Digital Human Security 2020

Human security in the age of AI: Securing and empowering individuals

Foreword

We at the ICT4Peace Foundation have been working on Information Communication Technologies (ICTs) and peace and security issues for the past 14 years. Much has changed in terms of the potential for international coordination, the speed of information, the rise of social media and the advancements made by AI. Over the past 18 months ICT4Peace has focussed particularly on the impact of Artificial Intelligence and Cybersecurity on society and individuals and resulting peace-time threats. How can we secure individuals' rights, data and privacy online, using traditional national security approaches when the challenges we face are inherently both local citizen-based, and international? One way forward could be to develop policies that consider the individual as the epicenter of the security challenge instead of only traditional territorial sovereignty. Human beings need to be the core focus of the IT and security agenda going forward.

The ICT4Peace foundation in cooperation with the Zurich Hub for Ethics and Technology (ZHET) held a series of informal workshops[1] in 2018 with leading thinkers on the impact of AI. As the events of the past months have signalled, it is clear we are at a turning point about how we want to manage and shape the future of the "Data Age".

Introduction

Since the creation of the World Wide Web in 1989 humanity has been charging full speed ahead on all things technological. Much of this has been positive for humanity: improving global interaction, helping to lift millions out of poverty, simplifying processes, facilitating medical advances, promoting peace, improving peace-keeping operations and humanitarian work, and allowing for contacts and connections in real time everywhere. Democratic governments and open-minded individuals were euphoric about the role of social media in the Arab Spring,

---

[1] Including with Professor Stuart Russell**,** *Professor* of Computer Science and Smith-Zadeh *Professor* in Engineering, University of California, Berkeley, USA, and Professor Ron Arkin**,** Regents' Professor, Director of Mobile Robot Laboratory School of Interactive Computing, College of Computing, Georgia Tech, USA. The ICT4Peace and Zurich Hub for Ethics and Technolog (ZHET) also participated at RightsCon Toronto 2018 on ***Artificial Intelligence: Lethal Autonomous Weapons Systems and Peace Time Threats*** with Todd Davies, Associate Director and Lecturer, Symbolic Systems Program, Stanford University, USA, Ron Deibert, Director, Citizen Lab, Munk School of Global Affairs, University of Toronto, Canada , Kyle Dent, Author and Researcher, Palo Alto Research Center, PARC, USA , Maarten Van Horenbeeck, Board Member and former Chairman, Forum of Incident Response and Security Teams (FIRST), USA and David Kirkpatrick, Founder, Techonomy Media, and author, The Facebook Effect, USA.

seeing it as a progressive force, uniting the oppressed and underprivileged, a tool for mobilization and the ultimate voice and power of the people. The principles of democracy and equality, “Liberté, Fraternité et Égalité”, “life, liberty and the pursuit of happiness”, were thought to be entrenched, virtually unchallenged and impossible to turn back in particular due to the power of technology.

The world today looks much more layered and complex. The euphoria is gone and has been replaced by angst. The tools that were meant to bring people closer together, to lessen even further the differences in colour, culture, religion, gender and sexuality, have been turned against the very principles humanity has been striving towards for generations, and have in part created even greater divisions. Against the volatile backdrop of rising inequality across the world, disparate actors are manipulating social media and the masses for their own personal political and material gain. Disinformation, misinformation, propaganda, is spread with lightning speed to targeted audiences reinforcing beliefs based on lies, perpetuating a vicious cycle where the truth is not only hard to find but for many uninteresting, irrelevant and in their world untrue. Advances in Artificial Intelligence are in part augmenting these trends. The full impact of these technological tools are yet to be really understood. What is sure is that the consequences of our actions today, the enormity of which we don’t or can’t fully comprehend, will be felt for generations to come.

We are witnessing a rise in demagoguery, a rise in racism and a turning away from joint global and regional institutions. Institutions that were created out of the destruction and devastation of World War, the lessons of which seem to have been forgotten. Institutions that may not be perfect but are a crucial forum for interaction, exchange, global discussion, and a means through which we can actually tackle global problems together such as climate change, poverty, conflict, transnational crime and trade. Protectionism, nationalism, racism and hatred have had their chance in history and always failed.

We need to start using technological tools and AI to fight against current negative trends, we need to have ethical technology that is not easy to manipulate and abuse. We need to re-build trust in the media and in our institutions, and use technology as a tool to promote peace and not hatred. In the dearth of leadership today, we need to think how great leaders like Martin Luther King or Nelson Mandela would have used technological tools to advance peace and the fundamental principles of human rights. How would the principles of King’s 1963 speech “I have a dream” translate into the digital age? We need to promote digital human security, starting with the individual, both in conflict and peace-time situations, and through all levels of society and government.

