CA State Legislature
Senate Select Committee on the Growing Impact of Artificial Intelligence in California
Hearing on 25 June, 2019: Overview of Policy Issues

Testimony of Joshua A. Kroll, PhD.
Postdoctoral Research Scholar, University of California, Berkeley School of Information

TRUST AND ACCOUNTABILITY IN AI SYSTEMS

Chairman Umberg, members of the Select Committee, thank you for the opportunity to speak to you today. My name is Joshua Kroll, and I am a computer scientist studying the relationship between technology and governance – with a focus on AI and automated decision-making – at the UC Berkeley School of Information, as a postdoctoral research scholar. My work focuses on the need to design computer systems to support human values such as fairness, privacy, security and legal compliance when computers are used to make life-altering decisions such as who wins an election, who gets access to financial products, whether someone should be incarcerated, or whether a diagnostic image shows a malignancy. As we move into a world where *artificial intelligence* increases the number, stakes, and speed of these decisions, it is essential that we design systems to support human values and to be subject to robust governance so that we can ensure people and organizations remain accountable. Here, I use the term “artificial intelligence” to mean any behavior by a machine that a person would consider to be intelligent, regardless of how that behavior is implemented.

I want to make three points today: **first**, while mechanical, computers are not objective; **second**, that there exist technical and nontechnical methods for controlling software-based systems, holding them to high standards of fitness for use (including appropriately accounting for values such as fairness and security), and verifying to regulators, oversight entities, and the public at large that these standards are being met; and **third**, while there is no universal way to understand the impacts and consequences of a particular computer system, governance regimes based on robust accountability hold the greatest promise of enabling the benefits of technology while minimizing the harms.

Artifacts Have Politics

To my first point, although computers are often described as mechanical and therefore objective, we must understand that all technical systems are, in truth, sociotechnical systems, meaning they include people, organizations, institutions, as well as legal and social norms.

The imagined objectivity of computers derives from their mechanical nature and their reliance on data about the world. But choices about which data to gather, which gathered data to use, and the nature of that use work together to capture existing societal inequities. Data result from a process of *measurement* and can systematically deviate from the underlying phenomenon of interest, known technically as the problem of *bias*. Measurement choices are just one example of the design choices made when AI systems are constructed or deployed into the world.

No decision is purely technical or entirely free of political considerations, and no computer system should be considered to reveal objective truth, even systems that are “based on data” about the world, as many AI systems are. Indeed, AI systems can and do reflect and reinforce existing social structures and biases, discriminating against vulnerable people and – often inadvertently – treating historically disadvantaged groups less well than historically advantaged groups.

Machines play forward the political choices made during their design and construction. When systems are based on historical information that reflects discrimination, the choice to use this information reflects a particular perspective on how such systems should work. For example, research has shown that face detection systems, which identify whether a photograph has a face in it, work less well for dark-skinned people than for white people as well as less well for women than for men. The performance for dark-skinned women is even worse, transforming a mostly functional (but occasionally erroring) system that works for nine out of ten people into a broken one that works only two times out of three. Other research (to which I contributed) demonstrated the potential for false matches in face recognition systems, which associate an identity with a detected face, including nearly twice as many false matches for dark-skinned faces as for light ones. Both kinds of unequal failure can be attributed to a combination of design and evaluation workflows that did not consider sufficiently diverse data and decisions in the early history of photography about how to build cameras to capture white faces accurately. Even if we stipulate that a face processing system worked perfectly, its use would tend to enable surveillance and policing, both practices deployed heavily against the minority groups for which the technology is already less accurate. Indeed, a bill limiting the use of face surveillance in police body cameras is currently pending before this body as AB1215.

