Research Paper

Heather M. Roff
International Security Department and US and the Americas Programme

May 2017

# Advancing Human Security through Artificial Intelligence

CHATHAM HOUSE
The Royal Institute of International Affairs

## Summary

- AI is one potential way to enable real-time, cost-effective and efficient responses to a variety of human security-related issues. AI applications related to search, classification and novel pattern recognition can help to correlate and extract content and meaning from multiple sources.

- AI planning applications can quickly, reasonably and reliably enable users to carry out complex and multi-stage actions in disaster relief operations. The need for this capability is illustrated by the UN's average estimated response times for new peacekeeping missions. The UN estimates that when a new crisis emerges – that is, a crisis that involves violence and mass threats to human rights – the estimated response time to plan and field a credible peacekeeping mission is six to 12 months.

- It is important to still use normative principles to interrogate the purpose and effects of AI applications as they relate to empowerment and human security. AI is as likely to disempower people as it is to empower them, and so it is necessary to use ethical principles to guide the creation and deployment of AI systems.

- Human security is not for an elite few, and so the capabilities of AI must be within everyone's grasp. When it comes to applications related to disaster relief, conflict prevention, human rights protection and justice, it is imperative that wider schemes of data sharing are employed by individuals, groups, NGOs and governments. However, it is simultaneously imperative that data, through sharing and acquisition, are also protected to the greatest possible extent.

# Introduction

*This is a draft of the author's contribution to a forthcoming Chatham House report on artificial intelligence, to be published in the autumn of 2017.*

Over the next two decades human security will be confronted by significant challenges. With continuing global warming there will be increased temperatures, rising sea levels and more extreme weather events.[1] These changes will lead to a scarcity of resources, particularly of water, food and energy.[2] The hardest hit areas of the globe are most likely to be those already suffering from various types of instability, violence and unrest, such as sub-Saharan Africa, Pakistan and parts of the Middle East.[3] The confluence of climate and political refugees will undoubtedly compromise local, regional and the international community's ability to secure individuals from fear and want.

Concomitantly, growing connectedness via social media and changes in labour and production due to advancing technology proliferation will also place new stresses on the world economy, as well as create new shifts in political and economic power. Microsoft predicts that by 2025, 4.7 billion people will use the internet – just over half the world's expected population at that time – and of that number, 75 per cent of users will be in emerging economies.[4] With an estimated 50 billion connected devices, all generating mass amounts of data, information will become an even more powerful tool for development, coordination, persuasion and coercion.[5] Moreover, these individuals will enter new (and old) economic market sectors, and be faced with increasing automation and the stresses of wage devaluation.

In this future world, increasingly divided on demographic, economic and technological lines, achieving human security will not be without its difficulties. Systemic challenges, such as climate change and war, and more localized threats like social, economic or political disruptions are almost certain.

One way to meet these challenges is through novel applications of technology, particularly artificial intelligence (AI). AI holds much promise to enable the international community, governments and civil society to predict and prevent human insecurity. With increased connectivity, more sophisticated sensor data and better algorithms, AI applications may prove beneficial in securing basic needs and alleviating or stopping violent action.

[1] Global Facility for Disaster Reduction and Recovery (2014), Understanding Risk in an Evolving World: Emerging Best Practices in Natural Disaster Risk Assessment, World Bank, https://www.gfdrr.org/sites/gfdrr/files/publication/Understanding_Risk-Web_Version-rev_1.8.0.pdf.

[2] Veldkamp, T.I.E, Wada, Y., de Moel, H., Kummu, M., Eisner, S., Aerts, J. C. J. H., and Ward, P. J. (2015), 'Changing Mechanism of Global Water Scarcity Events: Impacts of Socioeconomic Changes and Inter-annual Hydro-climatic Variability,' Global Environmental Change, 32: pp.18–29, http://www.sciencedirect.com/science/article/pii/S0959378015000308.