In 1994 the UNDP Human Development Report presented the concept of security as linked to humans rather than geographical entities and to development instead of weapons. http://hdr.undp.org/en/content/human-development-report-1994. The concept of human security included freedom from fear and freedom from want. Human security encompasses

food, a safe place to live, healthcare, economic well-being and education. It is now time to extend this to encompass technological issues that threaten human security; to consider the full impact of technology on the individual from fake news to the latest developments in AI.

The impact of AI, LAWS and Peace-time Threats

AI is a fundamental game-changer in the context of Digital Human Security. Only time will really tell how much is hype but we are still at a point where we can influence, shape and govern the outcome and the type of world we would like to live in. To do this we need the active involvement of all actors including individual citizens, civil society, business, government, the international community and academia. Like almost all advances, AI brings with it both positive beneficial aspects for society and negative, potentially dangerous developments.

We need to look closely at the role of AI including AI and peace-time threats, AI and cybersecurity, AI and ethics, AI and Lethal Autonomous Weapons Systems (LAWS) and the use of AI in peace negotiations and the non-violent transformation of conflict. We also need to ensure that individual citizens are being educated about AI and its implications. Only then can individuals take informed decisions about information they are presented with and about their own digital security. Finland is leading the way in educating its citizens through the development and launch of a free online course in May 2018. *"The Elements of AI course seeks to demystify AI by making it more accessible - it is targeted to anyone who is interested in learning more about AI with no prior mathematical or programming skills required. The initiative of the Finnish government aims to attract 1% of the population to take up the challenge and learn more about basics in AI topics such as machine learning and neural networks by the end of this year."* https://www.nesta.org.uk/blog/ai-all-how-finland-and-other-countries-are-delivering-free-accessible-digital-skills-training/
https://www.elementsofai.com

AI already permeates many aspects of our daily lives with algorithms continuously figuring out and storing our preferences, and allowing an unprecedented level of surveillance and personal data accumulation, in many cases negatively impacting our digital security. Microsoft has just called for the regulation of facial recognition technology, https://www.technologyreview.com/the-download/611651/microsoft-wants-the-government-to-regulate-face-recognition-software/ . Algorithms are being used to hire and fire people, to determine sentencing in judicial processes, to forecast global financial activities, for targeted advertising, loan approval, translation, self-driving cars and also have an enormous impact in the health industry. Robots are taking on an increasingly important role in many businesses, and may spell mass unemployment and job displacement. ICT4Peace and the Zurich Hub for Ethics and Technology have highlighted the "Peace Time Threats" that AI can pose for Society in the following paper by Regina Surber: https://ethicsandtechnology.org/artificial-intelligence-autonomous-technology-lethal-autonomous-weapons-systems-laws-peace-time-threats/.

In line with the moves in Silicon Valley to integrate ethics, Facebook's initiatives to create AI models and tools that deal with hate speech, we need to look at how AI can be used for complex decision support modelling in peace negotiations, and the ways in which information flows can support peace, or the non-violent transformation of conflict. This work also embraces Data for Good initiatives including work the ITU has done, https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx; work the private sector is doing in this field, e.g. Microsoft https://www.microsoft.com/en-us/ai/ai-for-good and IBM https://ai.xprize.org/AI-For-Good, and processes within the UN, https://news.un.org/en/story/2017/06/558962-un-artificial-intelligence-summit-aims-tackle-poverty-humanitys-grand

Potentially most dangerous for Digital Human Security are the advancements of AI in the defence industry with the development of LAWS. The danger of LAWS has been highlighted by the Campaign to Stop Killer Robots, https://www.stopkillerrobots.org, and the 2015 Open Letter from AI researchers including Stephen Hawking and Elon Musk. At the recent Paris Peace Forum in November 2018, the Secretary-General of the UN, António Guterres said, *"Imagine the consequences of an autonomous system that could, by itself, target and attack human beings. I call upon States to ban these weapons, which are politically unacceptable and morally repugnant."* https://www.stopkillerrobots.org/2018/11/unban/ .

