FUTURE OF HUMANITY INSTITUTE, UNIVERSITY OF OXFORD

AI: The Consequences for Human Rights

Initial findings on the expected future use of AI in authoritarian regimes

Paul de Font-Reaulx

House Foreign Affairs Committee
Tom Lantos Human Rights Commission Hearing:

Artificial Intelligence: The Consequences for Human Rights
Tuesday, May 22, 2018 – 3.00pm
2255 Rayburn House Office Building

This report is based on research conducted during a temporary stay at the Future of Humanity Institute. I thank my peers there for the invaluable input they have provided on this material, and particularly Jeff Ding for his remarkable help in bringing this report into existence. Any views expressed here are my own.

1. INTRODUCTION

This report provides some initial findings on how AI-technologies available within the coming decade are likely to affect the behavior of authoritarian regimes towards their citizens. In extension I hope it can contribute to a discussion of whether such technologies will make violations of human rights more or less likely in the future, and if so in what way.

Discussions of future technological impact are necessarily uncertain. My aim here is not to provide robust predictions, but to guide further discussions towards the most realistic future scenarios and away from those unlikely scenarios which might otherwise receive undue attention.

In this report I conclude primarily the following:

- Impactful uses of AI are unlikely to be effectively implemented in most authoritarian states within the coming decade, due to lacking digital infrastructure and dependence on elites who could undermine its efficacy. The discussion should therefore be limited to cases like China or the Gulf States where these limitations are less severe.
- If the AI technologies considered here were effectively implemented then that would make *political opposition more difficult*, *overt violence perpetrated by the regime less prevalent* and *produce some shifts towards totalitarian traits*. All of these effects are uncertain.
- The United States and the international community should focus on understanding the effects of AI on human rights better. Direct action at this stage is difficult due to the ubiquity of the underlying technology, but one option is to fund research into counter-technology to be used by political opposition in authoritarian regimes.

I elaborate on these points below. I start by outlining some technologies that we can plausibly expect authoritarian regimes to make use of, and then discuss why we might doubt that these would be effectively implemented. I then consider some implications relevant to human rights if they were effectively implemented. I conclude by commenting on the actions that international actors could take.

2. HOW CAN AI BE OF USE TO AUTHORITARIAN REGIMES?

Authoritarian regimes face a threat from a population that might mobilize into mass protests. If new technology would allow a regime to mitigate that threat, then we should expect the regime to make use of it. The recent developments in AI might provide a number of new such methods for authoritarian regimes to stabilize control. Below I consider four applications of AI which will plausibly be available to regimes within 10 years based on current research.

A) HIGH-PRECISION ALGORITHMS FOR IDENTIFYING DISSIDENTS

Advanced pattern-recognition systems developed in the last few years allows an actor to turn large amounts of data into useful information at an unprecedented scale. When individuals make decisions in their daily lives they provide different actors with such data, which includes for example revealing their location to various apps and their purchases to credit card providers. Sophisticated algorithms can be used to make reliable inferences from these data points to e.g. political inclinations.[1]

For instance, let us assume that someone who purchases only vegan food is statistically likelier to support gender equality policies. An actor that has data available on a person's purchases could then make a probabilistic claim about that person's political inclination. If they also have information that the person has been in the geographical vicinity of relevant political events, then they can infer with a high probability that the person has such views.[2]

An authoritarian regime with access to substantial data on its citizens could use such methods to infer highly relevant information about specific individuals. In particular it might allow the regime to gauge the degree to which a person supports the regime, and how much of a threat that specific individual poses. This makes repression less costly for a regime, as it can target a lower number of individuals and respond in an effective manner. It also makes it more attractive for a regime to rely on avoiding the spread of dissent as opposed to handling it once it does arise.

Knowledge that the regime has this capacity should dissuade individuals from engaging in regime-critical behavior. This is clearest in the case of protesters for instance. If you know that a regime can identify you with high reliability if you participate in a protest, and

[1] See e.g. Matz & Netzer (2017) and Sundsoy (2017).