[3] World Bank (2011). 'Climate Change Adaptation and Natural Disasters Preparedness in the Coastal Cities of North Africa', http://www.worldbank.org/en/news/feature/2011/06/04/north-african-coastal-cities-address-climate-change-and-natural-disasters. Petley, D. N. (2010), 'On the Impact of Climate Change and Population Growth on the Occurrence of Fatal Landslides in South, East and SE Asia', Quarterly Journal of Engineering Geology & Hydrogeology 43(4): pp. 487–96, http://qjegh.lyellcollection.org/cgi/doi/10.1144/1470-9236/09-001; Mueller, V., Gray, C., and Kosec, K. (2014), 'Heat Stress Increases Long-term Human Migration in Rural Pakistan', Nature Climate Change 4(3): pp. 182–85, http://www.nature.com/nclimate/journal/v4/n3/full/nclimate2103.html.

[4] Burt, D., Kleiner, A., Paul Nicholas, J., and Sullivan, K. (2014), 'Cyberspace 2025: Today's Decisions, Tomorrow's Terrain', Microsoft https://www.microsoft.com/security/cybersecurity/cyberspace2025/#chapter-1.

[5] In 2016, internet traffic reached 1.3 Zettabytes. A Zettabyte is 1,000,000,000,000,000,000,000 bytes of information. This is a 50 per cent increase from 2011–14. Thus it is likely that if this trend continues, then by 2025 traffic will increase to 4.38 Zettabytes. However, if there are advances in storage, more devices, and more streaming, then traffic will exceed this. See, http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html; http://www.tvtechnology.com/resources/0006/welcome-to-the-zettabyte-era/278852.

Section one of this paper lays out the principles of the United Nations' approach to human security, as well as more critical viewpoints. Section two argues that many of the conflict and development problems that the international community, states and civil society face can be ameliorated or solved by advancements in AI. In particular, algorithms adept at planning, learning and adapting in complex data-rich environments could permit stakeholders to predict and coordinate responses to many types of humanitarian and human security related situations. Finally, section three argues that to ensure broad access, transparency and accountability, especially in countries that may be prone to human security emergencies, the relevant AI ought to be open source and sensitive to potential biases.

## Human Security

Human security is a concept that takes the human – as opposed to the state – as the primary locus of security. As Sadako Ogata, former United Nations High Commissioner for Refugees, has written, 'Traditionally, security issues were examined in the context of "State security", i.e. protection of the State, its boundaries, its people, institutions and values from external attacks,' and that individuals were to only be secured by way of the state.[6] Yet, with changes in the post-Cold War era, where external threats to state security declined and internal threats of intra-state violence increased, many policymakers, practitioners and scholars required a new lens through which to understand these internal conflicts.

Indeed, in 1994, the United Nations Human Development Report concluded:

> without the promotion of people-centered development, none of our key objectives can be met – not peace, not human rights, not environmental protection, not reduced population growth, not social integration. It will be a time for all nations to recognize that it is far cheaper and far more human to act early and to act upstream than to pick up the pieces downstream, to address the root causes of human insecurity rather than its tragic consequences.[7]

From 1994 onwards, many different avenues for examining the concept of human security emerged.[8] Central to all, however, was the focus on the nexus between development, human rights (protection and promotion), and peace and security. The premise that people possess dignity logically entailed that they ought to be 'free from fear' and 'free from want'.[9] To establish what this expansive formulation meant, the Human Development Report 2000 identified seven elements comprising human security.[10]

[6] Ogata, S. (2015). 'Striving for Human Security', United Nations Chronical, No.1 and 2: p. 26.
[7] United Nations Development Program (1994), 'The Human Development Report', Oxford: Oxford University Press: p. iii.
[8] Cf: Human Security Now (2003), the United Nations Trust Fund for Human Security, 2005 World Summit Outcome Document, especially paragraph 143. General Assembly Resolution 66/290 of 2012 addressing explicitly human security and requesting additional reports on lessons learned from 'human security experiences at the international, regional and national levels.' A/Res/66/290.
[9] Ibid.
[10] The Human Development Report's findings as cited in: Paris, R. (2001), 'Human Security: Paradigm Shift or Hot Air?', *International Security*, 26(2): p. 90.