LAWS are dangerous for many reasons including our inability to defend against them; the fact that a small number of people could deploy large numbers of LAWS; and that they are relatively cheap and based on software technology that is vulnerable to hacking and accidental use. They are also not accountable before the law and therefore universally accepted "guidelines" of some kind are essential. The Future of Life Institute has recently released a pledge in July 2018, with 2400 signatures of scientists who are refusing to help develop or manufacture LAWS. https://www.theguardian.com/science/2018/jul/18/thousands-of-scientists-pledge-not-to-help-build-killer-ai-robots

At an international state level, the problematic of LAWS is being addressed in the context of the Group of Governmental Experts, (GGE) of the High Contracting Parties to the Convention on Certain Conventional Weapons at the UN in Geneva. The Chairman's Summary from the last meeting held in August, made the following points:

Emerging commonalities, conclusions and recommendations Possible Guiding Principles

26. It was affirmed that international law, in particular the United Nations Charter and international humanitarian law (IHL) as well as relevant ethical perspectives, should guide the continued work of the Group. Noting the potential challenges posed by emerging technologies in the area of lethal autonomous weapons systems to IHL,[1] the following were affirmed, without prejudice to the result of future discussions:

(a) International humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems.

(b) Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system.

(c) Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control.

(d) In accordance with States' obligations under international law, in the study, development, acquisition, or adoption of a new weapon, means or method of warfare, determination must be made whether its employment would, in some or all circumstances, be prohibited by international law.

(e) When developing or acquiring new weapons systems based on emerging technologies in the area of lethal autonomous weapons systems, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered.

(f) Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems.

(g) Consideration should be given to the use of emerging technologies in the area of lethal autonomous weapons systems in upholding compliance with IHL and other applicable international legal obligations.

(h) In crafting potential policy measures, emerging technologies in the area of lethal autonomous weapons systems should not be anthropomorphized.

(i) Discussions and any potential policy measures taken within the context of the CCW should not hamper progress in or access to peaceful uses of intelligent autonomous technologies.

(j) The CCW offers an appropriate framework for dealing with the issue of emerging technologies in the area of lethal autonomous weapons systems within the context of the objectives and purposes of the Convention, which seeks to strike a balance between military necessity and humanitarian considerations.

https://www.unog.ch/80256EDD006B8954/(httpAssets)/20092911F6495FA7C125830E003F9A5B/$file/CCW_GGE.1_2018_3_final.pdf

Progress is quite slow due to diverging viewpoints, and in particular due to resistance among key players to call for an outright ban or a mandatory minimum level of human involvement in life and death decision-making by LAWS. As a society, we need to remain vigilant and ensure that the 1949 Geneva Conventions and the principles of International Humanitarian Law (https://www.icrc.org/en/doc/assets/files/other/what_is_ihl.pdf) are not being contravened through the use of LAWS. Importantly we also need to consider the civilian impact of LAWS, both in terms of populations at risk in conflict situations but also the risk of individuals or entities acquiring these devices as weapons in peace-time situations for their own personal use and criminal activities.

AI and Cybersecurity

An additional area that needs to be addressed when considering Digital Human Security is the nexus of AI and cybersecurity, a potentially under-examined intersection of two powerful forces. We need to consider the human element, and focus on ensuring the security and safety of individuals and their data across networks. As Ron Deibert argues in his recent paper, Towards a Human Centric Approach to Cybersecurity, the individual needs to take center stage in policy development concerning cybersecurity. This approach "prioritizes human rights and civil society as the ultimate objects of security, with nation states in supporting roles". https://www.cambridge.org/core/journals/ethics-and-international-affairs/article/toward-a-humancentric-approach-to-cybersecurity/4E8819984202A24186BB0F52E51BC1E4

AI is at the core of many security technologies. Identifying what is causing an error or where an attack might be coming from will become even more difficult in the future, hence the rush to use artificial intelligence to identify attacks. One risk is that this could create a sense of false security. "Many products being rolled out involve *""supervised learning," which requires firms to choose and label data sets that algorithms are trained on—for instance, by tagging code that's malware and code that is clean….. The bad guys don't even need to tamper with the data; instead, they could work out the features of code that a model is using to flag malware and then remove these from their own malicious code so the algorithm doesn't catch it."* *https://www.technologyreview.com/s/611860/ai-for-cybersecurity-is-a-hot-new-thing-and-a-dangerous-gamble/*

We also need to push for explainable AI and the auditing of AI and move away from the black box problem that comes with deep learning and opaque technology. As we build more and more complex algorithms we may get to a place where turning them off and doing things manually might not work anymore. This is not in anyone's interest, particularly not when trying to prevent, identify, manage and defend against cyberattacks.