In a different context, research by journalists at ProPublica demonstrated that a tool used to establish the risk of recidivism in Broward County, FL falsely rated black arrestees as high risk for future crime at nearly twice the rate of white arrestees, even while correctly establishing the risk of future arrest for both groups. This counterintuitive result can be attributed to the fact that black people in Broward County are arrested at a substantially higher rate than white people, mathematically guaranteeing that calibrating ratings to risk would lead to unequal error distributions. ProPublica's study has spawned much research on bias and discrimination in AI systems. In California, the recently passed and now-recalled SB10 mandates the use of similar risk assessments to determine pretrial sanctions as part of the laudible elimination of money bail. But with few controls on whether these tools promote justice, due process, or the equitable administration of criminal procedure, risk assessment tools lead to important new governance questions. I contributed to a report by the AI research community on the appropriate governance of such risk assessment systems, coordinated and released by the multi-stakeholder Partnership on AI for the Benefit of People and Society, which lays out a baseline of ten issues with which responsibly governed criminal risk assessments must grapple.

Even decisions about what computer systems to build are political: studies have shown that simply reminding people of their court dates, for example by text message, significantly improves appearance rates. Without comparing such alternative approaches and starting from the perspective of what a system's goals are, we risk ceding control over important public functions to places beyond the reach of governance. Decisions about who will be released during the pretrial period and who will be detained stand to be influenced by data scientists with no connection to the public policymaking process more than by judges.

The fact that judges have the final say on these computer generated "recommendations" is cold comfort in light of research into *automation bias*, the deference of humans towards machine-generated decisions, and research revealing the manipulation of scores by court clerks or other professionals in the criminal justice system who gather and enter the data from which scores are calculated.

Trust, but Verify

My second point is that technical tools and nontechnical interventions can help us construct software that is reliable – both fit for purpose, and supportive of human values and governance goals. This requires a paradigm shift away from a mindset that we build computer systems and only then assess their impact on social and political values and consider modifying them. Instead, systems must be designed from the start to reflect their *requirements*, criteria established at the beginning of a system's lifecycle describing what it should and should not do, and explicitly including goals or constraints such as protecting privacy or advancing fairness. Common industry practices such as software testing, code review, and explicit tracking of bugs all serve to improve security and reliability and bring us closer to this goal. But even with such methods, bugs and vulnerabilities are perceived as inevitable. While no system can be perfect, computer science offers a suite of richer and more powerful tools, including cryptography, software verification, and a mathematical understanding of the capabilities and limitations of machine learning techniques, all of which support stronger claims about reliability, fitness for purpose, and the preservation of values such as privacy, security, or fairness. My own work has demonstrated how to build enhanced audit logs for automated decision-making systems, so that consumers who receive, say, a credit score can be certain that a complete justification for their score exists (that is, the formula for generating the score is recorded along with their credit history to which the formula was applied, and the formula applied to their personal credit history generates their particular credit score).

These tools can demonstrate that the same decision policy (for example, the same credit scoring formula) was applied to everyone. All this is possible while protecting the proprietary nature of the decision policy (for example, the credit scoring formula).

AI systems need not be faulty, unreliable black boxes which lie beyond human understanding. This is a political choice: Software systems can instead be designed to function correctly in all situations, or at least to contain failures to particular subsystems so that they do not impair overall reliability or fitness for use. The imagined inscrutability of computer systems stems from differences in knowledge and power dynamics between system controllers, system consumers, and system overseers. Although large software platforms often claim they cannot explain why a particular advertisement was shown, why a particular search result was ranked as it was, or when and how often a particular social network post will be viewed, the same companies are often highly certain of whether changes to their algorithms will increase or decrease revenues or change the behaviors of their users. Certain aspects of the computing ecosystem – for example, the Google home page – have become so consistently available that outages are now cause for news stories.

Just as airplanes, elevators, and bridges have been improved so that you and I can rely on them without the need for a deep or detailed understanding of how or why they work, so too can software be designed and built to meet standards of proper function in a convincing way. And just as the FAA, the Elevator Unit of the Occupational Safety and Health, or the Department of Transportation parlay those technical assurances into the trust of average citizens, so could an appropriate regulatory regime for software systems parlay the certainties of software-based systems into broad social trust.