### Table 1: Human Security Dimensions

| Object of Security | Content |
|---|---|
| Economic Security | Freedom from poverty |
| Food Security | Access to food |
| Health Security | Access to healthcare and protection from disease |
| Environmental Security | Protection from environmental pollution and depletion |
| Personal Security | Physical safety (e.g. freedom from torture, war, criminal attacks, domestic violence, drug use, suicide and traffic accidents) |
| Community Security | Survival of traditional cultures, ethnic groups and the physical security thereof |
| Political Security | Freedom to enjoy civil and political rights, freedom from political oppression |

Source: Author’s research

Following from this, the UN also framed human security as emerging from the achievement of ‘sustainable development’ and various established international development goals.[11] Human security should be seen as complementary to state security, and measures taken to uphold human rights and build local or regional security capacities through non-coercive measures will simultaneously generate greater stability and development.

However, despite the UN rhetoric, the notion is not without critics. Some claim that it is ‘so broad that it is difficult to determine what, if anything, might be excluded from the definition of human security.’[12] The problem, of course, is that if human security as a concept includes such extensive facets of human existence, in reality it means little and impedes the formulation of sound policy. Others point out that the two key elements that define human security have not been treated equally, with progress on the ‘freedom from want’ portion subjugated to issues related to war and violence, in an attempt to make ‘freedom from fear’ a reality.[13] Such prioritization reflects various realities of power politics, and demonstrates how some states view their obligations towards capacity-building in areas that have little, if any, strategic or economic interest for them. Indeed, even responses to global health crises appear to mirror power politics and national security interests.[14]

From a practical perspective, difficulties of adequately and appropriately responding to potential, emerging or ongoing human security crises are endemic. One might claim this is due to the fact that the concept is over expansive, but this objection notwithstanding, achieving human security may have more to do with the inability of various stakeholders, such as the UN, civil society and nation states to monitor, predict and react to a crisis. Since there are linkages between development, human rights and security, the number of different actors with varying priorities and knowledge bases is high. These actors become disconnected, and may even be forced to work against one

[11] A/RES/66/290.
[12] Paris (2001), ‘Human Security’, p. 90.
[13] Schittecatte, C. (2006), ‘Toward a More Inclusive Global Governance and Enhanced Human Security’ in Maclean, S., Blakc, D.R., and Shaw, T. M. (eds) (2006), *A Decade of Human Security: Global Governance and New Multilateralisms*, Ashgate Press: p. 132.
[14] O’Manique, C. (2006), ‘The “Securitization” of HIV/AIDS in Sub-Saharan Africa: A Critical Feminist Lens’ in Maclean, S., Blakc, D.R., and Shaw, T.M. (eds) (2006), *A Decade of Human Security: Global Governance and New Multilateralisms*, Ashgate Press: p. 165.

another for resources or funding. Lack of communication and information exchange between these actors may only exacerbate problems.

Thus, to counter some of these objections, especially in light of the challenges in the coming 10–15 years, it is necessary to devise novel approaches to ameliorate human insecurity and vulnerability. Specifically, by taking a closer look at how new AI applications can help a variety of stakeholders predict, plan and respond to human security crises.

## Securing the Human through AI

The expansive and interconnected set of factors that affect human security is not the only challenge to alleviating human insecurities.[15] There are three antecedent constraints on human security-related activities: the inability to know about threats in advance; the inability to plan appropriate courses of action to meet these threats; and, the lack of capacity to empower stakeholders to effectively respond. Tackling these constraints could save thousands of lives. The use of AI is one potential way to enable real-time, cost-effective and efficient responses to a variety of human security-related issues.

However, it should be noted that AI is not a panacea. As an inter- and multi-disciplinary approach to 'understanding, modeling, and replicating intelligence and cognitive processes by invoking various computational, mathematical, logical, mechanical, and even biological principles and devices,' it is effective at carrying out certain tasks but not all.[16] Much depends on the task at hand. For example, AI is very good at finding novel patterns in mass amounts of data.[17] Where humans are simply overwhelmed by the volume of information, the processing power of the computer is able to identify, locate and pick out various patterns. Moreover, AI is also extremely good at rapidly classifying data. Since the 1990s, AI has been used to diagnose various types of diseases, such as cancer, multiple sclerosis, pancreatic disease and diabetes.[18] However, AI is not yet able to reason as humans do, and the technology is far from being a substitute for general human intelligence with common sense.