AI, Ethics, Trust and Transparency

On the overarching ethical questions relating to AI and its use, we need to assess fundamental questions about the level of human involvement in all aspects of AI from LAWS to facial recognition technology to judicial sentencing. We need to take concrete steps to ensure that engineers and ethicists interact and learn about each other's fields, perhaps through mandatory course requirements at the university level and regular workshops. This will help to ensure an educated discussion and awareness of the key ethical concerns, preventing in part unwanted results when developing algorithms and new AI. Algorithms that are currently, often accidentally, programmed with bias need to be designed with values supporting human rights and protecting individuals. We also need to safeguard the quality of the information and data being used in developing AI and the underlying algorithms.

Trust is perhaps the most critical missing element in restoring the "good" in technological advances. Trust in the information I am receiving either as an individual or an organization. Trust in my online environment. Trust in the processes and institutions I use on a regular basis. Trust as a victim of conflict or as a member of a vulnerable population in the government, humanitarian organization or NGO trying to assist me. Trust in the data an organization is using to make critical decisions in peace-time and conflict. How can the international community work toward this goal?

One concrete example of restoring trust that could be replicated elsewhere is the **bot legislation** in California. "A bot is an automated software program that does *something*. Beyond this rudimentary description, bots vary tremendously. They moderate chat room discussions, scrape the web to collect information, and provide customer service on websites. They also pose as real people on social media, where they can cause serious mischief. It is this last capability that has made bots a part of our common vernacular." https://slate.com/technology/2018/08/to-regulate-bots-we-have-to-define-them.html
The bot legislation in California, signed by Governor Jerry Brown is "a new law that bans automated accounts, more commonly known as bots, from pretending to be real people in pursuit of selling products or influencing elections. Automated accounts can still interact with Californians, according to the law, but they will need to disclose that they are bots". *https://www.nbcnews.com/tech/tech-news/can-t-spot-bot-california-automated-accounts-have-reveal-themselves-n915556*

The idea of making it mandatory for a bot to identify itself was also raised by Professor Stuart Russell at an ICT4Peace Workshop on AI and Peace Time Threats in Zurich in June 2018 https://ethicsandtechnology.org/ict4peace-colloquium-with-professor-stuart-russell-at-the-zurich-hub-for-ethics-and-technology-zhet/ . This simple step would make real progress in ensuring greater Digital Human Security. Of course, there is much discussion about the definition of bots and the distinction between malicious bots and bots that perform other functions including the categorization of information. However, knowledge is power and if the user is simply informed that they are being contacted by a Bot then the user has the choice to continue or not. At the moment this option does not exist. A recent book co-authored by Tony Veale, Associate Professor at the University College Dublin, and Mike Cook, Senior Research Fellow at the University of Falmouth, UK, entitled "Twitterbots: Making Machines that Make Meaning," explores the bot theme in depth.

AI and Transparency

For trust to flourish, transparency is also needed. Transparency in sources of information, transparency in the medium and means through which individuals receive information and transparency in decision-making. AI and transparency is a complex problem as even the most renowned experts and roboticists working on AI don't always understand how deep learning works. AI is riddled with "black box" situations, whereby the researchers working on various

projects are not exactly sure how the "machine" reaches certain conclusions. For example a car developed by Nividia learned how to drive itself: "The car didn't follow a single instruction provided by an engineer or programmer. Instead, it relied entirely on an algorithm that had taught itself to drive by watching a human do it." The resulting problem is that no-one knows exactly how the car reaches the decisions it takes. As society increasingly uses deep learning in a wide range of fields, we need to make sure that we can ultimately control the technology we are creating. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

There are some new technological tools that exist that can improve transparency. For example. Canada is using blockchain technology to improve transparency in government funding, which provides a small example of how technology itself, in particular the blockchain could be part of the solution to improve Digital Human Security. Its National Research Council is using the Ethereum blockchain to proactively publish grants and contribution data in real time, complementing ongoing quarterly disclosures available through the Open Government website." https://nrc-cnrc.explorecatena.com/en/

> On the simplest level, blockchains are public ledgers that record transactions shared among many users. Once data is entered on a blockchain it is secure and unalterable, providing a permanent public record. This technology offers unprecedented levels of transparency and trust allowing public records to be searched, verified and audited at a level the world hasn't seen before.

