ICT for peace foundation

POLICY
BRIEF

ETHICAL AND POLITICAL PERSPECTIVES ON EMERGING DIGITAL TECHNOLOGIES

Regina Surber (Author)

Daniel Stauffacher (Editor)

GENEVA 2022

ICT4Peace Foundation

ICT4Peace Publishing, Geneva. March 2022
Copies available from www.ict4peace.org

ETHICAL AND POLITICAL PERSPECTIVES ON EMERGING DIGITAL TECHNOLOGIES

Regina Surber (Author)

Daniel Stauffacher (Editor)

ICT for peace foundation

CONTENTS

FOREWORD

ICT4Peace is pleased to publish[1] this compilation of policy briefs, op-eds and recorded talks by Regina Surber, a global thought leader and scholar.

Regina Surber's deep concern for human rights and the ethical dimension of the rapid technological developments in Artificial Intelligence, brought her to ICT4Peace's attention in 2016. In particular, when she encouraged and helped the Foundation to further pursue its initial work on AI, Lethal Autonomous Weapons Systems (LAWS), Emerging and Converging Technologies and Peace Time Threats.

Her pioneering publications and lectures that followed, helped to inform the international community on the risks of Autonomous Technology (AT) and Lethal Autonomous Weapons Systems (LAWS), and she contributed – through ICT4Peace - her message to the legal and policy debates within the international arms control framework of the United Nations Convention on Certain Conventional Weapons (UN CCW).

Regina, along with other colleagues at ICT4Peace, demonstrated at an early stage that LAWS are not the only manifestation of the security risks of AT, but also other emerging technologies, such as quantum computing, additive manufacturing, or biotechnology, such that these may – in her own words – converge into a new weapons landscape, and demonstrated, that these emerging technologies not only have effects during armed conflicts but also during peace time.

Through her research and writings, she demonstrated very well that new weapon systems do not always fit within our traditional concept of state sovereignty and do not only impact State security, but also affect human security as well. This is because these weapons systems impact numerous aspects of individuals' lives including, but not limited to, our data security, privacy, autonomy, or the (truth or falsehood of) available information. After the outbreak of Corona she was one of the few scholars who warned early on, that the many government's reliance on emerging technologies to contain the pandemic, may severely infringe on the right to privacy, and possibly mark the transition into a surveillance society.

[1] In cooperation with the Zurich Hub for Ethics and Technology (ZHET) (www.ethicsandtechnology.org), which Regina Surber helped co-found with ICT4Peace in 2016

It is for these reasons that she is calling for a rethinking and a reshaping of traditional architectures both on the level of international arms control and disarmament, as well as at the level of national and international governance. To support these processes she urges the integration of education and training on ethics and technology into educational systems around the globe.

Regina's important contribution to the better understanding of the potential threats of emerging and converging technologies and her human rights advocacy work is profoundly important for the international (human) security landscape, as this compilation demonstrates.

Daniel Stauffacher
Founder and President
ICT4Peace Foundation

POLICY BRIEFS

DECENTRALIZED DIGITAL NETWORKS

A Potential for a New Way of Human Cooperation?

Zurich Hub for Ethics and Technology and ICT4Peace 2021[1]

Global society is currently witnessing the development of '*decentralized digital networks*' and applications that run on them. The text briefly describes the main underlying idea of this pragmatic and revolutionary new technology as well as some of its societal potentials.

Key points:

- A decentralized digital network distributes information-processing workloads across multiple devices instead of relying on a single central server (centralized digital network).

- In contrast to centralized digital networks, decentralized digital networks enable a great degree of user privacy, are safer against cyberattacks, and data is ideally owned by the network user.

- In a centralized digital network, network users need to trust the central server 'agency.' In a decentralized digital network, the users need to trust the autonomous algorithm that enables user coordination within it.

- Applying decentralized technological processes can reduce or even eliminate the role of intermediaries across industries.

- Any digital application can be built on a decentralized digital network. Cryptocurrencies and the Web 3.0. are such applications. They have the

[1] https://ethicsandtechnology.org/wp-content/uploads/2021/08/2021_RSurber_Decentralized-networks_for-ICT4Peace-and-ZHET-1.pdf

potential to disrupt the current global financial system as well as the way humans exchange values on the web.

- Every aspect of human coexistence requires coordination of human activities. Until now, this coordination – especially economics and nation-states – has happened in a centralized manner. Decentralized digital networks offer a potential for creating a hierarchy-free digital world of secure human exchange of information, money, values, goods, etc.

1. Centralized digital networks

In order to understand the innovative potential of decentralized digital networks, it helps to first contrast them to centralized digital networks that currently undergird most of global society's interaction online.

Currently, most digital networks are *centralized*.[2] Centralized networks are arranged around one central server. In simple words, the central server 'verifies' all the data processing happening amongst the users in the network. The central server thereby aims to solve an important problem that arises in all networks composed of a great number of users: in a network of many users that do *not* know each other's identity – and in the digital space (but not only) knowing users' identity is per se difficult – the individual users *cannot trust* that the information they receive from each other is not deceptive, nor can they trust that the information they themselves send out is not intercepted before it reaches the intended recipient. Without mutual trust they cannot reach a consensus about a certain issue. E.g., how can I know the email *really* came from you? Or, as digital objects are easy to duplicate, how can I be sure that the money I was sent was not simultaneously sent to someone else? One solution is the establishment of a central agency, the central 'server'. It figures as a monitoring device for the information flow and as an authority for publishing correct information in the network. In the email example this is the email-provider. In the money transaction example, this is a payment processor, an automated clearing house, or a bank (ultimately central banks). As the previously un-centralized networks face a trust problem when wishing to reach a consensus, they became centralized, and thereby hierarchical.

[2] The majority of today's web services – incl. YouTube or our online banking accounts – are based on centralized networks.

1.1. The trust problem of centralized digital networks

In order for a centralized network to function, the members must *trust* that the central authority, with its capacity to monitor the information flow in the network and its power to decide which pieces of information are true and which are not, does not *itself* deceive the members of the network. To give an example: in the traditional financial system, banks are *trusted* to show clients their balances and transaction histories in an honest manner. If a bank did attempt to lie, or defraud, its customers, a central authority higher up in hierarchy – the central bank or government – is then again trusted to rectify the bank's breach of trust. It follows that, whereas the consensus problem of decentralized networks is solved by establishing a central authority, the *source* of the problem – the problem of *trust* – is *not* solved, but merely relocated: given that centralized systems also require trust in a central authority, they are again vulnerable to corruption – not by its individual members, but by the central authority those members had established.

1.2. Disadvantages of centralized digital networks

Centralized digital networks face a number of drawbacks: First (1), the central server constitutes the network's single point of failure; if it crashes, the entire network is likely to shut down. Second (2), as there is only a single point of failure, cyber attackers must only compromise one target in order to disrupt the network. Third (3), given its centralization, data ownership and computational resources are not distributed evenly among the network. Hence, data, knowledge, and, thus, power, is located at the central server agency which needs to be trusted not to abuse it. Fourth (4), as there exists only one central depository of user data, centralized networks always involve an inherent privacy risk. If the main server is attacked, taken offline, or itself corrupted, user data may be lost.

2. Decentralized digital networks

Decentralized digital networks are a conglomerate of connected, but separated digital entities or users that communicate with each other *without* a central server. A great example of decentralized digital systems is a 'blockchain'.

In a blockchain, *every* network user must ‘approve’ of anything that happens in the network. Whenever information is exchanged between two or more users in the network, this is recorded and stored on *each individual* computer device – i.e. with every user – in the network. The data record of all transaction information gathered during a certain time period is called a ‘block.’ With transactions unfolding over time, those ‘blocks’ are added to the ‘chain of data’ in the network – hence the name ‘blockchain.’ It follows that the ‘truth’ that all network users must agree upon *is* the blockchain – that is, the decentralized network itself.

It is not so that users manually approve a transaction on a blockchain. This is handled by an autonomous algorithm that runs the decentralized digital network. The upshot is that the protection against information manipulation and misuse is *enshrined in* the technical structure of the decentralized network itself. There is no central server needed anymore. The blockchain is a decentralized database the technology of which ensures that the above-explained trust problem between network users does not arise. This is why some argue that blockchain is the first ever digital *solution* to the trust problem. However, also with decentralized digital networks, trust is still needed – not in other network users, nor in a central agency – but *in the system itself*.

2.1. Advantages of decentralized networks

It helps to visualize the blockchain network as a ‘book’ that stores every information exchange that has *ever* taken place on the network, and blocks as ‘pages’ that continuously update the state of exchanges in the network. As *each* computer device in the network maintains a copy of each ‘page’, this makes it almost impossible for a single computer (user) to change a page in retrospect. Hence, decentralized digital networks have the advantage (1) that whatever is agreed upon within them is almost impossible to manipulate. In addition (2), decentralized networks enable a greater degree of user privacy, since information saved on the network is disseminated across multiple points instead of passing through a single point. (3) This also makes data flows more difficult to track across a network, and eliminates the risks of having a single target that malicious actors can go after. In decentralized networks, data ownership and computational resources are ideally shared equally across the network. In addition (4), network users must not trust in a single central agency to both publish data correctly and not misuse it. Furthermore (5), centralized networks require a trusted third party, a central authority, or a ‘middleman’, to secure information exchange and transactions. With the possibility of decentralized systems, those previously necessary intermediaries are no longer needed. This saves time and money, and has the potential to 'give back power' to the network users:

2.1.2. Societal and economic potentials of decentralized digital networks

Decentralized networks can undergird many, if not *any*, digital application. Today, there already exist, e.g., blockchain-based contracts, software ensuring the secure sharing of medical data, cross-border payment software, personal identity security software, anti-money laundering tracking systems, voting mechanisms, or supply chain and logistics monitoring. Currently, though, the two most groundbreaking potentials are cryptocurrencies and the Web 3.0:

2.1.2.1. Cryptocurrencies

Today, decentralized networks' potential is arguably being demonstrated strongest in the financial sector. The reason is that money is a prime example of the above-described trust problem. The root problem with all conventional currency is the trust that is required to make it work: governments and central banks must be trusted not to debase currencies. However, over the course of history, this trust has been breached many times.

Bitcoin was the very first currency that runs on a blockchain and that, hence, does *not* require trust in central monetary agencies. It is a *crypto*currency because it is secured by advanced cryptography. In very simple words, the algorithmic mechanism is the following: all members of the blockchain agree on every financial transaction occurring amongst them. Thereby, they verify who owns how many bitcoins at what time and establish a functioning money without a centralized authority.[3]

Cryptocurrencies are traded directly between two or more participants of a decentralized network. This is called 'peer-to-peer trading'. Peer-to-peer trading removes the centralized middleman, allowing the users of the platform to pay minimal or zero fees to use the service.

[3] Some large companies, e.g. Microsoft and AT&T, accept Bitcoin as a legitimate source of funds. Most countries have not clearly determined the legality of bitcoin, preferring instead to take a wait-and-see approach. Some countries have indirectly assented to the legal use of bitcoin by enacting some regulatory oversight. However, as of June 2021, El Salvador is the only country that recognizes bitcoin as legal tender.

In contrast, most traditional financial institutions charge fees and impose limits on the size, type, and number of transactions a client can execute. Additionally, some transactions in the classical centralized financial systems can take anywhere from 30-90 days to settle depending on the transaction type. Bitcoin transactions, in turn, can achieve final settlement in as little as one hour.

Central banks figure as gatekeepers of the current centralized money transaction process. They make money through interests and through the management of money and transactions. Hence, widely used cryptocurrencies and peer-2peer trading systems would make central banks, and banks in general, potentially obsolete. With decentralized financial systems, no bank nor corporation would make money out of human financial exchange – only humans themselves would profit.

This may be the reason why almost every central bank worldwide is currently trying to develop its own digital version of its fiat currency.[4] Those currencies are called *central bank digital currencies (CBDCs).* Regulated by a country's monetary authority, CBDCs are designed to replace traditional fiat and increase ease of use for those that deploy them. However, unlike blockchains, CBDCs are not decentralized. Hence, central banks must be trusted not to compromise money holders' privacy. Critics regard repercussions concerning financial privacy as well as censorship as a great risk inherent to CBDCs.

In short: the governmental resources invested in the development of CBDCs can be understood to reflect the disruptive power of applications running on decentralized systems.

2.1.2.2. Web 3.0

The term Web 3.0 refers to a vision of the 'third generation' of computing. It anticipates that technologies like blockchain will decentralize the internet, thereby disintermediating companies like Facebook, Amazon, Google, LinkedIn, and Apple to enable the online exchange of value, and allow users to own their data. Web 3.0 is designed to benefit all participants using a peer-to-peer model for websites, applications, and the internet as a whole. It aims to be an open, public, censorship-resistant, borderless, free internet. Analogous to a decentralized financial system: no corporation would make money out of human information exchange on the web – only the humans themselves would profit.

[4] As of August 2021, app. 83 countries are researching and developing CBDCs.

Every aspect of human coexistence requires coordination of human activities. Up to now, those activities – most notably economics and nation-states – have been coordinated in a centralized manner. The more global digitalization proceeds, more ways of human interaction go digital. Decentralized digital networks offer the potential for creating a hierarchy-free digital world of secure human exchange of information, money, values, goods, etc.

3. Key Questions

- Decentralized digital networks require trust in the algorithms that run the network. This raises a series of new questions, most importantly: can a technological artifact be a trustee? If so, what technological conditions must be in place?
- The existing global centralized human coordination mechanisms are challenged by a promising new technology favoring decentralization, and a peer-to-peer, safe and open digital exchange of values. Will those two mechanisms of human cooperation continue to coexist? Will there be a transition from centralized to decentralized human digital interaction? How would such a transition look like? Would it be socially disruptive, or could one pave the way for a smooth passage? How?
- Given the potentially tremendous influence of decentralized digital networks on every aspect of future human (co-) existence and cooperation, democratic legitimization of the algorithms that run those networks seems key. However, as those algorithms are very complex, their development and design could hardly be agreed to democratically. Hence, their application requires trust in the technologically knowledgeable. How to ensure that it is not abused?

CORONA PAN(DEM)IC: GATEWAY TO GLOBAL SURVEILLANCE

Springer Ethics and Information Technology 2021[1]

Abstract: The essay reviews the digital emergency measures many governments have adopted in an attempt to curb Covid-19. It argues that those 'virologically legitimized' measures may infringe the human right to privacy and mark the transition into a world of global surveillance. At this possible turning point in human history, panic and latent fear seem to fog much needed farsightedness. Leaving the current state of emotional paralysis and restarting to critically assess the digital pandemic management can serve as an emergency break against drifting into a new era of digital monitoring.

Keywords: Corona; Covid-19; Pandemic; Human Rights; Digital Technologies; Surveillance; Ethics

It is said that the 'corona crisis' may be the biggest crisis of the current generation. As of 28 September 2020, 32.7 million persons are said to have been tested positive on Sars-CoV-2 in more than 200 countries and territories, and 991.000 people are said to have died from Covid-19 (WHO 2020a). On 11 March 2020, the World Health Organization's Director General declared Covid-19 as a pandemic (WHO 2020b). By the end of January and early February 2020, a wave of panic of the previously unknown physical Covid-19 illness has spread across the planet.[2]

Governmental restrictions and human rights

In an attempt to contain the spread of the corona pandemic, and in order for national health care systems not to be overwhelmed by the potentially enormous influx of people suffering from the acute respiratory syndrome that Sars-CoV-2 may trigger,

[1] https://link.springer.com/article/10.1007/s10676-020-09569-5 and text published by Research Outreach: https://ict4peace.org/wp-content/uploads/2020/04/2020_RSurber_Corona-pandemic_final-1.pdf. Original text was published by ICT4Peace in April 2020: https://ict4peace.org/wp-content/uploads/2020/04/2020_RSurber_Corona-pandemic_final-1.pdf

[2] Sars-CoV-2 is a mutation of a corona virus known to cause severe disease in the human body; see e.g. Corman et al. (2018), Anderson et al. (2020). Note, however, that there exists disunity among scientists with regards the severity of Sars-CoV-2.

many governments have adopted emergency measures to secure public health and order.

