KA Draft Written Submission
May 22, 2018
US Congress Tom Lantos Human Rights Commission

Likely Impacts of Emerging Artificial Intelligence Software Agents On Internal Human Rights Conditions in Authoritarian States

*

Artificial Intelligence: The Consequences for Human Rights
Hearing Before the Tom Lantos Human Rights Commission, United States Congress
Randy Hultgren, M.C. and James P. McGovern, M.C., Co-Chairs
2255 Rayburn House Office Building, Washington DC
May 22, 2018

Written Submission By
Kenneth Anderson, Professor of Law
Washington College of Law, American University
Washington DC

*

I. Summary

My thanks to the Tom Lantos Human Rights Commission for inviting me to make this submission on a question that is likely to take on increased importance over time: how emerging technologies in artificial intelligence software agents are likely to impact the internal human rights conditions of authoritarian regimes. I would like to preface my remarks below by saying that I had the privilege of meeting occasionally with Rep. Lantos in the 1980s when I worked as an NGO human rights lawyer. He was a person of great personal integrity and a deep, principled commitment to issues of human rights. It is an honor to be invited to make this submission to the Tom Lantos Human Rights Commission.

The key conclusions of my submission are that emerging applications of AI technologies will have important implications for the internal conditions of human rights in some, perhaps many, authoritarian countries – but that these applications are, today, still largely "emerging" rather than "emerged." The uses and misuses of these applications of AI by authoritarian states beyond a handful of technologically sophisticated pioneer states (China, most importantly) are largely still to come, and the contours of how they might impact particular societies much dependent on

the specific characteristics of the applications, as well as the characteristics of the regime and society into which they are deployed, including the extent and sophistication of that society's digital infrastructure.

The policy implications for the United States government today are that it should be keenly observant of how such applications emerge, including their technological specifications, capabilities and limitations. In particular, it should absorb and take account of how these emerging AI applications are used and how they behave in both democratic and authoritarian societies, in order to understand ways in which these applications can be used and abused. This points toward taking account of what today is known as the field of "AI ethics" – interdisciplinary examination of the ways in which AI applications can be engineered and used in ethical ways, as well as ways in which, whether intentionally or unintentionally, these technologies wind up being used in unethical ways.

It seems unlikely to me that the US government will be able to prevent the spread of such AI software applications through long-standing methods of export controls, licenses, etc. There are too many potential producers of such applications; the US does not have a special lock on these technologies, at least in their generic (and customizable) forms. It can assist US-based global technology companies in establishing industry standards in design, deployment, and use that might provide important normative markers, whether formal or informal, for acceptable uses of such technologies. It might be able to assist or encourage the development of AI applications – or applications drawing on other emerging technologies, such as distributed ledger or blockchain, cyber, or combinations of these – that might be of assistance to beleaguered human rights defenders at risk in authoritarian regimes, either in protecting themselves or in the work of gathering information on human rights abuses. But there are limits on how much the US government (or any government) is likely to be able to do to constrain the spread of these technologies or their illegitimate uses by authoritarian regimes.

II. AI Software Applications in Machine Learning

AI technologies and applications covers a vast range of possibilities, and it is important to understand certain key differences, including what technologies are clearly "emerging" and the nature of their likely capabilities and limitations. It is also essential to focus on "real," even if

"emerging," technologies and applications of AI, rather than jumping to purely speculative possibilities for imaginary "AI."

The AI technologies and applications most relevant to the internal human rights conditions of authoritarian regimes are AI software agents – not physical, robotic machines. For that reason, as well as to avoid a range of very different normative and practical considerations, everything in this submission refers to pure software agents that run on computers, perhaps (and perhaps very likely) combined with cyber technologies – but not physical robots, such as autonomous weapon systems. As a general rule of thumb (and despite both the genuine successes but also hype surrounding self-driving vehicles), AI-enabled robotics is harder to do than pure AI software agents consisting purely of code; robotics involves sensors and motion/mobility in the physical world, and thus robotics requires whole fields of engineering not required by software programs alone.