AI and Recourse

Also of great importance is the concept of recourse. Recourse for individuals and organizations, or even countries who have been maligned by false information or AI augmented/manipulated information. Recourse for individuals who have been fired by an AI empowered system. Recourse against, for example, the rigid AI empowered structures being implemented in China under the national social credit system. Recourse against a machine is inherently difficult and needs to be considered by governments as a key public policy objective in reaching a new level of Digital Human Security.

Understanding the technology is also of critical significance as we move forward. At all levels of society there is a desperate need for programs to educate and increase the general understanding of the technology and the ethical implications of our decisions. Individuals and governments need a toolkit for the road ahead.

Conclusion

As we can see from the above examples, the governance of AI and its impact on Digital Human Security is of critical importance. It seems a daunting, complex and unwieldy task to try to figure out a path forward. There are many institutions working on, or having published guidelines for the ethical use of AI. In 2017, the Future of Life Institute issued guiding principles for AI (please see addendum 1 https://futureoflife.org/ai-principles/?cn-reloaded=1) These guidelines are incredibly useful and should be considered in detail when confronted with AI research and new technological advances.

What is perhaps also needed, or would certainly be useful, is a simple litmus test or prism through which to assess new technological advancements in both peace-time and conflict situations**. ICT4Peace and ZHET are proposing the following five points** outlined by the acronym T.R.U.S.T. that would serve this purpose and act as a preliminary checklist:

T-Trust Can I trust this information?)

R-Recourse (Do I have recourse ?)

U-Understanding and Education (Do I understand the technology and implications?)

S-Security/Safety (Am I safe? Is this a secure system? Secure interaction?)

T-Transparency (Can I see the methodology and processes through which this information is being provided or decision being taken?)

As a global society we have to decide what kind of world we want to live in and set the boundaries accordingly. Awareness, education and learning about AI and what it means for individual citizens, society and our institutions is critical. It is hard for governments to respond quickly and effectively to fast-moving technological change. The ICT4Peace and ZHET goal is to develop and highlight recommendations for policy-makers and business leaders about the governance of AI for good, using AI for peaceful purposes and preventing it's use for purposes that negatively impact human security.

The importance of global collaboration in promoting human security has never been more crucial. We as a global community are faced with serious challenges both in terms of the issues themselves, such as war, migration, climate change or the potentially destructive force of AI, but also in terms of the lack of real forward-thinking global leadership to address these issues. The interest in international cooperation from some of the key architects of the post WWII world order is waning and in fact could be described as actively destructive. Those countries, organizations, civil society and individuals that are committed to continued global progress, cross-border collaboration, supporting an inclusive and environmentally and technologically

viable future need to stand together and take action to ensure digital human security. We cannot afford to backtrack on key issues, and be led astray by negative, destructive and protectionist behaviour. The challenges we face are global and need to be addressed collaboratively. We as a foundation are committed to creating a better and more peaceful world through the constructive use of Information and Communication Technology.

ICT4Peace is a policy and action-oriented international foundation. The purpose and goal of the foundation is to save lives and protect human dignity through the improved use of Information and Communication Technology for good. Since 2003, ICT4Peace has explored and championed the use of ICTs and new media for peaceful purposes, including for peacebuilding, crisis management and humanitarian operations. Since 2007, ICT4Peace has promoted cybersecurity and a peaceful cyberspace through international negotiations with governments, international organizations, companies and non-state actors.

Barbara Weekes, Senior Advisor ICT4Peace Foundation and Board Member of the Zurich Hub for Ethics and Technology (ZHET) and Daniel Stauffacher, President, ICT4Peace Foundation

Heidelberg, 18 Dezember 2018.

Addendum 1

Guidelines for the ethical use of AI, Future of Life https://futureoflife.org/ai-principles/?cn-reloaded=1

Research Issues

1) **Research Goal:** The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.

2) **Research Funding:** Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:

- How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
- How can we grow our prosperity through automation while maintaining people's resources and purpose?
- How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
- What set of values should AI be aligned with, and what legal and ethical status should it have?

3) **Science-Policy Link:** There should be constructive and healthy exchange between AI researchers and policy-makers.

4) **Research Culture:** A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.

5) **Race Avoidance:** Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and Values

6) **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

7) **Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.

8) **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

9) **Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

10) **Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

11) **Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

12) **Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

13) **Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

14) **Shared Benefit:** AI technologies should benefit and empower as many people as possible.

15) **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

16) **Human Control:** Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

17) **Non-subversion:** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

18) **AI Arms Race:** An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

19) **Capability Caution:** There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.

20) **Importance:** Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.

21) **Risks:** Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

22) **Recursive Self-Improvement:** AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.

23) **Common Good:** Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.