Those emergency measures are, arguably, drastic. As of March 2020, almost the entire globe 'locked down':

Most governments have temporarily closed educational institutions, impacting 60% of the world's student population. Several other countries have implemented localized closures that may impact millions of additional learners (UNESCO 2020).[3]

As a result of the pandemic, around 70 countries across the world had imposed or still are imposing entry bans, quarantines and other restrictions for citizens or travelers to most affected areas (Salcedo and Cherelus 2020). As of 28 September 2020, around 70 countries and territories still impose global restrictions applying to all foreign countries, or prevented their citizens from travelling (IATA 2020). Many governments had also implemented curfews or urged people to stay at and work from home. In places where people were still allowed to leave their houses, gatherings of more than a handful of people were banned.[4] During the lockdowns, in many countries, doctors' offices and pharmacies remained open, but restaurants, bars and non-essential shops in the majority of places around the globe were ordered to close their doors. This threatened the existence of small companies, with some businesses already declaring insolvency as early as March 2020 (Allen 2020), and governments adopting economic support measures of unprecedented amounts (European Commission 2020). Further, especially in low-income countries, health access was restricted to almost Covid-19-only cases, disrupting the prevention and treatment of other noncommunicable diseases (WHO 2020c).

Whereas lockdowns are gradually eased and terminated in phases[5], fears of a second virus wave are currently spreading again due to surges in the number of confirmed cases in various regions. This pushes some countries to consider a retake

[3] By the end of April, 190 countries and territories had closed educational institutions. As of 1 August 2020, 106 countries still observe a nation-wide school closure (UNESCO 2020).

[4] An example is Switzerland, see e.g. Schweizerische Eidgenossenschaft BAG (2020).

[5] E.g. New Zealand, Spain, Germany, South Korea, India, Iran, Hungary, Singapore, Dubai, Panama, Peru, Thailand (Kaplan et al. 2020).

on restrictions[6] or even mandating second lockdowns.[7] The restrictions affect our human rights. Curfews and the ban on gatherings may infringe our freedoms of movement[8] and assembly.[9] The closing of educational institutions worldwide severely rephrases access to education, a right granted by the Universal Declaration of Human Rights (UDHR).[10] What is more, the requirement to shift to online-learning exposes education's digital divide: in poorer countries, children may not have the resources required to be digitally home-schooled (Thong 2020). Further, many health institutions and hospitals have been forced to triage patients in case of sudden overloads, which may impede the right to access to medical care.[11] In addition, the shutting down of public life has put jobs and livelihoods into severe jeopardy – possibly affecting our right to work.[12]

What is more, the listed emergency measures have forced a great majority of people to physically isolate and distance from loved ones. Millions of people also face economic turmoil, because they have lost, or are at risk of losing, income and livelihood.[13] Misinformation and general unknowingness about the virus create deep uncertainty about the future. This may probably entail a long-term upsurge in the severity and the number of mental health problems (UN 2020). In attempting to secure public *physical* health, governmental restrictions may well be read as potentially putting public *mental* health into jeopardy.

Global surveillance

Besides restrictions on physical movement that entail the above-mentioned potential risks to our freedoms of movement and assembly, our rights to access education and

[6] Belgium, e.g., is re-imposing drastic social distancing measures in order to avoid a new general lockdown. For Belgium citizens, contacts outside family circles must be limited to the same five people over the month of August 2020. See e.g. Van Dorpe and Furlong (2020).

[7] E.g. California (Somerville 2020) and Victoria in Australia (Picheta 2020).

[8] Art. 13 (1) Universal Declaration of Human Rights (UNDHR), Art. 12 (2) International Covenant on Civil and Political Rights (ICCPR).

[9] Art. 20 (1) UDHR, Art. 21 ICCPR.

[10] Art. 26 (1) UDHR.

[11] Art. 25 (1) UDHR.

[12] Art. 23 (1) UDHR.

[13] E.g., between March and May 2020, widespread layoffs and furloughs have prompted about 20% of the US labor force to file for unemployment benefits. See e.g. Tappe (2020).

health institutions, and our right to work, many governments also rely on emerging technologies in their ‘fight’[14] against the pandemic. Those ‘digital measures’ may severely infringe our human right to privacy,[15] and may mark the transition into a world of surveillance technology.

The adopted emergency measures that engage new technologies aim primarily at analyzing the spreading pattern of the virus and at monitoring and enforcing curfews. Through relying on digital strategies, governments follow the World Health Organization’s recommendation to trace contacts between their citizens (WHO 2020d).

The emergency measures engaging new technologies may be roughly divided into five groups:[16]

Contact tracing apps: Contact tracing apps are designed to support curbing the spread of Sars-CoV-2 by tracking individuals and those they have come into contact with. Usually, if a person was found to be infected, the people she has been recently in contact with are informed. Often, they are then asked to self-quarantine. As of 3 July 2020, roughly 50 countries have been using contact tracing apps in dealing with corona: Australia, Austria, Azerbaijan, Bahrain, Bangladesh, Brunei, Bulgaria, Canada (Alberta), China (Tangermann 2020), Cyprus, Czech Republic, France, Georgia, Germany, Ghana, Hungary, Iceland, India, Indonesia, Iran, Israel, Italy, Japan, Jordan, Kyrgyzstan, Latvia, Malaysia, Mexico, New Zealand, North Macedonia, Norway, Peru, Philippines, Poland, Qatar, Saudi Arabia, Singapore, Slovakia, South Africa, South Korea, Spain, Switzerland, Thailand, Tunisia, United States of America, United Arab Emirates, Ukraine, United Kingdom, Uruguay, and Vietnam.[17] According to research conducted by TopVPN.com, about one third of the apps rely on GPS technology, a third on Bluetooth, and another third use both Bluetooth and GPS (Woodhams 2020).

*Digital Track*ing: Digital tracking includes the use of aggregated mobile location data to track citizens during lockdowns, apps designed to help identify the location of those with Sars-CoV-2,[18] and the deployment of advanced mobile monitoring technologies. As of 3 July 2020, 31 countries around the world have adopted digital tracking

[14] Covid-19 has brought up war rhetoric, see e.g. Goninet (2020).

[15] Art. 12 UDHR, art. 17 ICCPR.

[16] This section partly relies on research conducted by Woodhams (2020).

[17] A list of contact tracing apps per country can be found in the appendix.

[18] See e.g. the tracking-app developed with the support of Swiss researches at the EPFL in Lausanne, Handelszeitung (2020).

measures. E.g., government officials across the US are relying on location data from millions of cellphone users to better understand the movements of Americans during the pandemic, and how those movements may be affecting the disease (Tau 2020). The British government is working with major mobile network O2 to analyze its users' location data (Martin 2020). Other countries whose governments retrieve or had retrieved their citizens' geolocation data are Argentina (Davidovsky 2020), Austria (Mijnssen 2020), Belgium (Cloot 2020), Brazil (Mari 2020), Bulgaria,[19] China (Davidson 2020), Ecuador (EcuadorTV 2020), Finland (Telia 2020), Germany (Reikowski 2020), Guatemala (Estrada Tobar 2020) Hong Kong (Hui 2020), India (Srivastava and Nagaraj 2020), Iran (Gilbert 2020), Israel (Reuters 2020a), Italy (Vodafone 2020), Jordan,[20] Kazakhstan (Gussarova 2020), Morocco (Chahir 2020), New Zealand (Andelane 2020), Pakistan (Jahangir 2020), Poland (Privacy International 2020), Russia,[21] Singapore (Baharudin 2020), South Africa (Business Insider SA 2020), South Korea (Kim 2020), Spain (GovLab 2020), Switzerland (Reuters 2020b), Taiwan (Chen 2020), and Turkey (HRW 2020a).

Physical Surveillance: In order to slow the spread of Covid-19, governments are also adopting increasingly extensive physical surveillance measures. Those measures include the deployment of facial recognition cameras equipped with heat sensors, surveillance drones used to monitor citizens' movements, and extensive CCTV (Closed Circuit Television) networks. As of 3 July 2020, 11 countries have been using physical surveillance technologies to address Covid-19. The West Australian police force (Spires 2020), the New York Police Department,[22] UK police forces,[23] Belgian police,[24] and Madrid's police force[25] are increasingly relying on the use of aerial

[19] See tweet by Dr. Vesselin Bontchev from 24 March 2020: https://twitter.com/VessOnSecurity/status/1242503942409519106?s=20. Accessed 2 April 2020.

[20] The Jordanian app 'Cradar' is designed to allow citizens to inform Jordanian authorities about unauthorized gatherings.

[21] See an announcement by the Russian Government ordering the Ministry of Communications to develop a new contact tracing system to help monitor citizens thought to have come into contact with those that have the virus (Russian Government 2020).

[22] See the tweet from Spectrum News NY1: https://twitter.com/NY1/status/1243502731408670720. Accessed 2 April 2020.

[23] See the tweet from the Derbyshire police: https://twitter.com/DerbysPolice/status/1243168931503882241. Accessed 2 April 2020.

[24] See the tweet from Raphael-Antonis Stylianou, the EU Commission's Online Communications Officer: https://twitter.com/Stylianou_EU/status/1241405641266249728?s=20. Accessed 2 April 2020.

[25] See the tweet by BBC World News: https://twitter.com/BBCWorld/status/12392671525464678

footage through drones in order to enforce ongoing lockdowns and monitor citizen movements. Since the corona virus outbreak, also Russia (Reuters 2020c) and China (Kuo 2020, Shen 2020) are relying on a host of extensive surveillance mechanisms, including both drones and facial recognition cameras. Other countries using physical surveillance are the Bahrain (McArthur 2020), France (BBC), India,[26] and the United Arab Emirates (Al Monitor 2020).

Censorship: Since the outbreak of the corona virus, there has been an acceleration in the spread of false information (Woodhams 2020). In order to control and contain mis- and disinformation, governments have sought to regulate online content and promote official facts and figures from international health organizations. However, as of 3 July 2020, 18 governments have used the rise of mis- and dis-information about Covid-19 to justify censorship practices that aim at silencing regime critics and at controlling the flow of information. E.g., Cambodian (HRW 2020b) and Ugandan (Unwanted Witness 2020) authorities have arrested social media platform users that spread info about the virus. In Niger (CPJ 2020), authorities have arrested a journalist due to his coverage of the virus. Egypt (Al Jazeera 2020a) has taken away the press credentials of a British Journalist due to his alleged bad faith in how Egypt is dealing with the virus. Iran (Paganini 2020) blocked access to the Farsi language edition of Wikipedia due to criticism on how its authorities are handling the pandemic. Further countries leveraging the risk of false information about corona for censorship purposes are Azerbaijan (RSF 2020c), Bangladesh (RSF 2020a), China (Ruan and Knockel 2020), Hong Kong,[27] Japan (Denyer 2020), Kenya (Woodhams 2020), Russia (RFE 2020), Singapore

(Mahtani 2020), Thailand (HRW 2020c), Turkey (RSF 2020b), Turkmenistan (RSF 2020e), Venezuela (Cincurova 2020), and Zimbabwe (RSF 2020d).

Internet shutdowns: During the spread of a novel virus, access to and a free flow of reliable and correct information is urgent. Still, the governments of Bangladesh (HRW 2019),[28] Ethiopia (AFP 2020a), India (Ganai 2020), and

43?s=20. Accessed 2 April 2020.

[26] India's CG Covid-19 ePass requires citizens to register for an electronic pass to authorize travel. Users have to provide a photograph and an ID proof (Aadhar number).

[27] See the tweet by Kenneth Roth from 19 March 2020: https://twitter.com/KenRoth/status/1240671686258802692?s=20. Accessed 2 April 2020.

[28] Since 2019, the Bangladeshi government has shut down internet connections in its Rohingya

Myanmar (Al Jazeera 2020b) have restricted internet access some areas of their territories.

The description of ongoing monitoring and surveillance measures leads to two observations. First, our right to privacy may be severely infringed. And second, for the first time in human history, technology may make it possible to monitor almost everybody, almost everywhere, almost all the time. In other words, the corona panic and pandemic may let us slide into a world of global surveillance. Most unfortunately, due to the level of fear and panic, we seem to accept or even take part in those measures without the usual reflex of questioning them.

Potential permanence and inefficacy of emergency surveillance measures

In an exceptional situation, states may need additional powers to secure public safety and health. National constitutions as well as international human rights treaties[29] contain clauses that allow governments to temporarily suspend some of their obligations during a time of crisis. In those situations, governments can invoke special powers that would normally be considered infringements on human rights, even without formally declaring a state of emergency.[30] However, those powers are not absolute. Emergency measures must be *legal*[31] and *proportionate*,[32] as well as

refugee camps (HRW 2019).

[29] Art. 4 (1) ICCPR.

[30] While many states have enacted what have been described as emergency laws in response to the pandemic, not all of these countries have actually declared a ‘state of emergency’ under law. Hence, governmental behavior is not uniform. E.g., Armenia, Estonia, Georgia, Latvia, Moldova and Romania have declared a state of emergency according to Art. 15 of the European Convention on Human Rights (ECHR). Other countries in Europe, e.g. Italy and Spain have declared states of emergency in accordance with their constitutional provisions (AFP 2020b); see also Armstrong (2020). Others, like the UK, have introduced what politicians have described as ‘emergency powers.’ The UK government, e.g., convinced parliament to pass **lengthy legislation** allowing extra powers in less than a week, see the Coronavirus Act 2020 of 25 March 2020 (UK Parliament 2020).

[31] The restriction must be contained in a national law of general application. This law must be in force at the time the limitation is applied. The law must not be arbitrary, nor unreasonable. Further, it must be clear and accessible to the public.

[32] The limitation it must be somewhat ‘appropriate’ to achieve its protective function. Further, it must be the least intrusive instrument amongst all those that might achieve the desired result.

necessary and *time-bound*. What is more, government authorities carry the burden of justifying the restrictions (OHCHR 2020).

Restrictions must be *necessary* for the protection of public health. Most importantly, emergency measures can qualify as necessary only if they are also *efficacious*. An instrument is efficacious if it produces the intended effect. An instrument that is incapable of producing the intended effect, is, hence, not efficacious and cannot be necessary for achieving that effect. It follows that, in order to determine whether surveillance mechanisms can qualify as necessary measures, one must determine whether those measures can actually provide *reliable* and *useful* location information, i.e. whether they are efficacious.

Especially measures tapping personal smartphone information could not prove fully efficacious. How can cell phones be tracked? Cell phone towers are one option, but they provide only a very rough measure that is not useful to determine whether, e.g., a six-foot-proximity threshold is abided by. GPS signals are finer, but they work only outside, and can, therefore, not determine whether two people, e.g., sat in the same train wagon. What is more, as GPS drains battery, many people have it turned off in the first place. A WIFI network or Bluetooth beacon to which a smartphone is connected is a further location indicator. Still, the fact that two cell phones are connected to the same WIFI or Bluetooth does not say that they are not keeping a six-foot distance (Landau 2020; Stanley et al. 2020). Given that the majority of contact tracing apps rely on Bluetooth and GPS, those observations raise the question of those apps' effectiveness and, hence, necessity.

Besides the requirements of legality, proportionality, necessity, and non-discrimination, emergency legislation must be *time-bound.*[33] Unfortunately, crises have a habit to fast-forward certain processes and instruments, whose consequences may not disappear once the crisis is over. Hence, the surveillance measures endangering, in particular, our human right to privacy may not be terminated once the pandemic is successfully contained. Although lockdowns are being terminated now, the above-listed apps and digital instruments are, largely, still in place. Hence, the requirement of time limitation may well be neglected.