A further narrowing of the field of AI to focus on the part most relevant to repression in authoritarian regimes today means drawing differences between "rules-based" AI and "machine learning" AI (in its several forms). What is normally understood as "computer programs" of the last few decades is computer software based around the execution of rules-based algorithms – the rules of arithmetic, for example, in a calculator; we tend to forget that this is what the vast array of computerized functions in technologically advanced societies consists of, rather than the "emerging" AI techniques of machine learning (ML). For exactly the same reasons, however, that computer programs to automate such tasks as making social security payments to millions of individuals, or enabling telecommunications networks, or so many other things, allow society to work better, computer programs also exist that can automate such things as screening calls across a phone network for specific phone numbers believed to be used, for example, by dissidents or human rights defenders. These applications of rules-based AI computer technologies are so normal that we hardly think about them, but in fact form the large bulk of ways in which software can be used to repress in an authoritarian state.

The newer AI software applications comprising ML and its subcategories are today receiving most of the attention, but they are largely still "emerging"; have special social and technological requirements to be used effectively; and have uses (whether for good or bad) that are narrower than the existing range of applications of ordinary computerization. ML technologies are all about pattern recognition – various techniques for extracting patterns out of large quantities of

data. The most important and most-discussed form of ML today is a type of so-called "artificial neural networks" (ANN) widely known as "Deep Learning" (DL). DL algorithms are largely at the heart of the current enthusiasm for AI technologies, and they are also at the heart of current controversies over AI applications and AI ethics. From the standpoint of both national security and human rights, DL has important implications because of the successes it has had in areas ranging from recent victories playing a strategy game such as Go to facial recognition software and related mass surveillance technologies.

DL successes have led to high hopes for the emergence of "predictive analytics" using "Big Data," among other things. In addition to applications such as AlphaGo or facial recognition software, DL has been used by private companies to create algorithms for, among other uses, purport to predict recidivism in the US criminal justice system (and already used in sentencing in some cases); individuals likely to be at risk from gang violence (used by some American police departments); buildings in a city likely to have a fire occur; and many more. Some of these algorithms work better than other prediction tools (including human experience and intuition); some of them don't; and with others, the lack of counterfactuals makes it difficult or impossible to know.

Indeed, a key and controversial aspect of DL algorithms is not just that they are hugely complex and opaque (true of code generally), but that they necessarily use probabilistic techniques that make it difficult to impossible to fully predict how the algorithm will behave ex ante or fully reconstruct how it did behave ex post. For this very reason, however, the "ethical AI" movement within the technology communities, at least in the open societies, has been pressing for new techniques and technological tools by which to evaluate how an algorithm acts, and to be able to assess whether a DL software program does what it is supposed to do and doesn't do what it's not supposed to do.

Important steps have been taken toward "Explainable AI," but there is still a distance to go in creating widely usable tools for "verification and validation, testing and evaluation" within the field of reliability engineering for these new forms of AI software. Moreover, one apparent finding in this field today is that, perhaps unsurprisingly, software can be made much more "explainable" – predictable up front or reconstructable afterwards – if it is *designed* to be explainable. This possibility of establishing norms for designing "Explainable AI" has implications for ways in which the US government, together with technology companies and

governments of open societies, might be able to influence how DL algorithms with applications to surveillance, in legitimate national security ways or as tools of internal repression, can be generally engineered in accordance with industry common standards for transparency and explanation. It is by no means a "fix" to the human rights risks of DL algorithms, but it would matter if the routine, commercial or standard government, AI applications were built using widely accepted, verified and validated, "explainable" techniques – states could build their own without such features, or China or Russia or their companies might sell them, but it would help if there was a common commercial design norm favoring transparency.

The last important feature of ML and DL systems that matters to their use legitimately or illegitimately is that they are only as good as the datasets on which they "train." ML is "learning" because the algorithm is able to process a large number of examples relevant to the intended task – facial recognition, for example – from which it can learn correct and incorrect, within a probability range. In general, the datasets need to be very large in order to generate "accurate" learning, and smaller datasets can easily "teach" the machine algorithm systemically bad patterns – or simply produce results with many false positives or false negatives. Moreover, "datasets" actually means data that is digitized (while it's true that technological societies have large digitized data sets for some things, other societies do not, and much key information is not captured digitally at all); accurate; and structured in such a way as to capture the intended features for analysis, and not inadvertently pushing the algorithm to learn unintended lessons.

The rash of straight-up racist ML outputs from otherwise non-racist datasets has alerted the technology community – less the application user community, so far – to ways in which ML programs can produce not merely incorrect, but socially abhorrent or illegal, results from datasets in which such outputs would not be obvious. There is almost certainly going to be a backlash, and perhaps regulation, in the US against untested and unverified algorithms that have negative impacts on individuals – lending decisions, for example, or criminal sentencing – in the US; Europe is already leading the way in terms of the regulation of the use of personal data and gradually emerging requirements of "Explainable AI."

