send patients with asthma home even though such people are in a high-risk category. Because the hospital automatically sent patients with asthma to intensive care, these people were rarely on the 'required further care' records on which the system was trained. A social-systems analysis would look at the underlying hospital guidelines, and other factors such as insurance policies, that shape patient records[9].

> *"Artificial intelligence presents a cultural shift as much as a technical one."*

A social-systems analysis could similarly ask whether and when people affected by AI systems get to ask questions about how such systems work. Financial advisers have been historically limited in the ways they can deploy machine learning because clients expect them to unpack and explain all decisions. Yet so far, individuals who are already subjected to determinations resulting from AI have no analogous power[10].

A social-systems analysis needs to draw on philosophy, law, sociology, anthropology and science-and-technology studies, among other disciplines. It must also turn to studies of how social, political and cultural values affect and are affected by technological change and scientific research. Only by asking broader questions about the impacts of AI can we generate a more holistic and integrated understanding than that obtained by analysing aspects of AI in silos such as computer science or criminology.

There are promising signs. Workshops such as the Fairness, Accountability, and Transparency in Machine Learning meeting being held in New York City next month is a good example. But funders — governments, foundations and corporations — should be investing much more in efforts that approach AI in the way we describe.

Artificial intelligence presents a cultural shift as much as a technical one. This is similar to technological inflection points of the past, such as the introduction of the printing press or the railways. Autonomous systems are changing workplaces, streets and schools. We need to ensure that those changes are beneficial, before they are built further into the infrastructure of everyday life. ■ SEE WORLD VIEW P.291

Kate Crawford *is a principal researcher at Microsoft Research in New York City, a visiting professor at the Massachusetts Institute of Technology in Cambridge, Massachusetts, and a senior research fellow at New York University, New York, USA.*
Ryan Calo *is an assistant professor of law and of information science (by courtesy), and faculty co-director of the Tech Policy Lab at the University of Washington, Seattle, Washington, USA.*
e-mails: kate@katecrawford.net; rcalo@uw.edu

1. Domingos, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (Allen Lane, 2015).
2. Barocas, S. & Selbst, A. D. *Calif. Law Rev.* **104,** 671–732 (2016).
3. Sweeney, L. *Discrimination in Online Ad Delivery* (2013); available at http://dx.doi.org/10.2139/ssrn.2208240
4. Armstrong, S. & Orseau, L. in *Uncertainty in Artificial Intelligence: Proceedings of the Thirty-Second Conference* (eds Ihler, A. & Janzing, D.) 557–566 (AUAI Press, 2016); available at http://go.nature.com/2drokil
5. Friedman, B., Kahn, P. H. & Borning, A. in *Human–Computer Interaction in Management Information Systems: Foundation* (eds Zhang, P. & Galletta, D.) 348–372 (M. E. Sharpe, 2006); available at http://go.nature.com/2dee8om
6. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford Univ. Press, 2016).
7. Lin, P. in *Autonomes Fahren: Technische, Rechtliche und Gesellschaftliche Aspekte* (eds Maurer, M., Gerdes, J. C., Lenz, B. & Winner, H.) 69–85 (Springer, 2015); available at http://doi.org/brdw
8. Saunders, J., Hunt, P. & Hollywood, J. S. *J. Exp. Criminol.* **12,** 347–371 (2016).
9. Caruana, R. *et al.* 'Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission' *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 1721–1730 (ACM, 2015).
10. Crawford, K. *et al. The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* (2016); available at https://artificialintelligencenow.com