In short, AI looks to find various ways of using information communication technologies, and sometimes robotics, to aid humans and complete tasks. How the AI is created (its particular architecture) and its purpose (its application) can vary significantly. For the purposes of this paper, however, the tasks that AI are particularly well suited to, in the human security domain, are related to planning and pattern recognition, especially given big data problem sets. Considering the current considerable capabilities in these areas, it is reasonable to estimate that in the coming years AI will be able to overcome the three constraints on human security-related activities mentioned earlier.

[15] Interestingly, one can think of the human security project as a secularized version of the Christian or Jewish notions of 'heaven' or the Islamic idea of 'Jannah', as well as other non-monotheistic religions. The important thing to note is that there is a notion that when one transcends to this place, one is free from all evils, fears, vulnerabilities, and needs. Paradise in whatever form is the promise that all ills from the human condition are removed. Human security, likewise, argues for the removal of these same things.

[16] Frankish, K., and Ramsey, W.M. (2014), 'Introduction' in Frankish, K. and and Ramsey, W.M. (eds) (2014), *The Cambridge Handbook of Artificial Intelligence*, Cambridge: Cambridge University Press: p. 1.

[17] Sagiroglu, S. and Sinanc, D. (2013), 'Big data: A review', Collaboration Technologies and Systems (CTS), 2013 International Conference, San Diego, CA: pp. 42–47. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6567202&isnumber=6567186.

[18] Amato, F., López, A., Peña-Méndez, E.M., Vahara, P., Hampi, A., Havel, J. (2013), 'Artificial Neural Networks in Medical Diagnosis', Journal of Applied Biomedicine, 11(2): pp. 47–58.

## Knowledge

The ability to generate knowledge is no easy feat. Knowledge is subtly different from mere data, which are just an amalgamation of discrete and observable facts or inputs that lack meaning without analysis and context. Only when sets of data are given meaning do they become information, which feeds into and builds knowledge.

There are two obstacles to developing knowledge to tackle human security challenges. The first comes from the considerable amount of data that future generations will generate. Everything from individual output from wearable devices to content created on new communication or social media platforms will saturate the world in an ocean of bits and bytes. There will be a requirement for a way to make these data, from the billions of new devices and millions of new users, intelligible.

Second, because human security crises can emerge from anywhere and result in various physical or economic social impacts, there will be an urgent need to disentangle discrete flows of data specific to the various vulnerabilities. Such data flows can be specific to one type of phenomenon, such as extreme weather events prediction,[19] or could even be more diffuse, such as searching and correlating various events or combining data sets to look for indicators of conflict onset or escalation.

AI applications related to search, classification and novel pattern recognition can help to correlate and extract content and meaning from multiple sources. For example, the Early Model-Based Event Recognition using Surrogates (EMBERS) application forecasts key events up to eight days before they happen with a 94 per cent accuracy rate. EMBERS is a '24x7 forecasting system for significant societal events using open source data including tweets, Facebook pages, news articles, blog posts, Google search volume, Wikipedia, meteorological data, economic and financial indicators, coded event data, online restaurant reservations (Open Table), and satellite imagery' to forecast events and notify users in real time.[20] This far outstrips current abilities in traditional political science for prediction and explanation of war, where scholars trudge through and manually code content analysis.[21] In the future, it is more likely that scholars or practitioners will use intelligent artificial agents to process real natural language to comprehend text, rather than merely looking for word frequencies and correlations thereby deepening the capabilities of programs like EMBERS even further.[22]

Another area for the use of AI in human security is health. There are various applications and abilities in this domain but a few are of particular note. First, AI's ability to classify and identify

[19] Bauer, P., Thrope, A., and Brunet, G. (2015), 'The Quiet Revolution of Numerical Weather Prediction', *Nature*, 525: pp. 47–55.

[20] Ramakrishanan, Naren, see, http://dac.cs.vt.edu/research-project/embers/.

[21] For example, some scholars do this simply by frequency analysis (or counting the times a particular word or phrase comes up). In one study, Thomas Chadefaux searched Google's database of English-speaking newspapers for a period of 99 years looking for the frequency of various key words associated with 'tensions'. He then correlated this to the Correlates of War dataset to examine if there were any relationship between reported tension escalation and war onset. Unsurprisingly, he found a correlation. Chadefaux, T. (2013) 'Early Warning Signals for War in the News', Journal of Peace Research, 51(1): pp. 5–18. More technologically savvy are those scholars like Mike Thelwall, who utilize a query program to social media pages, like Twitter, to record the frequency of various topics, the location, the language and even the gender or sentiment of the searched word, phrase or query. See, Thelwall, M. 'Sentiment Analysis and Time Series with Twitter', http://mozdeh.wlv.ac.uk/resources/TwitterTimeSeriesAndSentimentAnalysis.pdf.