Accountable Algorithms

This brings me to my third point, which is that humans and organizations must be held responsible for the behavior of computer systems, especially AI systems. Unfortunately, computer systems tend to undermine existing governance structures, challenging assumptions about the nature of failures, the ability to examine and analyze decision mechanisms, and the speed and scale of decision-making. Merely explaining software outputs to humans is unlikely to help, and placing a human into the workflow artificially limits the promise of autonomy these systems offer while simultaneously preventing the application of software to certain domains where it can be extraordinarily valuable. Further still, we know human decision-making can be flawed yet we treat it as a gold standard. Computers are machines. We can design them so that we know how they will behave in all cases. Thus, we can hold them to higher standards of performance and transparency than even human bureaucracies.

All too often, computer systems fail to reflect the values we desire them to uphold. Software is excellent at implementing clear-cut rules at high speed across large populations. However, computers – even with the advent of advanced AI – are less apt at dealing with situations that require reasoning beyond stated rules or interpretation in the application of those rules. Flexible standards, abstract principles, and the normal intentional ambiguity of contracts and legislation all lie beyond the mechanistic operation of software. While computer scientists can guarantee the fidelity of a software system to articulated requirements, the real world requires more flexibility. Instead, to support these high-minded goals, we must tie human institutions (such as courts, judges, and other oversight entities) into computer systems, designing *sociotechnical* systems which improve upon the governance structures we have today.

Yet policy must not mandate the design of technology. Instead, the humans and human organizations who create AI systems must be made responsible for the safe and appropriate operation of these systems. That is, policy interventions should specify what the controllers of computer systems are responsible for rather than the manner in which systems achieve those assurances, perhaps setting minimum standards or guidelines for procurement. Holding the creators and controllers of AI systems responsible for their behaviors preserves the flexibility for innovation while providing avenues for governance and oversight. Such responsibility will require the development of new best practices in AI development to tie technical and organizational controls into support of the operative values.

Software system developers are currently unanswerable for their decisions, even when these decisions cause harm. Stakeholders rarely have any public forum in which to express their needs and concerns. Protections for proprietary technology and intellectual property interests have tended to shield system designers from the need to respond to requests for information. One study applied open records laws to “smart city” automation projects in 42 jurisdictions in 23 states, representing deployments of six types of predictive AI systems. Only one county was able to provide a level of transparency consistent with decision-making by government rather than the private sector. Well governed AI requires a more robust public sphere around system design and deployment considerations.

Although it is often suggested as a solution, full transparency of software systems is generally neither necessary (because of the assurance techniques mentioned previously) nor sufficient (because software not designed to facilitate assurance may in fact be impossible to analyze). Further, transparency is often undesirable as it presents risks to individual privacy and trade secrecy, while facilitating the strategic “gaming” of automated systems. Technology is creating new opportunities for governance, subtler and more flexible than total transparency.

Conclusion

Of course, this discussion elides a number of important questions in the policy of artificial intelligence. The AI industry’s hunger for data about individual consumers and citizens leads to new threats to privacy. The use of statistical techniques to infer sensitive information such as health status without its being explicitly disclosed challenges the assumptions of existing data protection laws, such as the newly enacted California Consumer Privacy Act. For example, one major retailer reportedly developed a system for predicting which customers are pregnant based on their purchases, raising the specter of new forms of health-based discrimination from actors with no overt access to healthcare information.

There is also wide agreement that, as AI automates an increasing number of tasks, the number and nature of jobs remaining for humans will change dramatically. As new and more automated systems of work displace existing human-mediated systems of work, the effects on workers and the policy responses necessary to mitigate them remain important open questions.

In conclusion, as policy and engineering practice evolve to deal with the newfound importance of software systems, it is critical that we do not allow software systems to operate unchecked. By designing AI systems so that they better align with legal and policy objectives, we can improve the governance of software systems and, sometimes, of existing human-mediated systems by extension. Thank you, and I look forward to your questions.