Two considerations may support the danger of persisting digital surveillance: On the one hand, digital surveillance could create financial pay-offs. If anything in the

[33] Peter Micek, Acces Now, Technology and human rights in times of crisis, WebDebate, DiploFoundation and Geneva Internet Platform, March 26 2020; OHCHR.

world is growing exponentially today, it is the provision of and the access to personal data. This may fuel the AI industry and could partially support the economic recovery once the virus spread is curbed.[34] What is more, the 'digital pandemic shock doctrine' does not only cover virus containment strategies or the monitoring and enforcing of curfews. Forced social distancing and isolation, and the shutting down of every-day institutions such as schools and workplaces, are a breeding ground for technologies that aim at re-installing our entire social life in the digital space. Our months-long pandemic isolation may well be a lab for a permanent contactless future of telehealth, broadband, and remote learning – highly profitable for businesses developing those services.

On the other hand, surveillance technologies may persist if people spread the perspective of the next crisis being 'just around the corner'. The speed of the Covid-19 panic wave was enormous, and the paralysis of reflection it created severe. Pre-emptive fear may corroborate and consolidate national and global surveillance mechanisms, and may make us blind to our duty to question them.

Panic and legitimation

The thought driving our rather precipitous behavior can be summarized as follows: 'Any measure necessary to save humanity is legitimate.'[35] This clause must undergo severe scrutiny:

What does humanity mean? If 'humanity' is referring to the human species, the pressing question is: Could Sars-CoV-2 extinguish the human species? Probably not. If the claim that corona puts global human existence at risk lacks considerable evidence, it may not ground legitimacy of severely rights-infringing measures. The argument that corona could put the normal functioning of health institutions into jeopardy is better founded. Yet, whether hospital overcrowding justifies the rush into surveillance is doubtable. If 'humanity' refers to what may slumber within each individual person, then the justification for increased surveillance may be even more fragile: Surveillance

[34] The tech giants Amazon, Alphabet, Apple and Facebook are already recording quarterly profit (Lopatto 2020).

[35] In addressing the pandemic, 'saving humanity' is one of the most prominent rhetorical references of the planet's most powerful leaders, see e.g. the Indian Prime Minister Narendra Modi (Economic Times 2020).

may violate human rights that protect precisely this seed of humanity each of us carries within.

Is panic the biggest risk? As it seems improbable that corona extinguishes the human race – especially given the speed with which governments have managed to paralyze public life and stop the first wave[36] – global panic seems equally ungrounded. Also, if we agree on a civilian duty to reflect upon whether government measures could infringe human rights, our state of excessive emotionality, or the gradual slide into a latent and passive fearfulness, must again be replaced by a state of reason – especially given emergency surveillance's potential permanence. Put differently, not only the pandemic, but also the panic must stop. If not, we will be incapable to reasonably reflect on whether, and if so, how, to opt out of the path fear has been pushing us on to. What is more, the emerging picture of wide-spread mental health issues due to the demanded isolation and panic may now require an even stronger effort to reclaim both our individual willingness and capacity for clear-sightedness.

The responsibility lies with all of us. Any institution is only as strong as the reflected minds of its members and the reflected minds of the population it aims to represent. On the one hand, this conclusion must guide media professionals. Their responsibility to curb fear and provide well-balanced facts in order to push us back to reason is enormous. On the other hand, it must guide those of us whose primary needs are currently still met. If we want to move into a balanced future, we must both reconquer and *use* our individual reflective capacity, which requires time and quietness. This may have been one profound advantage of the demand to stay at home, as both were, or still are, more easily accessible. We can and must regard our more isolated and socially cautious lives as an invitation for introspection, an increased level of self-understanding, and a new prioritization of values.

[36] Note that the question here is not whether high-level decision-makers decide for shutdowns with *moral* easiness.

References

AFP (2020a). Ethiopia vows to end communications blackout as virus cases rise. The Guardian. 31 March 2020. https://guardian.ng/news/ethiopia-vows-to-end-communications-blackout-as-virus-cases-rise/. Accessed 30 July 2020.

AFP (2020b). Italy declares state of emergency over coronavirus. France 24. 31 January 2020. https://www.france24.com/en/20200131-italy-declares-state-of-emergency-over-coronavirus. Accessed 2 April 2020.

Al Jazeera (2020a). Egypt targets Guardian, NYT journalists over coronavirus reports. Al Jazeera. 18 March 2020. https://www.aljazeera.com/news/2020/03/egypt-targets-guardian-nyt-journalists-coronavirus-reports-200318155434068.html. Accessed 2 April 2020.

Al Jazeera (2020b). Internet blackout in Myanmar's Rakhine enters its second year. Al Jazeera. 21 June 2020. https://www.aljazeera.com/news/2020/06/internet-blackout-myanmar-rakhine-enters-year-200621065709404.html. Accessed 30 July 2020.

Al Monitor (2020). Dubai police test using surveillance cameras to detect coronavirus. Al Monitor. 19 May 2020. https://www.al-monitor.com/pulse/originals/2020/05/dubai-cctv-coronavirus-surveillance-police-temperature.html. Accessed 30 July 2020.

Allen, M. (2020). Fight to survive: Coronavirus fallout threatens existence of small companies. Swissinfo.ch. 18 March 2020. https://www.swissinfo.ch/eng/fight-to-survive_coronavirus-fallout-threatens-existence-of-small-companies/45625050. Accessed 2 April 2020.

Andelane, L., (2020). Coronavirus: New Zealanders arriving home asked to consent to police tracking their location. Newshub. 2 April 2020. https://www.newshub.co.nz/home/new-zealand/2020/04/coronavirus-new-zealanders-arriving-home-asked-to-consent-to-police-tracking-their-location.html. Accessed 30 July 2020.

Andersen, K., Rambaut, A., Lipkin, W., Holmes, E., and Garry, R. (2020). The proximal origin of Sars-coV-2. *Nature Medicine 26*, 450-452, https://doi.org/10.1038/s41591-020-0820-9.

Armstrong, M., (2020). Covid-19: Spain extends state of emergency until 11 April. Euronews. 22 March 2020. https://www.euronews.com/2020/03/22/covid-19-spain-extends-state-of-emergency-until-11-april. Accessed 2 April 2020.

Baharudin, H., (2020). Coronavirus: Singapore develops smartphone app for efficient contact tracing. Straitstimes. 20 March 2020. https://www.straitstimes.com/singapore/coronavirus-singapore-develops-smartphone-app-for-efficient-contact-tracing. Accessed 2 April 2020.

BBC (2020). Coronavirus France: Cameras to monitor masks and social distancing. BBC News. 4 May 2020. https://www.bbc.com/news/world-europe-52529981. Accessed 30 July 2020.

Business Insider SA (2020). South Africa will be tracking cellphones to fight the Covid-19 virus. Business Insider SA. 25 March 2020. https://www.businessinsider.co.za/south-africa-will-be-tracking-cellphones-to-fight-covid-19-2020-3?fbclid=IwAR2SuMq5K3QiaX5UPs0XQg0pAXDWLh4j8INxDqxr3ftj1l_1lfdbPNLTMOs. Accessed 2 April 2020.

Chahir, A., (2020). Morocco’s coronavirus surveillance system could tip into Big Brother. Middle East Eye. 29 May 2020. https://www.middleeasteye.net/opinion/risks-moroccos-coronavirus-surveillance-system. Accessed 30 July 2020.

Chen, S., (2020). Taiwan sets example for world on how to fight coronavirus. Abc News. 13 March 2020. https://abcnews.go.com/Health/taiwan-sets-world-fight-coronavirus/story?id=69552462. Accessed 2 April 2020.

Cincurova, S., (2020). Venezuela arbitrarily detaining reporters covering COVID-19: CPJ- Al Jazeera. 3 May 2020. https://www.aljazeera.com/news/2020/05/venezuela-arbitrarily-detaining-reporters-covering-covid-19-cpj-200503152328224.html. Accessed 30 July 2020.

Cloot, A., (2020). Coronavirus: le cabinet De Block dit ‘oui’ à l’utilisation des données télécoms. Le Soir. 12 March 2020. https://plus.lesoir.be/286535/article/2020-03-12coronavirus-le-cabinet-de-block-dit-oui-lutilisation-des-donnees-telecoms. Accessed 2 April 2020.

Corman, V., Muth, D., Niemeyer, D., and Drosten, C. (2018). Hosts and Sources of Endemic Human Coronaviruses, *Advances in Virus Research 100*, 163-188, doi:10.1016/bs.aivir.2018.01.001.

CPJ (2020). Journalist Kaka Touda Mamane Goni arrested in Niger over Covid-19 report. CPJ. 24 March 2020. https://cpj.org/2020/03/journalist-kaka-touda-mamane-goni-arrested-in-nige.php. Accessed 2 April 2020.

Davidovsky, S., (2020). Un mapa online permite ver en qué zonas del pais se cumple mejor la cuarentena. La Nacion. 30 March 2020. https://www.lanacion.com.ar/tecnologia/un-mapa-online-compara-nivel-movilidad-personas-nid2348911. Accessed 30 July 2020.

Davidson, H., (2020). Chinese City plans to turn coronavirus app into permanent health tracker. The Guardian. 26 May 2020. https://www.theguardian.com/world/2020/may/26/chinese-city-plans-to-turn-coronavirus-app-into-permanent-health-tracker. Accessed 30 July 2020.

Denyer, S., (2020). Japan sets aside $22 million to buff government's global image amid pandemic struggles. The Washington Post. 15 Aril 2020. https://www.washingtonpost.com/world/asia_pacific/japan-coronavirus-image-abe/2020/04/15/73bf1dee-7f00-11ea-84c2-0792d8591911_story.html. Accessed 30 July 2020.

Economictimes (2020). Like Lord Buddha, India committed to save humanity; help world in defeating Coronavirus: PM Modi. Economictimes. 7 May 2020. https://economictimes.indiatimes.com/news/politics-and-nation/like-lord-buddha-india-committed-to-save-humanity-help-world-in-defeating-coronavirus-pm-modi/videoshow/75592675.cms?from=mdr. Accessed 2 August 2020.

EcuadorTV (2020). El gobierno autoriza rastreo satelital para mejorar vigilancia epidemologica. EcuadorTV. 17 March 2020. https://www.ecuadortv.ec/noticias/covid-19/romo-vigilancia-epidemiologico-covid19-?. Accessed 2 August 2020.

Estrada Tobar, J., (2020). Alerta Guate, la APP para informar sobre el coronavirus, puede recopilar tu información personal por 10 años. Nomada. 24 March 2020. https://nomada.gt/pais/actualidad/alerta-guate-la-app-para-informar-sobre-el-coronavirus-puede-recopilar-tu-informacion-personal-por-10-anos/. Accessed 30 July 2020.

European Commission (2020). Policy measures taken against the spread and impact of the coronavirus. 17 July 2020. https://ec.europa.eu/info/sites/info/files/coronovirus_policy_measures_17_july.pdf. Accessed 2 August 2020.

Ganai, N., (2020). https://www.outlookindia.com/website/story/india-news-4g-internet-ban-in-kashmir-extended-in-the-interest-of-sovereignty-o f-india/356311. Accessed on 1 August 2020.

Gilbert, D., (2020). Iran launched an app that claimed to diagnose Coronavirus – instead, it collected location data on millions of people. Vice News. 14 March 2020.

https://www.vice.com/en_us/article/epgkmz/iran-launched-an-app-that-claimed to-diagnose-coronavirus-instead-it-collected-location-data-on-millions-of-people. Accessed 2 April 2020.

Goninet, Anne-Sophie, (2020). COVID-19 And The Casualties of War Rhetoric. Worldcrunch. 24 April 2020. https://worldcrunch.com/coronavirus/covid-19-and-the-casualties-of-war-rhetoric. Accessed 2 August 2020.

GovLab (2020). Data Collaboratives in Response to Covid-19, https://docs.google.com/document/d/1JWeD1AaIGKMPry_EN8GjIqwX4J4KLQIAqP09exZ-ENI/preview#. Accessed 30 July 2020.

Gussarova, A., (2020). Kazakhstan uses electronic surveillance to enforce quarantine. The Jamestown Foundation. 8 April 2020. https://jamestown.org/program kazakhstan experiments-with-surveillance-technology-to-battle-coronavirus-pandemic/. Accessed 30 July 2020.

Handelszeitung (2020). Coronavirus: Schweizer Forscher entwickeln Tracking-App mit. 1 April 2020. https://www.handelszeitung.ch/panorama/coronavirus-schweizer-forscher-entwickeln-tracking-app-mit. Accessed on 2 April 2020.

HRW (2019). Bangladesh: Internet Blackout on Rohingya Refugees. 19 September 2019. https://www.hrw.org/news/2019/09/13/bangladesh-internet-blackout-rohingya-refugees. Accessed 30 July 2020.

HRW (2020a). Mobile Location Data and Covid-19: Q & A. 13 May 2020. https://www.hrw.org/news/2020/05/13/mobile-location-data-and-covid-19-qa. Accessed 30 July 2020.

HRW (2020b). Cambodia: Covid-19 Clampdown on Free Speech. 24 March 2020. https://www.hrw.org/news/2020/03/24/cambodia-covid-19-clampdown-free-speech. Accessed 2 April 2020.

HRW (2020c). Thailand: Covid-19 clampdown on free speech. 25 March 2020. https://www.hrw.org/news/2020/03/25/thailand-covid-19-clampdown-free-speech. Accessed 2 April 2020.

Hui, M., (2020). Hong Kong is using tracker wristbands to geofence people under coronavirus quarantine. Quartz. 20 March 2020. https://qz.com/1822215/hong-kong-uses-tracking-wristbands-for-coronavirus-quarantine/. Accessed 2 April 2020.

IATA (2020). Covid-19 Travel Regulations Map. https://www.iatatravelcentre.com/international-travel-document-news/1580226297.htm. Accessed 30 July 2020.

Jahangir, R., (2020). Govts starts cell phone tracking to alert people at virus risk. Dawn. 24 March 2020. https://www.dawn.com/news/1543301/govt-starts-cell-phone-tracking-to-alert-people-at-virus-risk. Accessed 2 April 2020.

Kaplan, J., Frias, L., and McFall-Johnsen, M., (2020). Our ongoing list of how countries are reopening, and which ones remain under lockdown. Business Insider. 29 July 2020. https://www.businessinsider.com/countries-on-lockdown-coronavirus-italy-2020-3?r=US&IR=T. Accessed on 2 August 2020.

Kim, M. (2020). South Korea is watching quarantined citizens with a smartphone app. Technologyreview.com. 6 March 2020. https://www.technologyreview.com/2020/03/06/905459/coronavirus-south-korea-smartphone-app-quarantine/. Accessed 1 August 2020.

Kuo, L., (2020). 'The new normal': China's excessive coronavirus public monitoring could be here to stay. The Guardian. 9 March 2020. https://www.theguardian.com/world/2020/mar/09/the-new-normal-chinas-excessive-coronavirus-public-monitoring-could-be-here-to-stay. Accessed 2 April 2020.

Landau, S., (2020). Location Surveillance to counter Covid-19: Efficacy is what matters. Lawfareblog. 25 March 2020. https://www.lawfareblog.com/location-surveillance-counter-covid-19-efficacy-what-matters. Accessed on 2 April 2020.

Lopatto, E. (2020). In the pandemic economy, tech companies are ranking it in. The Verge. 30 July 2020. https://www.theverge.com/2020/7/30/21348652/pandemic-earnings-antitrust-google-facebook-apple-amazon. Accessed 1 August 2020.

Mahtani, S., (2020). Singapore introduced tough laws against fake news – Coronavirus has put them to the test. The Washington Post. 16 March 2020. https://www.washingtonpost.com/world/asia_pacific/exploiting-fake-news-laws-singapore-targets-tech-firms-over-coronavirus-falsehoods/2020/03/16/a49d6aa0-5f8f-11ea-ac50-18701e14e06d_story.html. Accessed 2 April 2020.

Mari, A., (2020). Brazil introduces surveillance tech to slow the spread of the coronavirus. Zdnet. 27 March 2020. https://www.zdnet.com/article/brazil-introduces-surveillance-tech-to-slow-the-spread-of-coronavirus/. Accessed 2 April 2020.