[22] One of the present state of the art AI companies, Google Deepmind, is already working to create a deep neural network that can read real documents, answer complex questions and do so 'with minimal prior knowledge of language structure'. Deepmind's AI is able to perform reasonably well at learning real natural language, as it can presently answer about 64 per cent of queries correctly. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). 'Teaching Machines to Read and Comprehend', http://arxiv.org/pdf/1506.03340v1.pdf.

images allows it to recognize patterns more quickly and accurately than people. This has been particularly true in the diagnosis of certain types of cancer.[23]

However, one need not be in a state-of-the-art facility or hospital to receive this type of care. Mobile phones are increasingly being used for bioanalytical science, including digital microscopy, cytometry, immunoassay tests, colorimetric detection and healthcare monitoring. The mobile phone 'can be considered as one of the most prospective devices for the development of next-generation point-of-care (POC) diagnostics platforms, enabling mobile healthcare delivery and personalized medicine.'[24] With advancements in mobile diagnostics, millions more people may be able to monitor and diagnose health-related problems, especially given the estimated increased use in mobile data and devices.

Moreover, with increased connectivity through social media, AI can leverage big data in ways that encourage the uptake of preventive measures. For instance, one application uses machine learning to estimate real-time problematic areas or establishments that may cause food borne illnesses.[25] This particular application alerts health inspectors in real time to potential outbreaks of food-borne illness so that they may take immediate action. In essence, the ability to know what is happening, when and where is the first step in addressing vulnerability.

## Planning

In addition to acquiring and contextualizing knowledge, it is also essential to have the ability to plan an appropriate response. Algorithms related to planning can quickly, reasonably and reliably enable users to carry out complex and multi-stage actions. The need for this facility can be illustrated by examining the UN's average estimated response times for new peacekeeping missions. The UN estimates that when a new crisis emerges – that is, a crisis that involves violence and mass threats to human rights – the estimated response time to plan and field a credible peacekeeping mission is six to 12 months.[26] There are two reasons for this; first, is the strict structure and process of formulating peacekeeping missions.[27] Second, attempts 'to develop better arrangements for rapid deployment have been repeatedly frustrated by austerity and a zero-growth budget.'[28] In short, politicking within the bureaucracy and money constraints limit the UN's ability to act swiftly.

Additionally, there are serious problems related to logistics once a mission is approved. The ability to rapidly and reliably estimate, plan and deliver equipment, supplies and services is 'a constant demand across all field operations'. As such, the UN created a Department of Field Support in 2007, whose focus is on developing a standardized approach to forecasting and planning for new

[23] Perry, P. (2016), 'How Artificial Intelligence will Revolutionise Healthcare', Big Think, http://bigthink.com/philip-perry/how-artificial-intelligence-will-revolutionize-healthcare.
[24] Vashist, S.K., Mudanyali, O., Schneider, E.M. et al. (2014), Anal Bioanal Chem 406: p.3263. doi:10.1007/s00216-013-7473-1.
[25] Sadilek, A., Kautz, H., DiPrete, L., Labus, B., Portman, E., Teitel, J., and Silenzio, V. (2016), 'Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media', Association for the Advancement of Artificial Intelligence, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.723.8225&rep=rep1&type=pdf.
[26] Langille, H.P. (2014), 'Improving United Nations Capacity for Rapid Deployment', New York: International Peace Institute: p. 1.
[27] The process for forming a new operation involves 6–7 steps, depending if one considers evaluation as part of the process. The steps involve: initial consultations with a myriad of actors in and outside the UN, technical field assessments, a Security Council resolution, the appointment of senior officials, then planning, deployment and reporting to the Security Council. See, http://www.un.org/en/peacekeeping/operations/newoperation.shtml. The mere fact that planning takes place after all of the political hurdles have been passed increases the length of time to get a mission off the ground. Given that many humanitarian emergencies are ongoing conflicts, much of the information required is already available and could be used to formulate initial operational plans or contingency plans.
[28] Langille (2014), Improving United Nations Capacity, p. 1.