[2] These techniques are already being put to use by US police for instance. The Chicago Police Department make use of algorithms to put together a 'Strategic Suspect List' of individuals who are statistically likely to be perpetrators or victims of gang shootings, and use this to inform them of the risk they are facing. Police in Los Angeles, New Orleans and New York have implemented similar techniques. (Ferguson 2017)

retroactively punish you if the protest is unsuccessful, then this would decrease your incentive to participate.

As will be discussed more below however, the effective implementation of such information-processing systems requires a high level of digital infrastructure and domestic organizational capacity which is unlikely to be found in most authoritarian states in the near future.

B) SOCIAL CREDIT SYSTEMS

Using the information-processing techniques considered above, a regime can systematize its response in the form of a social credit system or equivalent which rewards desirable behavior. This score can in turn be used as a supplement for financial credit for instance, or other benefits. Social credit systems are famously being explored by China today.[3]

Systematizing rewards and punishments in a way which is advantageous to a regime provides a cheap way of solidifying stability. It produces significant and consistent incentives for individuals to behave in a way commensurable with the benefits of the social credit system, which in turn dissuades regime-critical behavior which could constitute a threat. In the longer term there is some reason to believe that social credit systems could shape social norms to the benefit of a regime, which in turn might perpetuate long-term regime stability.

While China's Social Credit System has received much media coverage, there is reason to believe that this hype is somewhat overblown.[4] To the contrary of popular reports, it is unlikely that the Social Credit System will be fully implemented on a national obligatory basis by 2020. That does not mean however that this will not be the case in the future.

There are however other reasons to think that its spread could be limited. If a social credit system was highly centralized, then it would be fragile and constitute a security risk for the state due to its vulnerability to cyber-attacks. If it was made regional however, then the benefit in terms of stability would be contingent on regional elites, who might themselves constitute threats to the regime.

C) DISTORTION OF PUBLIC DISCOURSE

Advanced AI has allowed not just for the extraction of information from data, but also the creation of information made to confuse. In particular we have seen the development of 'bots' able to imitate humans in different contexts which are sometimes difficult to tell apart from

[3] For more on China's use of AI, see Ding's (2018) excellent overview.

[4] https://www.chinalawtranslate.com/seeing-chinese-social-credit-through-a-glass-darkly/?lang=en

real humans. As these technologies develop, we should expect it to become more difficult to distinguish bots from people in online interactions for instance.[5]

These bots can be used to shift the perception of public discourse in a way that is beneficial to the regime. By using bots to infiltrate forums online, e.g. comments on videos or news articles, a ruling coalition may use bots to create the impression that there is more widespread support for the regime than is actually the case. This could be used to dissuade would-be dissidents from taking action, giving the impression that they would be highly unpopular if they did. The other application of bots is to undermine the credibility of political opposition.[6]

The novelty of human-imitating bots is not the ability to influence sources of information for the population. This has often been done historically in the form of propaganda, and it is unclear to what degree it is effective, and to what degree it just causes the population to stop believing what they read. The difference with bots is that it might allow a regime to effectively infiltrate those forums which people might use as reference points to check whether national media is really credible, e.g. informal conversation online.

D) DRONES FOR ASSASSINATIONS

Military-style drones used to eliminate targets in warfare are already prevalent, and are currently being developed by numerous states. In the future we might also expect drones appropriate to assassinate targets to be developed. These could take the form of a cleaning robot made to explode once a target has been located,[7] or not to be bigger than an insect but able to inject a lethal agent into the target.[8]

Targeted assassination is intuitively an effective method of eliminating political opposition, and the threat of it a way to deter opposition. It is however not a novelty, and politically motivated assassinations have been conducted since ancient times. What is novel about employing drones is that identifying a perpetrator might be made more difficult due to untraceability. When there is ambiguity about who lies behind the death of some individual

[5] See e.g. Adams (2017) and Fariello (2017).

[6] Regimes like Russia already employ similar strategies today, by attempting to delegitimize opposition and portray them as extremists or criminals (Finkel & Brudny 2012)

[7] For this example, see Brundage et al. (2018, p. 27)

[8] For more on the development of coercive drones, see Scharre (forthcoming), Horowitz & Fuhrmann (2014) and Allen & Chan (2017).

the consequences are less severe in terms of e.g. popular dissatisfaction, as it is not clear who can be blamed.[9]

If untraceable drones become more easily accessible, then there is some risk that this will increase the prevalence of politically motivated killings by decreasing the risk involved in performing them. The effect can be strengthened if the assassination can be combined with a plausible cause of death, which might be provided by the information processing systems considered above (e.g. attributing death by a lethal agent to an existing heart disease).