Martin, A., (2020). Coronavirus: Government using mobile location data to tackle outbreak. Skynews. 19 March 2020. https://news.sky.com/story/coronavirus-government-using-mobile-location-data-to-tackle-outbreak-11960050. Accessed 2 April 2020.

McArthur, R., (2020) Bahrain launches electronic bracelets to keep track of active COVID-19 cases. MobiHealthNews. 8 April 2020. https://www.mobihealthnews.com/news/europe/bahrain-launches-electronic-bracelets-keep-track-active-covid-19-cases. Accessed 30 July 2020.

Mijnssen, I. (2020), https://www.nzz.ch/international/coronavirus-oesterreich-handy-daten-fuer-die-regierung-ld.1547014. Accessed on 1 August 2020.OHCHR (2020). Emergency Measures and Covid-19: Guidance. 27 April 2020. https://www.ohchr.org/Documents/Events/EmergencyMeasures_COVID19.pdf. Accessed 30 July 2020.

Paganini, P. (2020). https://securityaffairs.co/wordpress/98930/digital-id/coronavirus-iran-blocked-wikipedia-farsi.html. Accessed on 1 August 2020.

Picheta, R., (2020). Victoria declares ‘state of disaster’, locking down millions in Melbourne to fight a soaring coronavirus outbreak. CNN. 3 August 2020. https://edition.cnn.com/2020/08/02/australia/victoria-coronavirus-state-of-disaster-intl/index.html. Accessed 3 August 2020.

Privacy International (2020). Poland: Apps helps police monitor home quarantine. Privacy International. 19 March 2020. https://www.privacyinternational.org/examples/3473/poland-app-helps-police-monitor-home-quarantine. Accessed 2 August 2020.

Reikowski, K. (2020). Coronavirus: Telekom will Handydaten an das RKI übermitteln – droht jetzt noch ein Datenskandal? Merkur. 9 April 2020. https://www.merkur.de/welt/coronavirus-telekom-handydaten-rki-covid-19-krise-bevoelkerung-massnahmen-deutschland-mobilfunk-zr-13606412.html. Accessed 2 August 2020.

Reuters (2020a). Israel halts coronavirus cellphone surveillance, official says. Reuters. 9 June 2020. https://uk.reuters.com/article/us-health-coronavirus-israel-surveillanc/israel-halts-coronavirus-cellphone-surveillance-official-says-idUKKBN23G1MM. Accessed 1 August 2020.

Reuters (2020b). Swiss government asked Swisscom for data on people’s movements, says was not surveillance. Reuters. 26 March 2020. https://www.reuters.com/article/us-health-coronavirus-swiss-data/swiss-government-asked-swisscom-for-data-onpeoples-movements-says-was-not-surveillance-idUSKBN21D22Q. Accessed 30 July 2020.

Reuters (2020c). Moscow deploys facial recognition technology for coronavirus quarantine. Reuters. 21 February 2020. https://www.reuters.com/article/us-china-health-moscow-technology/moscow-deploys-facial-recognition-technology-for-coronavirus-quarantine-idUSKBN20F1RZ. Accessed 2 April 2020.

RFE (2020). CPJ calls on Russia to stop censoring news outlets on Covid-19. Radio Free Europe. 25 March 2020. https://www.rferl.org/a/cpj-calls-on-russia-to-stop-censoring-news-outlets-reporting-on-covid-19/30507738.html. Accessed 2 April 2020.

RSF (2020a). Bangladeshi journalists, cartoonist, arrested for Covid-19 coverage. Reporters Without Borders. 14 May 2020. https://rsf.org/en/news/bangladeshi-journalists-cartoonist-arrested-covid-19-coverage. Accessed 30 July 2020.

RSF (2020b). Turkish journalists arrested for reporting Covid-19 cases. Reporters Without Borders. 11 May 2020. https://rsf.org/en/news/turkish-journalists-arrested-reporting-covid-19-cases. Accessed 30 July 2020.

RSF (2020c). Azerbaijani reporter jailed for 30 days over coronavirus reporting. Reporters Without Borders. 22 April 2020. https://rsf.org/en/news/azerbaijani-reporter-jailed-30-days-over-coronavirus-reporting. Accessed 30 July 2020.

RSF (2020d). Five Zimbabwean reporters arrested while covering coronavirus lockdown. Reporters Without Borders. 10 April 2020. https://rsf.org/en/news/five-zimbabwean-reporters-arrested-while-covering-coronavirus-lockdown. Accessed 30 July 2020.

RSF (2020e). Coronavirus off limits in Turkmenistan. Reporters Without Borders. 31 March 2020. https://rsf.org/en/news/coronavirus-limits-turkmenistan. Accessed 30 July 2020.

Ruan, L., Knockel, J., and Crete-Nishihata, M., (2020). Censored Contagion: How information on the coronavirus is managed on Chinese social media. Citizenlab. 3 March 2020. https://citizenlab.ca/2020/03/censored-contagion-how-information-on-the-coronavirus-is-managed-on-chinese-social-media/. Accessed 2 April 2020.

Russian Government (2020). Decisions of the Russian Federation's Governmental Coordinating Committee for the Fight against the Spread of the Coronavirus Infection. http://government.ru/orders/selection/401/39243/. Accessed 3 August 2020.

Salcedo, A., and Cherelus, G. (2020). Coronavirus Travel Restrictions Across the Globe. The New York Times. 1 April 2020, https://www.nytimes.com/article/coronavirus-travel-restrictions.html. Accessed 20 July 2020.

Schweizerische Eidgenossenschaft BAG (2020). Neues Coronavirus: Massnahmen und Verordnungen. https://www.bag.admin.ch/bag/de/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/massnahmen-des-bundes.html#797337129. Accessed 2 April 2020.

Shen, X., (2020). Shanghai introduced QR codes on the subway in order to monitor its citizens, Abacus, Shanghai introduces QR codes on subway to track potential contact with corona. South China Morning Post. 28 February 2020. https://www.scmp.com/tech/article/3052880/shanghai-introduces-qr-codes-subway-track-potential-contact-coronavirus. Accessed 26 March 2020.

Somerville, E. (2020). California ordered into second lockdown after record daily rise of 11,800 coronavirus cases. Evening Standard. 21 July 2020. https://www.standard.co.uk/news/world/coronavirus-lockdown-california-cases-a4504271.html. Accessed 2 August 2020.

Spires, J., (2020). Western Australia police to use drones to blast messages out amid coronavirus. Dronedj. 30 March 2020. https://dronedj.com/2020/03/30/western-australia-police-drones-messages-coronavirus/. Accessed 2 August 2020.

Srivastava, R., and Nagaraj, A., (2020). Privacy fears as India hand stamps suspected coronavirus cases. Reuters. 20 March 2020. https://www.reuters.com/article/us-health-coronavirus-privacy/privacy-fears-as-india-hand-stamps-suspected-coronavirus-cases-idUSKBN21716U. Accessed 2 April 2020.

Stanley, J., and Stisa Granick, J., (2020). The Limits of Location Tracking in an Epidemic. ACLU. 8 April 2020. https://www.aclu.org/sites/default/files/field_document/limits_of_location_tracking_in_an_epidemic.pdf. Accessed 30 July 2020.

Surber (2020). Corona Pan(dem)ic: gateway to global surveillance? Geneva/ Zurich: ICT4Peace Foundation. https://ict4peace.org/wp-content/uploads/2020/04/2020_RSurber_Corona-pandemic_final-1.pdf. Accessed 3 August 2020.

Tangermann, V., (2020). In China, This Coronavirus App Pretty Much Controls Your Life. Futurism. 16 April 2020. https://futurism.com/contact-tracing-apps-china-coronavirus. Accessed 2 August 2020.

Tappe, A. (2020). 1 in 5 American workers has filed for unemployment benefits since mid-march. CNN Business. 7 May 2020. https://edition.cnn.com/2020/05/07/economy/unemployment-benefits-coronavirus/index.html. Accessed 30 July 2020.

Tau, B., (2020). Government Tracking: How People Move Around in Coronavirus Pandemic. The Wall Street Journal. 28 March 2020. https://www.wsj.com/articles/government-tracking-how-people-move-around-in-coronavirus-pandemic-11585393202. Accessed 2 April 2020.

Telia (2020). Telia's anonymized location data helps Finnish government fight the coronavirus. Teliacompany. 3 April 2020. https://www.teliacompany.com/en/news/news-articles/2020/telias-anonymized-location-data-helps-finnish-government-fight-the-coronavirus/. Accessed on 2 August 2020.

Thong, R. (2020). The Coronavirus exposes Education's Digital Divide. The New York Times. 18 March 2020. https://www.nytimes.com/2020/03/17/technology/china-schools-coronavirus.html. Accessed 2 April 2020.

UK Parliament (2020). Coronavirus Act 2020 of 25 March 2020. https://www.legislation.gov.uk/ukpga/2020/7/contents/enacted/data.htm. Accessed 2 August 2020.

UN (2020). United Nations Policy Brief: Covid-19 and the Need for Action on Mental Health. Executive Summary. 13 May 2020. https://unsdg.un.org/sites/default/files/2020-05/UN-Policy-Brief-COVID-19-and-mental-health.pdf. Accessed 30 July 2020.

UNESCO (2020). Covid-19 Educational Disruption and Response. https://en.unesco.org/covid19/educationresponse, https://en.unesco.org/covid19/educationresponse. Accessed 20 July 2020.

Unwanted Witness (2020). Coronavirus Covid-19 – Internet Censorship. Unwanted Witness. 23 March 2020. https://www.unwantedwitness.org/news-brief-ucc-tightens-social-media-censorship-as-uganda-registers-first-case-of-covid19/. Accessed 2 April 2020.

Van Dorpe, S., and Furlong, A., (2020). Belgium tightens restrictions amid fears of a second coronavirus wave. Politico.eu. 23 July 2020. https://www.politico.eu/article/belgium-coronavirus-restrictions-fear-of-second-wave/. Accessed 30 July 2020.

Vodafone (2020). Vodafone launches five-point plan to help counter the impacts of Covid-19 outbreak. Vodafone. 18 March 2020. https://www.vodafone.com/business/news-and-insights/company-news/vodafone-launches-five-point-plan-to-help-counter-the-impacts-of-the-covid-19-outbreak. Accessed 2 April 2020.

WHO (2020a). Coronavirus disease (COVID-19), Weekly Epidemiological Update –28 September, Data as received by WHO from national authorities by 10:00 CEST, 27 September 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200928-weekly-epi-update.pdf?sfvrsn=9e354665_2. Accessed 28 August 2020.

WHO (2020b). WHO Director-General’s opening remarks at the media briefing on Covid-19, 11 March 20. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed 2 April 2020.

WHO (2020c). Covid-19 and Noncommunicable Diseases. Survey conducted in May 2020. https://www.who.int/publications/i/item/9789240010291. Accessed on 2 August 2020.

WHO (2020d). Global surveillance for Covid-19 caused by human infection with Covid-19 virus. 20 March 2020. https://apps.who.int/iris/bitstream/handle/10665/331506/WHO-2019-nCoV-SurveillanceGuidance-2020.6-eng.pdf?sequence=1&isAllowed=y. Accessed 30 March 2020.

Woodhams, S. (2020). COVID-19 Digital Rights Tracker. Top10VPN.com. 20 March 2020. Updated on 3 July 2020. https://www.top10vpn.com/research/investigations/covid-19-digital-rights-tracker/. Accessed 30 July 2020.

Appendix: List of Contact Tracing Apps per Country

Australia	**COVIDSafe**
Austria	**Stopp Corona**
Azerbaijan	**e-Tabib**
Bahrain	**BeAware Bahrain**
Bangladesh	**Corona Tracer BD**
Brunei	**BruHealth**
Bulgaria	**ViruSafe**
Canada (Alberta)	**ABTrace Together**
China	**Close Contact Detector**
Cyprus	**CovTracer**
Czech Republic	**Mapy.cz, eRouska**
France	**StopCovid France**
Georgia	**Stop Covid**
Germany	**Corona-Warn App**
Ghan	**GH Covid-19 Tracker**
Hungary	**VirusRadar**
Iceland	**Rakning C-19**
India	**Aarogya Setu, SAIYAM - Track & Trace Together, COVID CARE, Covid Locator, Corona Watch, MahaKavach,COVID-19 Odisha, SMC COVID-19 Tracker, COVID-19 Quarantine Monitor Tamil Nadu, UP Self-Quarentine App, Uttarakhand CV 19 Tracking System**
Indonesia	**PeduliLindungi**

Iran	**Mask.ir**
Israel	**Track Virus, ‘The Shield’**
Italy	**Immuni**
Japan	**COCOA - COVID-19 Contact App**
Jordan	**AMAN App**
Kyrgyzstan	**Stop COVID-19 KG**
Latvia	**Apturi Covid Latvia**
Malaysia	**MyTrace**
Mexico	**CovidRadar.mx, Plan Jalisco Covid-19**
New Zealand	**NZ COVID Tracer**
North Macedonia	**StopKorona!**
Norway	**Smittestopp**
Peru	**PerúEnTusManos – Detén el avance del COVID19**
Philippines	**WeTrace**
Poland	**Kwarantanna domowa**
Qatar	**Ehteraz**
Saudi Arabia	**Tabaud**
Singapore	**Contact Tracer, TraceTogether**
Slovakia	**Zostaň Zdravý**
South Africa	**Covi-ID**
South Korea	**Corona 100m**
Spain	**COVID-19.eus**
Switzerland	**SwissCovid**

Thailand	**MorChana**
Tunisia	**E7mi**
USA	**Contact Tracer, SafePaths, HEALTHLYNKED COVID-19 Tracker, Healthy Together - COVID-19**
UAE	**TraceCovid**
Ukraine	**Action at Home**
United Kingdom	**NHS Covid-19**
Uruguay	**Coronavirus UY**
Vietnam	**Bluezone - Electronic mask**

ETHISCHE PROBLEME TÖDLICHER AUTONOMER WAFFENSYSTEME

Publiziert auf philosophie.ch, 21. Dezember 2020[1]

Tödliche Autonome Waffensysteme (LAWS, aus dem ‘Englischen Lethal Autonomous Weapons Systems’) sind eine neuartige Kategorie von Waffen, die, einmal aktiviert, ohne menschliche Mitwirkung ein Ziel identifizieren, suchen, auswählen und angreifen können.[2] Der Begriff ‘*System*’ bezieht sich auf das Zusammenspiel der Waffe mit einer integrierten Software, welche es der Waffe ermöglicht, all diese Funktionen ohne den Menschen – das bedeutet, technologisch eben hochgradig ‘*autonom*’ – auszuführen.[3] Autonome Software ist das Resultat jüngster Forschung vor allem in den Disziplinen KI und Robotik.[4] Weil diese Forschungsfelder rasant wachsen, werden wohl auch die Fähigkeiten eines Waffensystems in Zukunft immer raffinierter werden.

Ob ein Waffensystem nur dann ‘autonom’ genannt werden kann, wenn der Mensch gar keine

Mitwirkungsmöglichkeit mehr hat – z.B. auch keinen Stopp-Knopf mehr drücken kann – wird noch diskutiert.[5] Deshalb teilen sich die Meinungen, ob heute existierende Waffensysteme schon als LAWS gelten können. Selbstfliegende Flugzeuge mit hoch automatisierter Kampfsteuerung,[6] stationäre Kampfroboter als Grenzschutz,[7]

[1] https://www.philosophie.ch/beitraege/highlights/ethische-probleme-toedlicher-autonomer-waffensysteme

[2] Diese sind die vier sogenannten ‘kritischen Funktionen’ eines Waffensystems, oder der ‘Angriffszyklus’. Siehe ICRC, 2016, Convention on Certain Conventional Weapons, Meetings of Experts on Lethal Autonomous Weapons Systems (LAWS), April 11 – 15, 2016, Geneva, Switzerland, 1.