operations; human resource planning; supply chain logistics; and evaluative service centres for mission re-tasking and re-planning.[29]

While politics may get in the way of ensuring human security, there are technological solutions that may help. Specifically, advancements in planning algorithms are promising, particularly in emergency response situations. Emergency logistics scheduling is an application that deals with 'the need to identify, inventory, dispatch, mobilize, transport, recover, demobilize, and accurately track human and material resources in the event of a disaster.'[30] Depending upon the type of disaster or crisis, various linear or nonlinear, single- or multi-objective algorithms are presently available for this purpose. These algorithms can identify ideal station points,[31] routing paths for distribution and evacuation,[32] the amount of relief required,[33] and scheduling.[34]

In the coming years, these algorithms are likely to improve, and as they are able to undergo further modification, such as through genetic evolution, learning and adaptation, their abilities will sharpen. Indeed, there is no ostensible reason why the primary objectives of the UN's Department of Field Support could not be met by using planning AI, thereby automating the majority of its tasks. By doing this, time spent on forecasting and planning would fall, and its implementation would cut costs related to human resources and training and remove many redundancies and barriers in supply chain logistics. This would improve the efficiency of any service centres and potentially result in rapid deployment of forces at a reduced cost.

Governments, NGOs and civil society groups can also avail themselves of this AI. NGOs and civil society groups may in fact be best placed to trial these technologies, as they are often not hampered by political obstacles or byzantine bureaucratic rules. They may have greater flexibility to try out new approaches and improve confidence in those ideas. This would help immensely in various human security-related situations, such as complex humanitarian crises – those that are combinations of political and natural disasters – as they could examine vast amounts of data relating to available resources, use satellite imagery or images from surveillance aircraft to map affected terrain, as well as find survivors, and thereby estimate the necessary resource requirements given limitations on time, goods and manpower.

## Empowerment

As human security has such a broad definition there are almost limitless ways in which AI can help individuals to be more secure. The key is that such applications empower actors and enable them to make better decisions. Determining how AI can do this without exacerbating existing inequalities or unintentionally creating situations of insecurity is also a consideration. As UN resolution 290

[29] Ibid, p. 13.
[30] Chang, F-S., Wu, J-S., Lee, C-N., and Shen, H-C. (2014), 'Greedy-Search-Based-Multi-Objective Genetic Algorithm for Emergency Logistics Scheduling', Expert Systems with Applications, 41: p.2947.
[31] Chang M-S., Tseng, Y-L, and Chen, J-W. (2007), 'A Scenario planning approach for the flood emergency logistics preparation problem under uncertainty', Transportation Research Part E: Logistics and Transportation Review, 43: pp. 737–754.
[32] Zhang, X., Zhang Z., Zhang Y., Wei, D. and Deng, Y. (2013), 'Route Selection for Emergency Logistics Management: A Bio Inspired Algorithm', Safety Science, 54: pp. 87–91.
[33] Zhou, Q., Huang, W., and Zhang, Y. (2011), 'Identifying Critical Success Factors in Emergency Management Using a Fuzzy DEMATEL Method', Safety Science, 49: pp. 243–252.
[34] Sheu, J.B. (2007), 'An Emergency Logistics Distribution Approach for Quick Response to Urgent Relief Demand in Disasters', Transportation Research Part E: Logistics and Transportation Review, 43: pp. 687–709.

states, 'Human security calls for people-centred, comprehensive, context-specific and prevention-oriented responses that strengthen the protection and *empowerment* of all people and all communities,' acknowledging that it also 'equally considers civil, political, economic, social and cultural rights.'[35]

Empowerment is thus not easily achieved. If all human rights are of equal value, then trade-offs between them are not easily resolved. Furthermore, it is unclear how AI might contribute or detract from such rights. For example, AI's ability to find patterns in big data is an asset in diagnosing diseases such as cancer, but it may not be desirable when the pattern that it finds is controversial in some way, such as if it is obviously racist, sexist or extremist. Such patterns may well exist because of the available data or because of existing inequalities or systemic biases in a society. AI could merely be making visible the tyranny of the majority in this situation by classifying particular people, groups or behaviours in various categories.[36] Take, for instance, the Microsoft Twitter bot that within 24 hours of being deployed on Twitter was turned into a racist, sexist and genocidal chatbot due to the amount of these types of phrases being 'fed' to the bot by other Twitter users. Microsoft had to deactivate the bot immediately. When it was accidentally activated a few weeks later, it once again began making inappropriate tweets.[37]