Finally, as psychology professor and AI expert Gary Marcus has noted, DL algorithms perform far better at pattern recognition involving vast quantities of "primitive" data – pixels, for example, in facial recognition – than they do in higher level cognitive tasks. ML algorithms are about pattern extraction – correlations across large datasets – and not identifying causation or

causes, or even the direction of causality in a correlation extracted from data. A ML learning algorithm developed in order to help predict who in an ICU was likely to die, using vast amounts of medical records, for example, achieved a remarkably good success rate in its predictions – so successful that its designers took a look inside the algorithm's black box. They discovered that the algorithm had focused with relentless literalness and no human common sense on a feature of the ICU medical records that had a box for the ICU physician to check, "Call hospital chaplain."

III. Likely Uses of AI Software Agents by Repressive Regimes

The uses of AI software agents, particularly ML and DL algorithms, by repressive, authoritarian regimes to monitor and control their own populations are likely to track the legitimate policing, intelligence, and national security uses of them. Essentially the same facial recognition software will be used – and available – to security services in open societies engaged in legitimate uses and in authoritarian societies where goal is to prevent dissent by identifying dissenters at an early stage. This means that attempts to restrict the technologies' use to "legitimate" purposes will always be somewhere between difficult to impossible.

Moreover, one of the engineering difficulties of ML algorithms is that testing and evaluating to ensure that, legitimate or not, the software identifies the correct persons and doesn't draw in large numbers of false negatives or false positives is a difficult task. Off the shelf tech, even if sophisticated on its own terms, would certainly require extensive customization for any particular application in any particular society. If, however, you are an authoritarian regime that cares deeply about identifying dissenters, but doesn't much care if you identify too many people as dissenters who aren't – false positives – you might not be worried about off the shelf software that isn't customized with any sophistication.

The biggest limiting factors on the uses of such ML algorithms today in repressive societies outside the most important technological giants – China, e.g. – are likely two. One is simply that a society doesn't have enough digital infrastructure on which people routinely or necessarily depend, such as payment systems or banking, to use such digitally-based software on most people in the society. Additionally, with regards specifically to dataset driven ML algorithms, the digital infrastructure might not generate sufficiently large datasets that would run to the relevant information sought, e.g., facial recognition without widely used digital infrastructure ranging from ubiquitous video monitoring to Facebook users sufficient to make it

likely that the relevant targets will be found. Thus, one reason why the use of ML algorithms specifically by many non-technologically sophisticated countries will not be immediate is that the population broadly doesn't provide the inputs for digitized databases. On the other hand, dissenters are often not the agricultural peasants, for example, but rather elites and those with access to the world through digital means. They do participate, and their tools of dissent are overwhelmingly likely to be digital. AI automation software combined with cyber monitoring tools can be a potent regime weapon to identify and surveil such dissenters – note, however, that these tools are already widely available, because they are not ML or DL algorithms, just ordinary computer programs monitoring digital communications.

[More TK in final version]

IV. Conclusions for US Government Policy

This submission has suggested that ML and DL tools that can be used for human rights violations and suppression of dissent within a particular authoritarian society are not yet widely available – but almost certainly will be. It would be a mistake to generalize from the example of China – gigantic and technologically sophisticated – to the many other authoritarian countries in the world. Those countries might not have the digital infrastructure at this point such that any form of computerized surveillance would be effective – less so in the case of the cutting edge ML algorithms requiring large digital datasets. That said, digital sophistication will often be on the way in many authoritarian states – and the ability to monitor dissent by channeling members of society through digital tools under control or surveillance by an authoritarian government actually creates an incentive to invest in authoritarian-friendly versions of digital and cyber systems.

This submission has also argued that the appropriate role for the US government at this stage is to be sure to inform itself of possible ways in which such technologies could be abused – in part by paying close attention to the issues today of AI ethics of design. They are likely to also be ways in which such technologies are abused in authoritarian societies. The US government and governments of democratic countries might be able to work with their technology companies to come up with ways to limit the use of such technologies for illegitimate ends of human rights abuse and repression. This is likely easier said than done, however, because in many cases, the line between legitimate and illegitimate use of the technology will turn on the intent of the

government in using it, to ends of legitimate policing and national security or illegitimate identification and suppression of dissenters in an authoritarian regime.

[This draft is not finalized; it will be extended, along with references, in its final version.]

END