[3] Für eine gute Diskussion technologischer Autonomie, siehe z.B. Watson, David P., und Scheidt, David H., 2005, Autonomous Systems, Johns Hopkins APL Technical Digest 26(44), 368-376.

[4] Sie basieren aber auch auf anderen Disziplinen, z.B. der Mathematik, Psychologie und Biologie. Siehe z.B. Atkinson, David J., 2015, Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines, Foundation of Autonomy and Its (Cyber) Threats: From Individuals to Interdependence: Papers from the 2015 Spring Symposium, 6-17, 7.

[5] GGE.

[6] Siehe z.B. Dassault nEUROn, Dassault Aviation, https://www.dassault-aviation.com/en/defense/neuron/introduction/ (Zugriff 10. Oktober).

[7] Siehe z.B. Samsung SGR-A1, J. Kumagai, A Robotic Sentry For Korea's Demilitarized Zone,

Schwarmdrohnen[8] und hoch automatisierte Software zur Abwehr von Cyberangriffen[9] sind allerdings schon heute in Gebrauch.

Verglichen zu Waffensystemen, die von Menschen gesteuert werden, haben LAWS einen entscheidenden militär-ökonomischen Vorteil: Da, z.B., ein autonomes Kampfflugzeug nicht mehr von einem Piloten bedient werden muss, fallen jegliche Schutzmechanismen für den Menschen weg. Deshalb kann ein LAWS viel kleiner sein als von Menschen bediente Systeme – und deshalb enorm viel günstiger.[10]

Seit 2014 diskutiert ein spezielles UNO-Gremium sicherheitspolitische, operative, und vor allem international-rechtliche Probleme des Gebrauchs von LAWS im Krieg.[11] Der gemeinsame Grundpfeiler dieser Debatten ist allerdings ein zutiefst *ethisches* Problem: LAWS werfen die Frage auf, ob die Entscheidung über Leben und Tod eines Menschen auf Maschinen oder Software ausgelagert werden darf. Darf der Mensch die Kontrolle über diese Entscheidung *bewusst* abgeben? Wie kann man über diese Frage nachdenken?

Zuerst gilt es zu erörtern, ob LAWS aus ethischer Perspektive gar vorteilhaft sind. In diesem Zusammenhang wird einerseits angeführt, dass LAWS moralische und rechtliche Prinzipien besser respektieren könnten, da die Algorithmen,[12] welche der

in *IEEE Spectrum*, vol. 44(3), 16-17, March 2007, doi: 10.1109/MSPEC.2007.323429.

[8] Siehe z.B. STMs Kargu-Drohnen, https://www.stm.com.tr/en/kargu-autonomous-tactical-multi-rotor-attack-uav (Zugriff 10. Oktober 2020).

[9] Siehe z.B. Monstermind, Zetter, Kim, 2014, Meet MonsterMind, the NSA Bot That Could Wage Cyberwar Autonomously, Wired.com, https://www.wired.com/2014/08/nsa-monstermind-cyberwarfare/ (Zugriff 10. Oktober 2020).

[10] Dies birgt ein grosses sicherheitspolitisches Problem: Künftig könnten LAWS auf Kartoffelgrösse schrumpfen und als sich-selbst-koordinierende Kampfschwärme verwendet werden. Siehe z.B. Russel, Stuart, 2018, The new weapons of mass destruction?, The Security Times, February 2018, https://www.the-security-times.com/wp-content/uploads/2018/02/ST_Feb2018_Doppel-2.pdf (Zugriff 10. Oktober). Weitere Sicherheitspolitische Probleme sind das Risiko eines Rüstungswettlaufs sowie dasjenige der Proliferation, siehe z.B. Surber, Regina, 2018, AI: Autonomy, Lethal Autonomous Weapons, and Peace-Time Threats, Geneva: ICT4Peace Foundation.

[11] Group of Governmental Experts on Lethal Autonomous Weapons Systems, https://www.unog.ch/80256EE600585943/(httpPages)/5535B644C2AE8F28C1258433002BBF14?OpenDocument (Zugriff 10. Oktober). LAWS werden auch zu Friedenszeiten verwendet, siehe z.B. Heyns, Christof, 2016, Human Rights and the use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement, Human Rights Quarterly 38, 350-378.

[12] Ein Algorithmus ist eine mathematische Spezifikation dafür, wie man eine eine Klasse von mathematischen oder computerwissenschaftlichen Problemen lösen kann.

autonomen Software zugrunde liegen, enorm genau rechnen können und es deshalb dem LAWS erlauben könnten, einen Angriff sehr *präzise* auf militärische Objekte und Personen zu richten. Dies würde Risiken für Zivilisten minimieren.[13] Dieses Argument hängt allerdings vom Design und von der Kapazität der Software ab. Ob heutige Systeme tatsächlich zwischen einem Soldaten und einem Zivilisten unterscheiden können, ist höchst umstritten.[14] Andererseits könnte man anführen, dass durch LAWS weniger Soldaten sterben, da sie durch Maschinen und Software ersetzt würden. Dieses Argument gilt aber gleichermassen für ferngesteuerte Waffen, da auch diese den Soldaten vom Schlachtfeld entfernen und sein Leben dadurch besser schützen. Ethische Argumente für LAWS scheinen also eher schwach, und können deshalb den jetzt zu diskutierenden ethischen Problemen kaum die Schwere nehmen.

Wie oben angeführt, stellen LAWS die ethische Herausforderung, dass Entscheidungen über Leben und Tod eines Menschen von einer Software getroffen würden. Grob gibt es zweierlei Arten, wie man über dieses Problem nachdenken kann. Beide Gedankengänge führen zu ethischen Argumenten *gegen* LAWS.

Das erste Argument beruft sich auf die Menschenwürde. Der Kerngedanke ist, dass nicht nur zählt, *ob* ein Mensch getötet wird, sondern auch *wie*. Zwei sich feindlich gegenüberstehende Soldaten teilen, weil sie beide Menschen sind, dieselbe Erfahrung des eigenen Lebens und seines Wertes. Sie besitzen also, weil sie Menschen sind, das mögliche Bewusstsein für die Tragweite ihres Tötens. Für ein LAWS ist ein menschliches Zielobjekt aber gerade eben nur das: ein Objekt – ein Datenpunkt. Man kann deshalb sagen, dass das Töten durch ein LAWS für den Getöteten entwürdigend ist,[15] weil die Art dieses Tötens den Wert seines menschlichen Lebens minimiert.[16]

Das zweite Argument bezieht sich auf die moralische Verantwortung für das Töten durch ein LAWS. Wenn jemandem Gewalt angetan wird, muss bei irgendeinem Menschen die moralische Verantwortung liegen *können*. Wenn ein LAWS einen

[13] Siehe z.B. Arkin, Ronald, 2013, Lethal Autonomous Systems and the Plight of the Non-Combatant, AISIB Quarterly, July 2013, https://www.unog.ch/80256EDD006B8954/%28httpAssets%29/54B1B7A616EA1D10C1257CCC00478A59/$file/Article_Arkin_LAWS.pdf (Zugriff 10. Oktober 2020).

[14] Siehe z.B. ICRC, 2018, Ethics and autonomous weapons systems: An ethical basis for human control? Geneva, 3 April 2018, https://ict4peace.org/wp-content/uploads/2022/03/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf (Zugriff 10. Oktober 2020).

[15] Heyns, Christof, 2017, Autonomous weapons in armed conflict and the right to a dignified life: An African perspective, South African Journal on Human Rights 33(1), 46-71.

[16] UN Doc. A/HCR/34/47, § 109.

Angriff berechnet und ausführt, dann müssen zwei Dinge diskutiert werden, die für moralische Verantwortung wichtig sind: Absicht und Handlungsfähigkeit.[17] Da generell angenommen wird, dass eine Software kein Träger von moralischer Verantwortung sein kann, muss es sich also um *menschliche* Absicht und *menschliche* Handlungsfähigkeit handeln.[18] Vor allem die menschliche *Absicht* scheint bei LAWS aber schwierig zu garantieren: Einige Experten argumentieren, dass die menschliche Absicht *direkt* mit dem durch das LAWS resultierenden Angriff verbunden sein muss, damit menschliche moralische Verantwortung überhaupt vorhanden sein könnte. Der einzige Ort, an dem menschliche Absicht für einen LAWS-Angriff gesucht werden kann, wäre beim militärischen Befehlshaber.[19] Damit diese direkte Verbindung zwischen Absicht und Resultat aber gewährleistet sein könnte, müsste der Befehlshaber genau wissen, *wie* das LAWS funktioniert, und die Konsequenzen des von der Software ausgelösten Angriffs genau verstehen. Hier allerdings liegt der Hund begraben: Der ‘Output’ eines LAWS ist *nicht vorhersehbar*. Ein Mensch kann per Definition nicht wissen, wie sich ein LAWS genau verhält, da es den Angriff *selbst* initiiert. Es können zwar gewisse Zielgruppen und Zeiträume für einen Angriff vorprogrammiert werden. Innerhalb dieser Kategorien allerdings hat der Mensch keine ‘Wahl’ mehr – das LAWS berechnet, zielt und ‘schiesst’ von selbst. Ausserdem ist es bei technologisch sehr hochstehenden LAWS nicht einmal mehr möglich, rückwirkend den Rechenprozess zu verstehen, der zu einem bestimmten Angriff geführt hat. Denn LAWS Software integriert oft neueste KI Algorithmen, die auf ‘Machine Learning’ (ML)[20] basieren. Diese Algorithmen sind aber derart komplex, dass der Mensch den von ihnen berechneten Prozess gar nicht verstehen *kann*.[21] LAWS sind also aus zweierlei Gründen unvorhersehbar: aufgrund ihres Wesens, selber Handlungsmöglichkeiten zu sehen und zu wählen, und weil ihre Komplexität es dem Menschen verbietet, ihre Entscheidungsprozesse rückwirkend überhaupt zu verstehen. Folglich ist es kaum möglich zu sagen, wessen Absicht ein

[17] Leveringhaus, Alex, 2016, Ethics and Autonomous Weapons Systems, London: Palgrave Pivot.

[18] Sparrow, Robert, 2007, Killer robots, Journal of Applied Philosophy 24(1), 62-77; Roff, Heather, Killing in War: Responsibility, Liability and Lethal Autonomous Robots, in Allhoff, Fritz, Evans, Nicholas, and Henschke, Adam (eds.) Routledge Handbook of Ethics and War: Just War Theory in the 21st Century, 2014, New York: Routledge.

[19] Siehe z.B. ICRC, 2018.

[20] Machine Learning (ML) ist ein moderner probabilistischer Ansatz für Künstliche Intelligenz. ML befasst sich mit Algorithmen, die ‘lernen’ können, auf Basis gewisser Datenmengen eigene Voraussagen zu berechnen. Dies erlaubt es den Algorithmen, sich selbst durch ‘Erfahrung’ (d.h. neue Dateninputs) stetig zu verbessern.

[21] Dies wird oft als das sogenannte ‘Black Box Problem’ bezeichnet. Für genauere Ausführungen, siehe z.B. Surber, 2018.

LAWS widerspiegelt. Die Grundidee der ‘Autonomie’ sowie die Komplexität von LAWS geben also folgende Antwort auf die Frage *Wer ist verantwortlich?*: Niemand.

In Bezug auf die Frage moralischer Verantwortung stellt sich ein weiteres, sowohl ethisches als auch psychologisches Problem: Die obigen Ausführungen zeigen, dass das Töten mit Hilfe eines LAWS keine menschliche Entscheidung, sondern eine technologische Berechnung ist. Deswegen kann für dieses Töten niemand mehr wirklich verantwortlich sein. Nun *ist* es aber eine *menschliche Entscheidung, LAWS zu entwickeln und zu verwenden*. Der Mensch trägt deshalb die moralische *Verantwortung* dafür, dass er durch LAWS moralische Verantwortung für Gewaltanwendung *abtritt*.

Das Kernproblem ist deshalb Folgendes: Mit der Entwicklung von LAWS verkleinert der Mensch den Raum für mögliche menschliche Verantwortung auf der Welt. *Darf er das?* Diese Frage differenziert zu betrachten, würde das Ziel des vorliegenden Texts übersteigen. Vorläufig scheint aber folgende Behauptung plausibel: Damit sich die Welt ihren aktuellen Herausforderungen (Klima, Viren, etc.) stellen kann, sollten die Menschen ihre Verantwortung für Gewaltanwendung sowie für andere ihrer weitreichenden Entscheidungen wohl nicht in die unsichtbare Komplexität von Softwareprozessen abtreten, sondern diese Verantwortung bewusst ergreifen.

PHILOSOPHIE JENSEITS DER RATIO?

Menschliches Transformationspotential im Spiegel des Strebens nach Künstlicher Superintelligenz

Publiziert auf philosophie.ch 7. Dezember 2020[1]

Die Grundidee des Forschungsfeldes Künstliche Intelligenz (KI) ist es, Software zu kreieren, die intellektuelle Aufgaben ohne den und an Stelle des Menschen lösen kann.[2] Viele Theorien und Methoden der KI-Forschung basieren auf Theorien des menschlichen rationalen Denkens.[3] Einige sind gar an die Struktur des menschlichen Gehirns angelehnt.[4] Das KI-Forschungsfeld sucht also Wege, wie die Menschheit eine bis anhin ihr eigene Fähigkeit – das rationale Denken – der Technologie offerieren kann. Deshalb könnte man das KI-Forschungsziel auch eine Imitation des rationalen Denkens nennen. Das Ziel vieler Forscher ist es, das menschliche rationale Denken komplett zu imitieren. Diese sogenannt 'Starke KI' könnte jede intellektuelle Aufgabe so lernen und verstehen wie ein Mensch.[5] Überstiege die künstliche gar die menschliche

[1] https://www.philosophie.ch/beitraege/highlights/philosophie-jenseits-der-ratio

[2] Heute kann KI, z.B., fast passgenaue Filmvorschläge generieren (z.B. Netflix), medizinisch wichtige von weniger wichtigen Körperwerten unterscheiden, siehe z.B. Amisha et al., 'Overview of Artificial Intelligence in Medicine', *Journal of Family Medicine and Primary Care* 8, no. 7 (July 2019): 2328–31, https://doi.org/10.4103/jfmpc.jfmpc_440_19, fotorealistische Bilder und Videos erschaffen, siehe z.B. Generative Adversarial Networks: Antonia Creswell et al., 'Generative Adversarial Networks: An Overview', *IEEE Signal Processing Magazine* 35, no. 1 (January 2018): 53–65, https://doi.org/10.1109/MSP.2017.2765202. oder menschliche Sprache analysieren und kreieren, siehe z.B. Natural Language Processing: K. R. Chowdhary, 'Natural Language Processing', in *Fundamentals of Artificial Intelligence*, ed. K.R. Chowdhary (New Delhi: Springer India, 2020), 603–49, https://link.springer.com/chapter/10.1007/978-81-322-3972-7_19. though having a large content of knowledge, but it is becoming increasingly difficult to disseminate it by a human to discover the knowledge/wisdom in it, specifically within any given time limits. The automated NLP is aimed to do this job effectively and with accuracy, like a human does it (for a limited of amount text

[3] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall Series in Artificial Intelligence (Englewood Cliffs, N.J: Prentice Hall, 1995), 6-7.

[4] Das menschliche Gehirn dient als Inspirationsquelle von sogenanntem 'Deep Learning'. Siehe z.B. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, Adaptive Computation and Machine Learning (Cambridge Massachusetts, London England: The MIT Press, 2016).

[5] Russell and Norvig, *Artificial Intelligence*., 29.

Intelligenz, spräche man von Künstlicher Superintelligenz (KS).[6] Starke KI und KS sind hypothetische Zustände, über deren Realisierbarkeit die Wissenschaftler streiten.[7] Vor allem die KS tun viele als Scifi ab.