In the most concerning of cases, AI could actually disempower people. This was demonstrated by an algorithm used to predict recidivism rates, which incorrectly scored black defendants in the US along all metrics, such as the likelihood of re-offending or committing violent acts.[38] These estimates were considered as evidence in sentencing recommendations, and because of systemic race and gender biases against classes of individuals those being sentenced were unfairly and systematically sanctioned.

Thus, it is necessary to interrogate the purpose and effects of AI applications as they relate to empowerment and human security. To facilitate this, one might think of utilizing particular principles as normative guides. An example might be applying something like a Rawlsian principle of justice – that aims to give the greatest benefit to the least-advantaged members of society – to AI applications.[39] This would provide a general and high-level principle with which to test various context-specific cases to estimate the likely effects of a given AI application. To succeed, it would require AI application developers to adopt an attitude that reflects both their technical know-how and a consideration of broader social elements. In particularly sensitive applications, such as in those related to potential human rights transgressions, further scrutiny would be warranted to ensure that the data provided was robust and diverse, as well as designed to be mindful of the value

[35] A/Res/66/290. Italics added.

[36] Take for example the recent case where an image classification algorithm on Google classified images of African-American individuals as gorillas. Google apologized for the incident. BBC News (2015), 'Google apologises for Photos app's racist blunder', 1 July 2015, http://www.bbc.com/news/technology-33347866 .

[37] Kosoff, M. (2016), 'Microsoft's Racist Millennial Twitter Bot Strikes Again', Vanity Fair, http://www.vanityfair.com/news/2016/03/microsofts-racist-millennial-twitter-bot-went-haywire-again.

[38] White defendants were estimated to be less likely to reoffend and be of lower risk of committing future crimes (of both a nonviolent and violent nature). Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016), 'How we Analyzed the COMPAS Recidivism Algorithm', Pro Publica, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[39] Rawls, J. (1971), *A Theory of Justice*, Harvard University Press, p. 75. This is otherwise known as the 'difference principle'. The principle has come under much scrutiny for many reasons in its history. For more, see Amartya Sen's excellent essay 'Equality of What?' lays out many of the problems and a potential solution, Sen, A. (1979), 'Equality of What?', The Tanner Lecture on Human Values, http://tannerlectures.utah.edu/_documents/a-to-z/s/sen80.pdf. Sen's contention is that a difference principle is not sufficient because it does not sufficiently address individual people's needs and capabilities. He instead would see a capabilities approach. For more on the capabilities approach, see: Nussbaum, M. (2011), *Creating Capabilities: The Human Development Approach,* Harvard University Press.

of these rights. Recent work by the United States Federal Trade Commission and the White House on the need for further regulation of big data and algorithmic-based decisions, such as through best practices, codes of conduct and even existing or new laws, is also important.[40]

## Equitable, Transparent and Accountable

Ultimately, as more data are used to influence decisions, and as algorithms are increasingly utilized to shape, guide or make these decisions, humans must be vigilant in asking for transparency, accountability, equity and universality from these applications. These are all elements of ensuring a just distribution and accounting of the benefits of AI. From a policy standpoint, it is essential to know what data are used, an AI model's guiding assumptions, as well as the kinds of practices that developers utilize. An understanding of these components will allow for accurate estimates of the likely effects of an AI application.