It might however be that effective designs of drones with these capabilities will take many more years to produce, but their availability within 10 years cannot be ruled out. Furthermore, it might turn out that they provide little marginal benefit beyond currently existing methods of assassination, in which case we should not expect their availability to have significant implications.

3. REASONS TO DOUBT THAT AI WILL BE EFFECTIVELY IMPLEMENTED

In the sections above I have presented some of the probable effects of new technologies if they were to be effectively implemented by an authoritarian regime, and subsequently why such a regime would be interested in making use of them. Whether or not such technologies will be effectively implemented is currently an area of high uncertainty however, and there are a few reasons to doubt that they will.

THREATS FROM ELITES

One might think that the greatest threat against a ruling coalition is provided by unruly masses. The available data does not support this claim however. According to political scientist Milan Svolik 68% of authoritarian removals 1945-2008 have been by coup d'états lead by internal elites[10] (mainly the military)[11]. Because most of the technologies considered here seem more obviously applicable to minimizing the threat from popular protests, we should not expect such technologies to allow any given coalition to remain in power indeterminately, because the threat from other elites will remain.

---

[9] An example of this is the 2015 murder of Boris Nemtsov in Moscow. In this case some blame Chechnyan terrorists, which is considered a plausible explanation in the population. Other opposition politicians however attribute it to the Kremlin.

[10] Svolik (2012, p. 5), though Kendall-Taylor & Frantz (2014) present some evidence that protests have become more of a threat to authoritarian regimes today than during most of the 20th century.

[11] Svolik (2012, p. 149)

ORGANIZATIONAL CONSTRAINTS

Several of these technologies, particularly widely adopted information-processing and social credit systems, require a network of actors cooperating in cohesion. This might be unrealistic in most authoritarian states, where the dependence on relations with other elites is vital and often volatile.[12]

For instance, in order to have access to credit card information a regime would need to closely cooperate with a domestic credit card provider (or other actor with access to the data). Such private actors might however ally with a political opponent instead of the ruling coalition.[13]

Unless a ruling coalition can maintain effective control and the continued functioning of new technology, it will not be of much use, and the elite-dynamics of most authoritarian states might make this difficult.

INFRASTRUCTURAL CONSTRAINTS

Beyond the organizational requirements of effectively managing technology, there are also conditions of digital infrastructure which need to be satisfied. For instance, unless there is a functioning payment system with card in the state there obviously will not be much credit card information available to any domestic actor. A similar claim can be made about other sources of data.

For this reason we should not expect AI technology to receive widespread application in states that are unlikely to develop the necessary digital infrastructure to make effective use of it within the near-mid future. Subsequently I believe we should limit these discussions to states such as China or the Gulf States with that infrastructure, as opposed to including authoritarian states like Kyrgyzstan or Zimbabwe who are unlikely to develop it in the near future.[14]

---

[12] Bueno de Mesquita & Smith (2012) provide a good overview of these authoritarian dynamics.

[13] This happened for example 2004 in the Orange Revolution of Ukraine when Petro Poroshenko who owned the TV channel 5 Kanal allied with opposition leaders Viktor Yushchenko and Yulia Tymoshenko against the ruling coalition of Kuchma and Yanukovych.

[14] The judgments on infrastructure are based on the World Bank 2016 Logistics Performance Index, which can be found here:

https://lpi.worldbank.org/international/global?order=Infrastructure&sort=asc

4. WHAT ARE THE EXPECTED CONSEQUENCES IF THE TECHNOLOGY IS IMPLEMENTED?

While we should take the above points into account to avoid alarmism, they are not sufficient to rule out that authoritarian states will put these technologies to use in the future. In this section I note three areas where such developments could plausibly have implications relevant to human rights. I encourage the reader the remember that these developments are conditional on the technology being effectively implemented – which it might not be for the reasons considered in the previous section – and that they are in any case highly uncertain.