Ob KS je existieren wird, steht hier nicht zur Debatte. Man kann eine Welt mit KS heute nicht beobachten und soll hier nicht über Scifi-Szenarien fantasieren. Die Basis der nachfolgenden Gedanken ist eine andere Beobachtung: Die KI-Forschung *strebt nach KS*. In anderen Worten: Der Endzweck der KI-Forschung ist KS. Denn die Forschung will rationales Denken immer besser künstlich herstellen. Diese immer bessere Kreation von KI würde der Idee nach nur da gestoppt, wo diese KI so gut wird, dass sie als ‘Starke KI’ oder dann als KS das Kreieren selbst übernehmen würde.

Die Beobachtung dieses Strebens nach KS scheint legitim: Neugier und Forschungstrieb des Menschen sind nicht neu. Sie haben schon vor dem Ausprobieren der Atombombe nicht halt gemacht. Auch die KI-Forschung hat mit hoch automatisierten Kampfdrohnen-Schwärmen die Kriegspraxis von investitionskräftigen Staaten inzwischen im Griff.[8] Die Forschung schreitet auch enorm rasant voran und wird kräftig finanziert. Experten schätzen die globalen KI-Investitionen per 2025 auf 180 Milliarden Franken.[9] Diese Beobachtungen werfen eine wichtige Frage auf:

[6] Nick Bostrom, ‘Ethical Issues in Advanced Artificial Intelligence’, in *Machine Ethics and Robot Ethics*, by Wendell Wallach and Peter Asaro, ed. Wendell Wallach and Peter Asaro, 1st ed. (Routledge, 2020), 69–75, https://doi.org/10.4324/9781003074991-7. Der Moment, in dem KI die menschliche Intelligenz übertrifft, wird oft als Technologische Singularität bezeichnet, ‘TECHNOLOGICAL SINGULARITY by Vernor Vinge’, accessed 18 September 2020, https://frc.ri.cmu.edu/~hpm/book98/com.ch1/vinge.singularity.html.

[7] Berühmte Singularisten sind, z.B., Raymond Kurzweil: Ray Kurzweil, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, A Penguin Book (London: Penguin Books, 1999)., Ray Kurzweil, *The Singularity Is near: When Humans Transcend Biology* (New York: Penguin books, 2006).oder Bill Joy: ‘Why the Future Doesn’t Need Us | WIRED’, accessed 18 September 2020, https://www.wired.com/2000/04/joy-2/. Skeptiker sind Jaron Lanier, Chefstratege by Microsoft, Mitch Kapor, Gründer von Mozilla, Microsoft-Mitbegründer Paul Allen, oder Jaan Tallin, Mitbegründer von Skype. Für eine gute Diskussion und Gegenüberstellung von Singularisten und Skeptikern siehe z. B. Kurt Andersen, ‘Enthusiasts and Skeptics Debate Artificial Intelligence’, Vanity Fair, accessed 18 September 2020, https://www.vanityfair.com/news/tech/2014/11/artificial-intelligence-singularity-theory.

[8] V.a. die Türkei machte mit ihren kürzlich erworbenen Schwarmdrohnen des Typs ‘Kargu’ kürzlich Schlagzeilen: STM, ‘STM - KARGU’, STM, accessed 18 September 2020, https://www.stm.com.tr/kargu-autonomous-tactical-multi-rotor-attack-uav.

[9] ‘Global AI Investment to Top £150 Billion by 2025’, *Outside Insight* (blog), 31 July 2019, https://outsideinsight.com/insights/global-ai-investment-150-billion-2025/.

Welches Licht wirft das Streben nach KS auf den Menschen?

Seine rationale Denkfähigkeit hat dem Homo sapiens einen herausgehobenen Status der Evolution beschert.[10] Deshalb ist seine Spezies auf der Erde bis jetzt fast nur mit sich selbst konfrontiert.[11] Mit KS würden die Menschen intellektuell, eventuell gar physisch und biologisch,[12] aber etwas kreieren, wogegen sie womöglich verlieren würden. Ein Streben nach KS wäre also ein Streben nach Veränderungen der Konkurrenzbedingungen der menschlichen Spezies. Diese Veränderungen wären derart, dass das Merkmal 'Ratio', welches den Menschen höchst adaptiv und kaum angreifbar macht, sein Überleben wohl nicht mehr garantieren könnte. Der Mensch müsste sich also verändern, um zu überleben. Und diese Veränderung müsste *von seiner Ratio unabhängig* sein. Die Frage ist also Folgende:

Gibt es für den Menschen neue Entwicklungsmöglichkeiten, die nichts mit rationalem Denken zu tun haben? Bildlich gesprochen wäre es so: Wenn das Streben nach KS ein Spiegel wäre, in dem sich der Mensch betrachten wollte, könnte er dann in sich selbst neue, vom rationalen Denken unabhängige Entwicklungsmöglichkeiten erkennen? Könnte der Mensch biologisch und intellektuell so verharren, wie er ist, und über sein rationales Denken hinauswachsen? Könnte er eine Art Intellekt erwerben, der sich *qualitati*v vom rationalen Denken unterscheidet?

Und, wenn ja, *wie*? Wäre es ein Weg der *Evolution*, auf welchem der Mensch ein neues biologisches Merkmal entwickeln könnte? Leider kann man das im Voraus kaum

[10] 'Evolution des Geistes: Wie der Mensch das Denken lernte', accessed 18 September 2020, https://www.spektrum.de/magazin/wie-der-mensch-das-denken-lernte/828592.

[11] Der Kampf zwischen Menschen und Mikroben scheint allerdings noch nicht entschieden. Gleichzeitig scheint es auch plausibel zu sagen, dass die menschliche Ratio den Menschen gegen Erreger nicht schlecht ausrüstet, hat er doch in der Geschichte der Epidemien bis jetzt einige nennenswerte Erfolge erzielt (z.B. Pest-Impfung, erfolgreiche Eindämmung der Cholera-Pandemien, Impf- und Bekämpfungsprogramme von Pocken).

[12] Mithilfe von Tissue Engineering oder 'Gewebezucht' wird biologisches Gewebe künstlich hergestellt, siehe z.B. Aldo R. Boccaccini et al., 'A Composites Approach to Tissue Engineering', in *26th Annual Conference on Composites, Advanced Ceramics, Materials, and Structures: B: Ceramic Engineering and Science Proceedings* (John Wiley & Sons, Ltd, 2008), 805–16, https://ceramics.onlinelibrary.wiley.com/doi/10.1002/9780470294758.ch90. Composites Approach to Tissue Engineering\\uc0\\u8217{}, in {\\i{}26th Annual Conference on Composites, Advanced Ceramics, Materials, and Structures: B: Ceramic Engineering and Science Proceedings} (John Wiley & Sons, Ltd, 2008

beantworten. Denn neue Merkmale entstehen stets nur zufällig.[13] [14] Die Evolution ist kein kreativer Prozess. Sie hat kein Ziel. Hier wird aber ganz klar nach einer möglichen Entwicklung *hin zu einem Ziel* – dem Transzendieren des rationalen Denkens – gefragt. Wenn dieser Entwicklungsweg existiert und beschritten werden soll, muss also *der Mensch* irgendwie *tätig* werden. Deshalb wäre es kein Weg der Evolution, sondern eher einer der 'gewollten *Transformation'* des menschlichen Wesens.

Die Frage, wie der Mensch über sein rationales Denken hinauswachsen könnte, muss man also aktiv angehen. Eine grosse Hürde ist Folgende: *Kann* der denkende Mensch überhaupt über diese Frage nach*denken*? Einerseits sollte die Existenz eines a-rationalen Entwicklungswegs rational denkend zumindest annehmbar sein. Sonst wäre dieser Text selbst unsinnig. Aber andererseits: kann man denn einen a-rationalen Entwicklungsweg rational denkend *beschreiten*?

[13] Neue Merkmale sind stets eine Art Nebenprodukt von 'Fehlern', die bei der Weitergabe von Erbgut von Generation zu Generation auftreten. Wenn die Samenzelle des Vaters und die Eizelle der Mutter miteinander verschmelzen, entsteht die allererste Zelle des neuen Menschen. Sie enthält das Erbgut (Genom). Dieses enthält – verschlüsselt – die gesamte Bauanleitung für den späteren menschlichen Körper, zu dem sie werden wird. Bei jeder Zellteilung wird das gesamte Erbgut kopiert und an die beiden neuen Zellen weitergegeben. Bei dieser Kopie können Fehler auftreten.'Evolution des Geistes'.

[14] Ideen des sogenannten 'Transhumanismus' könnte man vielleicht als eine Art 'künstliche Evolution' bezeichnen. Der Transhumanismus diskutiert Ideen, wie der Mensch mittels neuer Technologien künstlich aufgewertet werden und mit potentieller KS in Konkurrenz treten könnte. Bis jetzt werden grob zwei Ideen diskutiert: Entweder der Mensch verschmilzt mit technologischen Implantaten und wird zu einer Art Hybrid-Wesen, das mit KS konkurrieren könnte. Siehe z.B. 'Elon Musk Unveils Plan to Build Mind-Reading Implants: "The Monkey Is out of the Bag"', the Guardian, 17 July 2019, http://www.theguardian.com/technology/2019/jul/17/elon-musk-neuralink-brain-implants-mind-reading-artificial-intelligence. Robert M. Geraci, *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*, Reprint Edition (New York; Oxford: Oxford University Press, 2012). Oder der Forschung gelingt es, das menschliche Genmaterial derart zu manipulieren, dass die menschliche Biologie schwieriger auszurotten wäre. Siehe hierzu z.B. Raya Bidshahri, 'The Power to Upgrade Our Own Biology Is in Sight—But Is Society Ready for Human Enhancement?', *Singularity Hub* (blog), 15 February 2018, https://singularityhub.com/2018/02/15/the-power-to-upgrade-our-biology-and-the-ethics-of-human-enhancement/. Diese Ideen bergen jedoch zwei Probleme: Einerseits wären solche Upgrades wohl mit enormen Kosten verbunden, weswegen sie sich nicht alle Menschen leisten könnten. Ein Ungleichgewicht zwischen technisch-biologisch verbesserten Menschen und denjenigen, die es nicht sind, wäre ethisch sehr schwierig zu rechtfertigen. Andererseits bergen vor allem Ideen einer Genmanipulation mit dem Ziel eines stark verlängerten Menschenlebens das Risiko einer Übervölkerung der Erde. Ohne Veränderung im Weltwirtschafts- und Finanzsystem würde dies wohl auf grosse Teile der Erdbevölkerung enormen existentiellen Druck ausüben.

Damit dies möglich wäre, müsste der Mensch aus seinem rationalen Denken herauskommen können, indem er rational denkt. Ist das machbar, oder kann der rationale Mensch stets nur an der 'Wand seines rationalen Denkens' entlang denken? Müsste der Mensch über sein rationales Denken vielleicht nicht rational nachdenken, sondern sein Denken auf eine andere Art und Weise *betrachten*? Wie käme er aber dahin? Müsste das rationale Denken vielleicht einer Art intuitivem Transformationswillen Platz machen? Ist ein solcher im Menschen selbst schon verborgen angelegt und er müsste sich diesem gegenüber 'lediglich' öffnen? Oder müsste der Mensch diesen Willen selbst kreieren? Und wie könnte man ein solches neues geistiges Element, wenn es denn gefunden wäre, in eine global verständliche Sprache übersetzen? Wer wäre der Übersetzer?

Eine Kaskade von Fragen – aber soll man sie sich stellen? Man könnte das Transformationsproblem mit der Begründung, dass KS wohl sowieso nie existieren wird, ja auf die lange Bank schieben oder als Spielerei abtun. Gleichzeitig könnte eine Auseinandersetzung mit dieser Frage aber auch ein grosses Potential für den Menschen und die Menschheit haben:

Es scheint plausibel zu sagen, dass das heutige menschliche Selbstverständnis ein globales Zusammenleben kreiert hat und noch immer kreiert, das viele Krisen hervorbrachte[15] und immer noch hervorbringt. Natürlich täte man ihm unrecht, wenn man den Menschen auf sein rationales Denken reduzieren würde. Er besitzt auch eine hohe emotionale und körperliche Intelligenz.[16] Aber an einer ganzheitlichen Wahrnehmung und einem (Aus-) Schöpfen all dieser Merkmale scheint doch etwas im Weg zu stehen. Denn würde der Mensch in Anbetracht dieser ausserordentlichen Eigenschaften mit sich selbst und anderen sonst nicht respektvoller umgehen?

[15] Und manche natürlich wunderbar bewältigte.

[16] Siehe z.B. John D. Mayer, Richard D. Roberts, and Sigal G. Barsade, 'Human Abilities: Emotional Intelligence', *Annual Review of Psychology* 59, no. 1 (January 2008): 507–36, https://doi.org/10.1146/annurev.psych.59.103006.093646. define EI, and describe the scope of the field today. We review three approaches taken to date from both a theoretical and methodological perspective. We find that Specific-Ability and Integrative-Model approaches adequately conceptualize and measure EI. Pivotal in this review are those studies that address the relation between EI measures and meaningful criteria including social outcomes, performance, and psychological and physical well-being. The Discussion section is followed by a list of summary points and recommended issues for future research.","container-title":"Annual Review of Psychology","DOI":"10.1146/annurev.psych.59.103006.093646","ISSN":"0066-4308, 1545-2085","issue":"1","journalAbbreviation":"Annu. Rev. Psychol.","language":"en","page":"507-536","source":"DOI.org (Crossref

Vielleicht wäre der Weg der Transformation, den die Menschen im Hinblick auf die angestrebte KS einschlagen könnten, eine Möglichkeit für eine ganzheitlichere Eigen- und Fremdwahrnehmung? Vielleicht bietet die angestrebte KS dem Menschen eine Chance, in eine zufriedenere Vernünftigkeit, gar in eine kultivierte Einsicht, wachsen zu können?[17]

Was wäre in diesem Zusammenhang die Rolle einer gelungenen Philosophie?

Vorab müsste man ein klares Argument dafür aufstellen, dass man ein solches menschliches Potential einfach mal in Betracht ziehen sollte. Dazu kann man vorweg nur anführen, dass das Recht zu behaupten, dass dieses Potential nicht existiert, nicht stärker ist als das Recht zu behaupten, dass es existiert.

Dann könnte die Philosophie den Menschen auf diesem Weg auch begleiten und unterstützen, indem sie die oben skizzierten Fragen systematisch aufwirft, bearbeitet, und Antworten sucht. In diesem Sinne wäre gelungene Philosophie gleichzeitig ein tiefes Ernstnehmen einer aktuellen Epoche und ein reflexives Instrument auf dem Weg eines menschlichen 'Werdens' zu einer allenfalls neuen Manifestation des menschlichen 'Seins'.

Diese Auseinandersetzung mit dem beschriebenen Spiegelbild könnte dann vielleicht auch für die Frage, ob KS je existieren wird, aufschlussreich sein. Wer weiss, womöglich würde ein neues menschliches Selbstverständnis den Drang zum Immer-Besser ein wenig relativieren?

[17] Da dieses Potential wohl von jedem einzelnen Menschen ergriffen werden könnte, könnte es auch alle nationalen, sozialen, religiösen oder andere Grenzen und Hierarchien transzendieren. Womöglich wäre alles, was aus diesem Potential erwächst, von diesen Grenzen ebenso unabhängig.

MANAGING THE RISKS AND REWARDS OF EMERGING AND CONVERGING TECHNOLOGIES: INTERNATIONAL COOPERATION, NATIONAL POLICY AND THE ROLE OF THE INDIVIDUAL

Paper Published on 23 April 2019 by the ICT4Peace Foundation, Geneva[1]

1. Broadening the perspective: beyond Lethal Autonomous Weapons Systems (LAWS)

Emerging technologies, such as machine learning, deep learning, robotics, biotechnology, additive manufacturing, and others, offer tremendous potential for good. However, as any other technology, they can be misused for negative purposes. An exemplary case with increasingly prominent media coverage are lethal autonomous weapons systems (LAWS), whose legal and ethical aspects, and challenges to peace and security are discussed within

the United Nations Convention on Certain Conventional Weapons (UN CCW).[2] Discussions on how to regulate the development and use of LAWS are of utmost importance. However, there are (at least) three other highly crucial aspects of the emerging technology landscape about which the global community must urgently gain more awareness and that also need to be properly addressed:

First, it is not only mathematical models for artificial intelligence (AI) and robotics, and it is not only LAWS, that are changing the landscape of international armed conflict.