As it is known that technology is not value neutral, but is in fact a human creation guided by (often implicit) particular values, policymakers ought to ensure that such technologies and their benefits are accessible to everyone in an open source format. Human security is not for an elite few, and so the capabilities of AI must be within everyone's grasp.[41] When it comes to applications related to disaster relief, conflict prevention, human rights protection and justice, it is imperative that wider schemes of data sharing are employed by individuals, groups, NGOs and governments. However, it is simultaneously imperative that data, through sharing and acquisition, are also protected to the greatest possible extent. Health, in particular, is one immediate area of focus for privacy, transparency and accountability policies, best practices and regulation.[42]

These concerns are important to human security activities. Take, for instance, the movement to require biometric identification to receive humanitarian aid. While the intention is to track individuals to reduce fraud, this data could also be used for political oppression. The UN High Commission for Refugees, for example, argues for increased biometric identification, but this data is shared by a variety of actors for multiple purposes.[43] Yet there is no discussion of data protection, and there is a gap in regards to policy guidance or compliance. Thus, in situations of complex humanitarian disasters, where refugee safety is certainly of concern, regulations need to be established.[44]

[40] United States of America, Executive Office of the President (2016), 'Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights', https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf. The Federal Trade Commission (2016), 'Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues', https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report.

[41] The OpenAI Initiative is one attempt at making AI open source and available to all. Unfortunately, however, the amounts of data required for robust AI is usually beyond the reach of private individuals and remains in the hands of governments and corporations. https://openai.com/about/.

[42] Google Deepmind, for example, has instituted a group of independent reviewers to oversee Deepmind Health applications. The reviewers are trusted public figures, and form a multi-stakeholder perspective. They are not paid, have no conflicts of interest, and are put in place to scrutinize Deepmind Health's 'work, its handling of data, investigate data sharing agreements, understand the product roadmaps and critique the handling of health data' from the UK's National Health Service, https://deepmind.com/health/independent-reviewers.

[43] Ismail, Y. (2006), 'Fingerprints Mark New Direction in Refugee Registration', UNHCR, http://www.unhcr.org/456ede422.html.

[44] For instance, one refugee from Myanmar was quoted as saying that 'I don't know what this is for, but I do what UNHCR wants me to do.' This sort of trust in the UN system's ability to protect refugees and their digital information is questionable. The same article cites that 84,000 biometric UK prison records were compromised when they were left on an unencrypted USB stick in an unlocked drawer. See Currion, P. (2015), 'Eyes Wide Shut: The Challenges of Humanitarian Biometrics', IRIN, http://www.irinnews.org/opinion/2015/08/26/eyes-wide-shut-challenge-humanitarian-biometrics.

In short, AI that enables human security must, by its very nature, ensure that it is aimed at minimizing human insecurity, maximizing human empowerment, and is as equitable, transparent and accountable as possible. The consequences of algorithms misclassifying or failing to plan appropriately could be catastrophic. Therefore, good policy, regulation and accountability measures need to be in place. These may range from pre-emptively instituting a set of best practices to remedial or coercive measures after the fact. Whatever the situation, humanity must not walk into a future increasingly influenced by AI and claim the equivalent of an AI 'Twinkie Defense.'[45] Therefore, AI that is sensitive to context, vulnerability and capacity-building, but guided by good judgment, foresight and principles of justice would be most beneficial for all.

[45] The 'Twinkie Defense' was used by Dan White's defence lawyer to argue that when he murdered Harvey Milk he was suffering from mental impairment due to too much sugar consumed from eating Twinkies and was thus not responsible for his act, https://www.law.cornell.edu/wex/twinkie_defense.

## About the author

Dr Heather Roff is currently a senior research fellow in the Department of Politics and International Relations at the University of Oxford, a research scientist in the Global Security Initiative at Arizona State University, and has held faculty positions at the Korbel School of International Studies at the University of Denver, the University of Waterloo, and the United States Air Force Academy. She is also an associate research fellow at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge, a research fellow at New America in the Cybersecurity Initiative and the Future of War Project. Her research interests include the law, policy and ethics of emerging military technologies, such as autonomous weapons, artificial intelligence, robotics and cyber, as well as international security and human rights protection.

## Acknowledgments

The author would like to thank the many sponsors of her research that helped to inform this piece, including the Future of Life Institute, the government of Canada, and Google Deepmind.

Independent thinking since 1920

Chatham House does not express opinions of its own. The opinions expressed in this publication are the responsibility of the author(s).

ISBN 978 1 78413 203 3

This publication is printed on recycled paper.

The Royal Institute of International Affairs
Chatham House
10 St James's Square, London SW1Y 4LE
T +44 (0)20 7957 5700  F +44 (0)20 7957 5710
contact@chathamhouse.org  www.chathamhouse.org
Charity Registration Number: 208223