PROSPECTS OF POPULAR OPPOSITION

If these AI-technologies are effectively implemented by an authoritarian regime, then this will plausibly have a significant *negative* effect on the prospects of organizing popular opposition.

Regimes that have a substantial informational advantage on their citizens can both take action against dissidents before they constitute a serious threat and credibly threaten to take action against those who demonstrate dissent. This means that individuals who demonstrate high-probability dissent can be detained or otherwise debilitated early on to mitigate their influence, and should disincentivize individuals who otherwise would have considered opposing the regime.

This problem becomes most clear in the ability of the opposition to mobilize for protests. This is by itself difficult, as it requires coordination on the part of many people, and the incentives to participate are not strong when the protesting will get done whether or not you participate and put yourself in harm's way. If any brooding protest can be undercut before it spreads, then popular mobilization will be made even more difficult.

The regime might handle protests in even more subtle ways however. Suppose that the planners of the protest can be kept under strict surveillance (e.g. through the data they leave behind), distorting information released as to the purpose of the protest (e.g. to make it seem conducted by extremists), and more or less subtle punishments guaranteed for participation (e.g. identification by facial recognition causes significant drop in social credit). In that case a regime can even allow a protest to occur without it constituting any real threat to stability, and might even increase popular support for the regime among the general population.

In summary, unless counter-technology is developed to benefit political opposition in authoritarian regimes, I believe there is a high probability that popular opposition will be made more difficult as a consequence of AI-technology employed by the regime if effectively implemented.

PREVALENCE OF VIOLENCE

While these technological developments might increase the regime's power relative to its citizens, there is reason to believe that it would also cause a *decrease* in the use of violence by the regime.

Violent repression is often a costly measure by a regime. This is partly because the use of violence can act as a focal point for protests to mobilize beyond what can be controlled by a regime.[15] In other cases violence can lead to international repercussions or even intervention.[16] Another reason however is that repressing protests requires giving power to the coercive forces, particularly the military. Doing so however increases the risk that the ruling coalition will be subject to a military coup.[17]

For these reasons there is an incentive for a regime to rely on other means of mitigating the threat of popular protest than the use of violence. The technologies above mainly help a regime minimize the likelihood that uncontrolled dissent will occur at all, which is a cheaper way to stay in power than to expose itself to repercussions with the use of violence.

Therefore we can expect authoritarian regimes who effectively implement these AI-technologies to rely less on violence in order to stay in power. That is not the same as a decrease in repression however, which can take other forms. It should for instance make us expect an increase in subtle forms of repression on the basis of highly personal information, e.g. by blacklisting individuals from certain government-supplied services.

DEVELOPMENTS TOWARDS TOTALITARIAN DYNAMICS

Intuitively we might expect the implementation of these technologies to shift the dynamic between society and the state in ways that would make give it totalitarian traits.

Authoritarian regimes are non-ideological and function mainly as a form of organization between citizens and the regime. In totalitarian regimes on the other hand the state plays a more intimate role in people's lives, and by influencing norms and ideology they blur the distinction between state and society.[18]

---

[15] E.g. in Romania 1989 Nicolae Ceausescu was deposed by popular protests following a prior crackdown on protesters.

[16] This was for instance seen in the intervention in Libya 2011.

[17] This is what happened to Milton Obote of Uganda in 1971 when he gave increased powers to his army chief Idi Amin in order to suppress dissidents, who then performed a coup against him.

[18] See Linz (2000). Kazakhstan is a typical example of an authoritarian regime, while North Korea is a typical example of a totalitarian regime.

Given that the efficacy of some of the technologies above relies on promoting certain kinds of behavior, while discouraging others, we could intuitively expect their implementation to influence norms in society and how people behave in their every-day life. This is particularly the case when benefits and punishments can be based on highly personal choices and behaviors (e.g. choice of reading, conversation topics or drinking habits) that regimes did not have information on before, but now do due to the data that individuals inadvertently provide. Under such circumstances individuals would have an incentive to adapt these more personal choices and behaviors to avoid repercussions.