[1] https://www.philosophie.ch/beitraege/highlights/philosophie-jenseits-der-ratio

[2] For ICT4Peace's in-depth analyses, see Surber, Regina, 2018: Artificial Intelligence: Autonomous Technology (AT), Lethal Autonomous Weapons Systems (LAWS) and Peace-Time Threats, Geneva: ICT4Peace Foundation, available at: https://ict4peace.org/wp-content/uploads/2018/02/2018_RSurber_AI-AT-LAWS-Peace-Time-Threats_final.pdf; Weekes, Barbara, 2018, Digital Human Security 2020, Geneva: ICT4Peace Foundation, available at: https://ict4peace.org/wp-content/uploads/2019/08/ICT4Peace-2018-Digital-Human-Security.pdf.

Different emerging technologies, such as quantum computing, additive manufacturing, or biotechnology, may *converge into a new weapons landscape*, which requires a breaking-up of the traditional weapons 'silos' of nuclear weapons, cyber-weapons/-attacks, biological weapons, or, more recently, LAWS.

Second, emerging technologies, such as LAWS, do not only have an effect on the individual and society during armed conflict but importantly also outside of war scenarios. These technologies raise broader social and human rights concerns relating to (data) privacy, bias and fairness, justice, and even existential risks for humanity. These concerns are prevalent independent of armed conflict. The paper highlights four of those highly transformative aspects: the new information landscape, the growing irrelevance of the human behind thedata, life-enhancement technologies, and how biomedicine is slowly creating a new understanding of human health.

And third, the UN Group of Governmental Experts' (GGE) debate on LAWS focuses on peace and security implications of emerging technologies and LAWS for traditional territorial state sovereignty. But the challenges arising from emerging technologies do not fit within our traditional concept of borders and state sovereignty and do not only affect the state as a collective construct. The challenges arising from emerging technologies are also *inherently local and citizen-based*, precisely because they affect an individual's data security, privacy, autonomy, or the (truth or falsehood of) information available. Therefore, it is key to bring individual human beings back into the epicenter of security concerns,[3] an urgency also highlighted by Sweden's Foreign Minister Margot Wallström at a recent arms control conference.[4]

These three aspects require a rethinking and a reshaping of traditional architectures both on the level of international arms control and disarmament, as well as on the level of national governance. Further, they require an integration of early ethical training into educational systems around the globe.

[3] 'Digital Human Security'.

[4] Statement by Margot Wallström, Capturing Technology – Rethinking Arms Control, Berlin, 16 March 2019.

2. Risks of emerging technologies beyond armed conflict[5]

Emerging technologies can have highly subtle, potentially permanent, and, therefore, very transformative effects on society. Those effects raise questions about (a) the self-understanding of the human being, (b), the role and make-up of social regulation, and (c) the perception society has of the individual.

These effects are structural, manifold, (still and potentially ever-) evolving, and, hence, require an immensely broad observational focus in order to be identified. Further, they require a holistic understanding of the interplay of emerging technologies. As the core of emerging technologies are technologically highly complex, and as they are developed at a very rapid pace, there exist exceptionally broad and currently unsolvable uncertainties about the trajectories of their future development, which in turn makes it difficult to delineate a clear risk environment. However, the beginning of certain social transformations resulting from emerging technologies can arguably already be observed.

2.1. Information: The blurring of truth

We live in a world where almost everyone has access to certain pieces of information. Those can be manipulated to offer exactly the piece of information that one individual, or a group of individuals, want or need to hear. The world has already witnessed incidents of mass information manipulation campaigns, targeting national elections and political parties, thereby undermining democratic processes.[6] In addition to *general* mass manipulation through widely spread disinformation, *individualized*[7] mis- or disinformation can also create an interesting landscape of perception: when people have access to different individualized news, a common reference point for knowledge is lost. Truth becomes something (even more) subjective and fluid. Further,

[5] For further examples of peace-time threats, see Surber, Regina, 2018, 16-18.

[6] See e.g., Hern, A., 2018, Cambridge Analytice: how did it turn clicks into votes?, The Guardian, available at: https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie (accessed on 23 April 2019).

[7] Cambridge Analytica has made lucrative use of those technological developments, see e.g. Hall, Jessica, 2017, Meet the weaponized propaganda that knows you better than yourself, Extremetech.com, March 1, 2017, accessible at: https://www.extremetech.com/extreme/245014-meet-sneaky-facebook-powered-propaganda-ai-might-just-know-better-know (accessed on February 15, 2018).

what is true, nowadays often depends on 'likes'. Therefore, quantitative support and not qualitative substance, seems to be the arbiter of truth. As a consequence, the borders between reality and artificial creation with regards to knowledge through individual research are blurring. This raises questions such as 'how might this affect social cohesion?', 'are we still *knowingly* shaping our (e.g. democratic) environment?', or 'do we need a human right to true information?'

2.2. Human data and AI

We live in a world where the individual human is arguably fading into irrelevance behind the vast economic and political possibilities of his/her data. Data can be willingly leveraged for economic and political interests, or for humanitarian purposes, e.g. when states try to attract tech companies that invest in AI by

offering them access to their citizens' data.[8] Or, when a "great power" trains its AI algorithms in developing countries to diversify its datasets.[9] Or, when refugees receive humanitarian aid only when giving away biometric data.[10] Also, data can unwillingly increase existing global inequalities, especially through insensitive choices in training data for AI applications in the medical sector. In the Global South, medical data is often scarce and 'bad'.[11] Hence, citizens from those resource-poor

[8] Moody, Glyn, 2017, Detailed medical records of 61 million Italian citizens to be given to IBM for its 'cognitive computing' system Watson, Privacy News Online, available at: https://www.privateinternetaccess.com/blog/detailed-medical-records-61-million-italian-citizens-given-ibm-cognitive-computing-system-watson/ (accessed on 23 April 2019).

[9] Council on Foreign Relations, 2018, Exporting Repression? China's Artificial Intelligence Push into Africa, available at: https://www.cfr.org/blog/exporting-repression-chinas-artificial-intelligence-push-africa (accessed on 23 April 2019).

[10] Indrajit, Sneha, 2017, The Cybersecurity Risks of Using Biometric Data to Issue Refugee Aid, The Henry M. Jackson School of International Studies, University of Washington, available at: https://jsis.washington.edu/news/cybersecurity-risks-using-biometric-data-issue-refugee-aid/ (accessed on 23 April 2019).

[11] Mate KS, Bennett B, Mphatswe W, Barker P, Rollins N., 2009, Challenges for routine health system data management in a large public programme to prevent mother-to-child HIV transmission in South Africa. PLoS One. 4(5): e5483; Carrell, D. S., Schoen, R. E., Leffler, D. A., et al., 2017, Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings, Journal of the American Medical Informatics Association 24(5), 986-991: 988-989; Fraser, Hamish S. F. et al., 2010, Implementing medical information systems in developing countries: what works and what doesn't, American Medical Informatics Association (AMIA) Symposium 2010, 232-236, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041413/pdf/amia-2010_sympproc_0232.pdf (accessed on

environments are generally excluded from clinical trials and from developments of AI systems for health care.[12] As differences in disease incidence between different ethnic groups or 'races' are scientifically well-established,[13] those AI health applications might not fit for a population subset underrepresented in the training data. Consequently, both conscious data geopolitics as well as missing consideration of existing inequalities when designing new technologies can lead to the exploitation of vulnerable communities and, thereby, enhance global inequality – something that the international community wants to reduce (SDG 10).

2.3. Life-enhancement technologies: From augmenting to invading

Life- or human-enhancement technologies (LETs or HETs respectively) may represent an a priori more 'physical' way of transformation. LETs/HETs aim to improve human physical, psychological or intellectual capabilities, and rely on a range of emerging technologies such as genetic modification or body implants. In principle, they could extend capacity beyond the typical range of human experience, e.g. not only restore missing eye-sight to normal, but make us see for miles. This rapidly advancing scientific field raises pressing social questions, e.g. 'what if LETs/HETs become mandatory, e.g. for police officers?', 'what if they recreate or augment inequality, because only 1% of society can afford them?'[14], or 'how autonomous is an individual who is 'modified' by deep-brain stimulation?'[15]

6 March 2019).

[12] Wahl, Brian, Cossy-Gatner, Aline, Germann, Stefan, and Schwalbe, Nina R, 2018, Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settinge?, BMJ Global Health, available at: https://gh.bmj.com/content/bmjgh/3/4/e000798.full.pdf (accessed on 20 February 2019).

[13] Coakley, Meghan, et al., 2012, Dialogues on Diversifying Clinic Trials: Successful Strategies for Engaging Women and Minorities in Clinical Trials, Journal of Women's Health 21(7): 713-716; see also e.g. Basu, D., Lopez, I., Kulkarni, A., and Sellin, J. H., 2005, Impact of race and ethnicity on inflammatory bowel disease, American Journal of Gastroenterology 100(10): 2254-2261.

[14] Whitman et al., 2018, What Americans Think of Human-Enhancement Technologies, Scientific American Blog Network, available at: https://blogs.scientificamerican.com/observations/what-americans-think-of-human-enhancement-technologies/ (accessed on 23 April 2019).

[15] Maslen, H., Pugh, J., and Savulescu, J., 2015, The Ethics of Deep Brain Stimulation for the Treatment of Anorexia Nervosa, Neuroethics 8(3), 215-230.

Besides tremendous ethical pressure to discuss those questions and many more, LETs/HETs also have a further aspect: In the future, possibilities to not only modify, but enhance our physical bodies and our cognitive functions might be more and more technological instead of biological – we might have body implants the size of micrometers.[16] In other words, technology might move from augmenting the human to invading the human body, with further implications when considering the IoT (Internet of Things). This might raise challenging issues with regards to hacking, and may require new methods to secure the physical integrity of the human being.

2.4. Health: A changing definition?

Advances in biomedicine and biotechnology might eventually lead to ever earlier diagnostics: through implanted monitoring devices, we might be capable of constantly controlling our bodily processes and notice slightest deviation from a pre-set 'healthy norm'. Further, as those controlling devices are individually-tailored and potentially implanted, health management might slowly move into the private sphere and more within the sphere of (perceived?) responsibility of individuals without any in-depth medical knowledge. Through constant individual supervision of bodily changes, the understanding of what is 'healthy' and what is (potentially) 'ill' might not depend on our individual physical and sensory feeling and awareness, but on our health monitoring devices. 'Feeling healthy' or 'feeling ill' might fall behind 'monitored health' and 'diagnosed disease'. Besides raising the question of whether a 'healthy' human is a human that can *feel* healthy or ill, biomedical research and developments seem to imply a *new understanding* of the term '*health*' (and 'illness'). Consequently, biomedical research might change our understanding of what it means to be healthy. As biomedical research and developments evolve at a highly rapid pace, we risk that the changed health landscape they produce sets a (new) definition of the (healthy human being *without* us having time to reflect upon this question, let alone guide research towards our *chosen* understanding of what is 'human'.

[16] Prof. Simone Schürle, Biomedicine, Personal Statement, 11 April 2019.

3. A three-fold strategy for future policy development

3.1. The danger of convergence: the underexamined interplay of emerging technologies and dual-use applications

Discussing emerging technologies in the area of LAWS, as the UN GGE's mandate requires, is a highly important task. However, it is crucial for the international community to understand that it is not only Machine Learning (ML) or Deep Learning (DL) and robotics, and that it is not only LAWS, that challenge international peace and security. Other emerging technologies, such as biomedicine, biotechnology, additive manufacturing, quantum computing or micro- and nanotechnology also (a) offer new ways of using traditional weapons, (b) enhance traditional weapons' lethality, accuracy, reach, and speed, or (c) may be used to create new weapons. Different emerging technologies may converge into a new weapons landscape, which requires a breaking-up of the traditional weapons 'silos' of nuclear weapons, cyber-weapons/-attacks, biological weapons, or, more recently, LAWS. Unfortunately, there currently exists a lack of a holistic understanding of emerging technologies, as well as a lack of understanding of the interplay of emerging technologies and the resulting security risks of potential dual-use applications.

For example, AI and robotics are drivers for autonomous weapons. But AI and robotics also make access and production of pathogens– bacteria and viruses for example – much easier because they can automate steps in the design process of a pathogen. Therefore, they can influence the production and proliferation of biological (and chemical) weapons. What is more, pathogens could potentially be deployed using autonomous drones, created through, e.g., 3D printing (additive manufacturing).[17] Further, autonomous intelligent agents are of great interest in the cyber domain. ML algorithms now offer the means to handle the incredible processing speed and the enormous amount of data used in cyber-operations, which the human cannot handle. In addition, they offer the flexibility that is needed to navigate within the fast-changing cyber environment, because they have the capability to learn and adapt. This makes cyber-operations cheaper, easier, and hence, more militarily lucrative.[18]

[17] Brockmann, K., Bauer, S., Boulanin, V., and Lentzos, F., 2019, New Developments in Biotechnology, Stockholm International Peace Research Institute (SIPRI), in: Capturing Technology. Rethinking Arms Control, Conference Reader, 25-32.

[18] King, M., and Rosen, J., 2018, The Real Challenges of Artificial Intelligence: Automating Cyber

What is more, quantum computing might change approaches to data security, because it offers novel ways to break encryption. This could have a game-changing effect for cyber-operations.[19] Cyber operations can be (and already are) used to sabotage nuclear weapons systems. Command and control-, alert- or launch systems of nuclear weapons could be targeted through cyber-attacks, and this could lead to accidental nuclear conflicts. This can have a 'game-changing' effect on the perceived value of nuclear weapons.[20]

There is a need to understand how emerging technologies converge into new weapons systems and weapons enhancements, which also leads to interconnection of 'classical' weapons categories. Separately analyzing and regulating different and currently pre-set weapons categories might not prove to be effective (anymore).

It could be advisable to create permanent international scientific expert groups for different weapons areas or technological sectors, that can continuously inform diplomatic debates, and that also regularly exchange on how their technological fields are converging.

3.2. The role of government: Re-claiming the regulation and safeguarding of basic rights and ethical principles in a digital world

National governments need to understand that the 'digital world' is an infrastructure like any other – if not the most important one. Currently, major tech companies are starting to create ethical principles (privacy, data security, transparency).[21] Those are

Attacks, Wilson Center, available at: https://www.wilsoncenter.org/blog-post/the-real-challenges-artificial-intelligence-automating-cyber-attacks. (accessed on 23 April 2019).

[19] Usas, A., 2018, The quantum computing cyber storm is coming, CSOOnline, available at: https://www.csoonline.com/article/3287979/the-quantum-computing-cyber-storm-is-coming.html (accessed on 23 April 2019).

[20] Van der meer, S., Cyber Warfare and Nuclear Weapons: Game-changing Consequences?, in: Meier, O., and Suh, E. (eds.), 2016, Reviving Nuclear Disarmament, Paths Towards a Joint Enterprise, Working Paper of the Research Division 'International Security', German Institute for International and Security Affairs, 37-38.

[21] See e.g. Artificial Intelligence Principles at Google: https://ai.google/principles/ (accessed on 23 April 2019), at Microsoft: https://www.microsoft.com/en-us/ai/our-approach-to-ai (accessed on 23 April 2019), or at IBM: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf (accessed on 23 April 2019).

principles that strive to safeguard basic rights, like the right to privacy or the right to physical integrity. Those rights are often guaranteed by national constitutions. Classically, if a new development or law risked to limit or violate a basic right, it needed to pass through parliament. However, now, with regards to potential risks of basic rights by emerging technology applications, the tech sector is taking on the task of deciding on the legality of limits and potential violations of those basic rights – and not governments. What is more, those ethical guidelines set up by representatives of the tech industry, are necessarily inspired by competitive thinking, and are developed under time pressure of global business. Whether or not this is 'ethical washing', i.e. marketing, or real added value, remains an open question. Generally, it is highly important not to 'abuse' ethical considerations and principles as a means to an end, but as an end in themselves.