However, it is not clear that regimes would like these developments to occur, as the perceived infringement on privacy might cause dissatisfaction and dissent. Furthermore, other factors than technological development (e.g. leadership or culture) might be much more important to determine these dynamics.[19] Therefore, whether or not AI-technology has a significant effect on developments towards totalitarianism cannot be determined with any confidence now. By studying e.g. the Xinjiang region in China in the coming years however we should be able to discern the directions of such developments more clearly.

5. WHAT CAN THE UNITED STATES DO?

Any decision to take action should be made with the awareness that these developments are highly uncertain. These are many other factors which I and others might have failed to identify which could increase or decrease the impact of AI on authoritarian governance and human rights. In this report I have remarked on some reasons to be skeptical of a major impact. These reasons are not decisive however, and leave open the possibility that AI will have a significant impact possibly in the ways outlined above.

For this reason one major focus should be on understanding where these developments are heading. We have little available data at this point, but more will be available in the coming years. Understanding both what technological developments seem plausible based on the recent research, and how authoritarian regimes make use of the existing technology will be vital. Special attention should therefore be given to cases like Xinjiang in China where advanced AI technology is employed for the purpose of mitigating the risk of opposition.

If the US or other intentional actors would desire to take more direct action, it is not clear what would be effective. The necessary technology is often dual-use in nature, meaning that no international constraints can be put on it even in principle without also constraining other non-authoritarian applications of it. Constraining development is also difficult, as it can be

[19] The Soviet Union under Stalin for example was a typical example of a totalitarian regime, while under Khrushchev and Brezhnev it developed away from these totalitarian traits and became more of an authoritarian regime. Under the same period however there were many technological developments which would have facilitated totalitarian-style governance (e.g. the spread of the television).

done by private actors domestic to the authoritarian regime. The Social Credit System in China for instance is being developed by the Chinese company Sesame Credit. Those states that have the domestic capacity for development can in turn export this to other states.

Another alternative is to fund research into counter-technology to increase the power of political opposition in authoritarian regimes. This could be technology that provides decentralized communication that the regime is unable to influence, or software that provides misleading data to distort the information that the regime has on individuals. Whether such technology would be of more use to political opposition than the technologies considered here to a regime is currently an open question that likely deserves further research.

REFERENCES

Adams, Terrence. “AI-Powered Social Bots,” 2017. https://arxiv.org/abs/1706.05143.

Allen, Greg, and Taniel Chan. “Artifical Intelligence and National Security.” Belfer Center for Science and International Affairs, 2017.

Brundage et al., Miles. “Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation,” 2018.

Bueno de Mesquita, Bruce, and Alastair Smith. *The Dictator’s Handbook: Why Bad Behavior Is Almost Always Good Politics*. PublicAffairs, 2011.

Ding, Jeff. “Deciphering China’s AI Dream”, tech. rep. Future of Humanity Institute, University of Oxford, 2018.

Ferguson, Andrew Guthrie. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. NYU Press, 2017.

Finkel, Evgeny, and Yitzhak M. Brudny. “Russia and the Colour Revolutions.” *Democratization* 19, no. 1 (February 2012): 15–36. https://doi.org/10.1080/13510347.2012.641297.

Gusterson, Hugh. *Drone: Remote Control Warfare*. MIT Press, 2016.

Horowitz, Michael C, and Matthew Fuhrmann. “Droning On: Explaining the Proliferation of Unmanned Aerial Vehicles,” 2014, 30.

Kendall-Taylor, Andrea, and Erica Frantz. “How Autocracies Fall: The Washington Quarterly: Vol 37, No 1,” 2014.

Linz, Juan J. *Totalitarian and Authoritarian Regimes*. Lynne Rienner Publishers, 2000.

Matz, Sandra C, and Oded Netzer. “Using Big Data as a Window into Consumers’ Psychology.” *Current Opinion in Behavioral Sciences* 18 (December 2017): 7–12.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. Norton, Forthcoming.

Schneier, Bruce. *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. Norton, 2015.

Sundsøy, Pål. “Big Data for Social Sciences: Measuring Patterns of Human Behavior through Large-Scale Mobile Phone Data,” 2017.

Svolik, Milan. *The Politics of Authoritarian Rule*. Cambridge University Press, 2012.