Based on those observations, it would be advisable to create forums and mechanisms for increased dialogue between governments and the tech industry in order for governments to catch up on technological advances, and to develop appropriate policies to meet new social and political needs. It would also be constructive to create continuous polity-technology interfaces, e.g. through state departments for technology, that would generate the knowledge and understanding that governments need in the digital age.

3.3. Adapting education to the digital age: A bottom-up approach

As emerging technologies – as arguably any other technology – are dual-use, criminalizing them will also limit their tremendous potential for good. Hence, bans or prohibitions are not a practicable long-term strategy. As long as an individual (or a state) feels insecure, or has the potential need to harm another, dual-use tools will be used for this end. Consequently, we need to strive towards an altering of the human (or state) wish to harm. This goal requires understanding and a tremendous level of individual awareness of the new technological environment we live in, the social and ethical implications of new technologies, as well as awareness of individual responsibility for those implications. As this required transformation is located at the individual level, ICT4Peace calls for a bottom-up, educational approach.

Steps to raise awareness about those issues could be a promotion of responsible technological research, e.g. via fixed ethical guidelines for different technological fields, and/or a promotion of value-added design. Value-added design offers an approach to treat values, in addition to safety, as design specifications. For example, systems can be designed to maximize users' privacy. Determining the liability and responsibility as a design specification can sensitize engineers to the risks and societal impacts of the technologies they develop.[22] Further, it seems highly advisable to sensitize young researchers about ethical questions and social implications of their own research. Education must offer a toolkit about how to approach ethical questions relating to technological research and developments, in order for graduates to have the competence to answer these questions in their later day-to-day work. The sensitization of students regarding ethics, social questions and individual responsibility must, arguably, be included even at an earlier age prior to university. The reasons are two-fold. First, exposure to ethical reasoning at an undergraduate or graduate age might be 'too late'. Young adolescents choose an academic field, such as one of the MINT subjects, often also because those fields are so clearly delineated from philosophy and social sciences. Hence, the importance of ethical reasoning must be taught at an earlier age, so that it becomes *natural* to also study MINT subjects through an ethical lens. And secondly, as technological tools start to increasingly shape our environment without our input, very early stage reflection on individual human power, responsibility, and control is necessary.

Children and young adults have to learn through updated and technologically savvy educational programs that the way our society is built today is based on ideas and developments that we as humans have developed over hundreds of years. And they have to learn that those ideas and developments can be influenced and changed – by humans.

[22] Wallach, W., and Marchant, G., 2019, Toward the Agile and Comprehensive International Governance of AI and Robotics, Proceedings of the IEEE 107(3), 505-508.

ARTIFICIAL INTELLIGENCE: AUTONOMOUS TECHNOLOGY (AT), LETHAL AUTONOMOUS WEAPONS SYSTEMS (LAWS), AND PEACE TIME THREATS

Paper Published on 21 February 2018 by the ICT4Peace Foundation, Geneva.[1]

List of Abbreviations

AAAI	Association for the Advancement of Artificial Intelligence
ACM	Association for Computing Machinery
AI	Artificial Intelligence
AT	Autonomous Technology
CBM	Confidence Building Measures
CCW	Convention on Certain Conventional Weapons
DNA	Deoxyribonucleic Acid
EPSRC	Engineering and Physical Science Research Council
EURON	European Robotics Research Network
FLI	Future of Life Institute
GAN	Generative Adversarial Network
GGE	Group of Governmental Experts

[1] https://ict4peace.org/wp-content/uploads/2019/08/ICT4Peace-2018-AI-AT-LAWS-Peace-Time-Threats.pdf

HRW	Human Rights Watch
IDSIA	Instituto Dalle Molle di Studi sull'Instelligenza Artificiale / Dalle Molle Institute for Artificial Intelligence Research
IEEE	Institute of Electrical and Electronics Engineers
IHL	International Humanitarian Law
IHRL	International Human Rights Law
LAWS	Lethal Autonomous Weapons Systems
UN	United Nations

1. Introduction

The main purpose of this paper is to inform the international community on the risks of Autonomous Technology (AT) for global society. AT can be said to be the essence of Lethal *Autonomous* Weapons Systems (LAWS), which have triggered a legal and policy debate within the international arms control framework of the United Nations Convention on Certain Conventional Weapons (UN CCW) that is now entering its fifth year. Since LAWS highly challenge existing International Humanitarian Law (IHL) due to their capacity of replacing a human operator on a weapons platform, the CCW's tasks of, i.a., ensuring that the concepts of legal accountability and human responsibility do not become void, and assessing whether LAWS are legal under IHL, are of utmost importance.

However, LAWS are not the only manifestation of the security risks of AT. This paper will demonstrate further ways of the actual and potential weaponization of AT that are currently not yet fully addressed by the UN organizations. Moreover, AT not only poses risks to global society if *weaponized*, but can pose tremendous systemic risks to global society and humanity also when *not weaponized*. This potentially dangerous transformative power of AT, which is beyond the scope of the CCW's mandate, will be the thematic core of this paper. Based on a risk assessment of *not-weaponized* AT, the paper will present thought-provoking impulses that can shape an international interdisciplinary debate on the risks of AT specifically and of emerging technologies more generally.

In addition, this paper highlights risks underlying the application of terms originally referring to *human* traits to technological artefacts, such as ‘intelligence’, ‘autonomy’, ‘decision-making capacity’ or ‘agent’. It will argue that this unquestioned so-called ‘anthropomorphism’ leads to a premature revaluation of technology and a simultaneous potential devaluation of human beings, and will present ideas for linguistic substitutes.

At the same time, the paper will illustrate that the ‘classical’ understanding of ‘autonomy’ as human ‘*personal* autonomy’ has, in fact, donated its meaning to the current technological use of the term. However, this fact risks to be obfuscated by the broadening pool of diverse definitions and understandings of ‘autonomy’ for technological artefacts. Thereby, the paper will unearth the current paradigm shift in human technological creation and self-understanding that underlies the ongoing debate on AT and LAWS: The fact that humans are creating technological artefacts that may lose their instrumental character because we gradually give away control and responsibility for the outcomes of their usage. Locating the core challenge of AT, AI and any emerging technology in this still subtle but pervasive change in the understanding of the human-technology relationship, this paper will also provide conclusions and recommendations that are of a more general and long-term character.

The paper will be structured as follows: Chaper 2 and 3 will describe the current understandings, uses as well as the risks of those uses of the terms ‘Artificial Intelligence’ (AI) and ‘Autonomous Technology’ (AT). Chapter 4 will introduce the term ‘Lethal Autonomous Weapons System’ (LAWS), which will lead over to chapter 5 on the international discussions within the UN CCW and this UN debate’s limitations. Chapter 6 will present further ways of weaponizing AT that are ignored by the UN CCW, yet need immediate attention. Chapter 7 shows threats of AT for global society during peace-time. Chapter 8 presents three arguments for shaping an international debate on AT, AI and LAWS. Chapter 9 concludes and presents a list of recommendations. Eleven lists of principles for ethical/ responsible research on AI, AT and Robotics can be found in the Annex.

2. Artificial Intelligence (AI)

AI are two letters that represent the financially most lucrative scientific field that currently exists.[2] Moreover, they represent something that is often regarded as the fuel of the fourth industrial revolution, which takes place with an unprecedented pace compared to any other in human history.[3] However, the question what AI really *is* most often receives a rather vague and elusive answer. The reason for this lack of clarity may by two-fold.

First, the term 'Artificial Intelligence' includes the term 'intelligence.' 'Intelligence' originally has been used as a characteristic of humans. However, there neither exists a general understanding of this natural trait, nor a standard definition, despite a long history of research and debate.[4]

Precisely due to the growing research on AI, there exist strong incentives to define what the term 'intelligence' shall mean. This need is especially acute when artificial systems are considered that are significantly different to humans. This is

[2] The Economist, 2017, Coding Competition, The Battle in AI, The Economist Online, December 7, 2017, available at: https://www.economist.com/news/leaders/21732111-artificial-intelligence-looks-tailor-made-incumbent-tech-giants-worry-battle?frsc=dg%7Ce (accessed on December 11, 2017).

[3] See e.g. Wan, James, 2018, Artificial Intelligence is the fourth industrial revolution, Lexology.com, January 18, 2018, available at: https://www.lexology.com/library/detail.aspx?g=fccf419c-6339-48b0-94f9-2313dd6f5186 (accessed on January 31, 2018); Kelnar, David, 2016, The fourth industrial revolution: a primer on Artificial Intelligence (AI), Medium.com, December 2, 2016, available at: https://medium.com/mmc-writes/the-fourth-industrial-revolution-a-primer-on-artificial-intelligence-ai-ff5e7fffcae1 (accessed on January 31, 2018); Wright, Ian, 2017, Artificial Intelligence and Industry 4.0 – Taking the Plunge, Engineering.com, October 19, 2017, available at: https://www.engineering.com/AdvancedManufacturing/ArticleID/15871/Artificial-Intelligence-and-Industry-40--Taking-the-Plunge.aspx (accessed on January 31, 2018). Some experts also say that we are currently in the middle of a digital revolution, see e.g. Helbing, Dirk, 2017, A Presentation on Responsible Innovation and Ethics in Engineering, November 11, 2017, available at: https://www.youtube.com/watch?v=Jyv3QpRp9LA (accessed on February 7, 2018). Research by McKinsey suggests that AI could potentially transform global society '[...] ten times faster and 300 times the scale, or roughly 3000 times the impact,' Dobbs, R., Manyika, J. and Woetzel, J., 2015, The four global forces breaking all trends, McKinsey&Company, available at: https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-four-global-forces-breaking-all-the-trends (accessed on February 3, 2018).

[4] Helbing, Dirk, 2018, Personal Interview, February 9, 2018. For a list of 70 definitions of 'Intelligence' see Legg, Shane, and Hutter, Marcus, 2007, A Collection of Definitions of Intelligence, Frontiers in Artificial Intelligence and Applications Vol 157, 17-24.

the reason why researchers at the Swiss AI Lab IDSIA (Instituto Dalle Molle di Studi sull'Intelligenza Artificiale) created a single definition based on a collection of 70 definitions of 'Intelligence' by dictionaries, psychologists and AI researchers. They state that *'intelligence measures an agent's ability to achieve goals in a wide range of environments.'* This general ability includes the ability to understand, to learn and to adapt, since those are the features that enable an agent to solve a problem in a wide range of environments.[5]

It must be highlighted that the driving force behind the above-mentioned definition was to create a definitional reference point useful for *both human* as well as *technological artefacts*.[6] This ignores the fact that the term 'intelligence' was originally used to refer to a natural *human* capacity; and without a clear understanding of this human trait, we could possibly risk a revaluation of technology and a devaluation of human beings.[7]

And second, a reason for confusion about the meaning of AI may lie in the fact that the term AI is used to refer to two distinct but interrelated things. The distinction of those two possible understandings of AI will here be highlighted by two definitions of AI. However, we do not claim for these definitions to gain universal validity, as they would merely increase the existing pool of possible choices of such definitions. Yet, they should provide the reader with a first sense of caution when dealing with the application to technological artifacts of originally 'human terms' such as 'intelligence' or 'autonomy'. It is so that, at first glance, it might seem accurate and comprehensive to apply originally human terms to technological artefacts, since the latter are increasingly capable to perform 'actions' that resemble those of humans. However, the elaborations in this paper will show that this could prove to be risky.

On the one hand, AI refers to a scientific field, whose modern history started with the development of stored-program electronic computers,[8] but whose intellectual roots can already be found in Greek mythology.[9] As a scientific field, AI can be regarded as

[5] Legg and Hutter, 2007, 8.

[6] Legg, Shane, and Hutter, Marcus, 2006, A formal measure of artificial intelligence, Proc. 15th Annual Machine Learning Conference of Belgium and The Netherlands, 73-80, 73.

[7] Helbing, 2018.

[8] A computer that stores program instructions in electronic memory.

[9] See e.g. on the bronze man Talos from Crete, who can be regarded as incorporating the idea of an intelligent robot: Appollodorus, The Library, Book 1, Chapter 9, Section 26, Frazer, Sir James George (trnsl.), 1921, Cambridge, MA: Harvard University Press; London: William Heinemann Ltd; Apollonius Rhodius, Argonautica, Book 4, Section 1638 et seq., Seaton,

the attempt to answer the question of how the human brain gives rise to thoughts and feelings. AI as a research field began with the idea that '[...] every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.'[10] Therefore, AI refers to '[...] the *study* of the computations that make it possible to perceive, reason, and act';[11] it is the '[...] *effort* to make computers think [...];'[12] and it is the '[...] *art* of creating machines that perform functions that require intelligence when performed by people.'[13]

Bearing in mind the above-mentioned risk of devaluating humans in creating a definition of (artificial) intelligence without a human reference, AI shall here be understood as

(1) a scientific undertaking that is aiming to create software or machines that exhibit traits that resemble human reasoning, problem-solving, perception, learning, planning, and/ or knowledge.

Core parts of research on AI include: 'Knowledge engineering,' which aims at creating software and machines that have abundant information relating to the world; 'machine learning', which is the modern probabilistic approach to AI and that studies algorithms that 'learn' to predict from data; 'reinforcement learning', a sub-discipline of machine learning and currently the most promising approach for general intelligence that studies algorithms that learn to act in an unknown environment through trial and error; 'deep learning'[14], a very fast-moving and successful approach to machine learning based on neural networks, which has enabled recent breakthroughs in computer vision and speech recognition;[15] 'machine perception', which deals with the capability of using sensory inputs to deduce aspects of the world, 'computer vision',

R.C. (trnsl.), 1912, London: William Heinemann; Cohen, J., 1966, Human Robots in Myth and Science, London: Allen and Unwin.

[10] McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E., 1955, A proposal for the Dartmouth summer research project on artificial intelligence, 1.

[11] Winston, Addison-Wesley, 1992, Artificial Intelligence 3rd ed., Boston, MA: Longman Publishing Co, emphasis added.

[12] Haugeland, John, 1985, Symbolic Computation: Artificial Intelligence: The Very Idea, Cambridge, MA: The MIT Press, emphasis added.

[13] Kurzweil, Raymond, 1990, The Age of Intelligent Machines, Chapter 1: The Roots of Artificial Intelligence, 2, emphasis added.

[14] For a more detailed description of deep learning, see p. 5.

[15] Leike, J., AI Safety Syllabus, 80.000hours.org, available at: https://80000hours.org/ai-safety-syllabus/ (accessed on February 3, 2018).

the capability of analyzing visual inputs; and 'robotics', which deals with robots and the computer systems for their control.[16]

On the other hand, AI is also referred to the 'knowledge' or 'capacity' *embedded* in software or hardware architecture that are the result of the research on AI (1). Such capacities of software or hardware, e.g. the capacity to 'recognize' faces or voices or to 'drive' without a human behind a steering wheel, can be understood as artificially created intelligence – or AI. In this sense, AI can be regarded as a resource or a commodity, because it can be traded. Tech Giants around the world are rivalling, e.g. on the brilliance of algorithms.[17] Therefore, AI can be regarded both as a formless potential foundation of wealth as well as a resource for political leverage.[18]

In this sense, AI can also be understood as

(2) the formless capacity embedded in software and hardware architecture which enables the latter to exhibit traits that resemble human reasoning, problem-solving, perception, learning, planning, and/ or knowledge.

Current AI in the second sense of the term (2) is known as 'narrow' or 'weak' AI, in that it is designed to perform a narrow task, such as *only* driving a car or *only* recognizing faces. The long-term goal of many researchers, however, is to create so-called 'general' or 'strong' AI, sometimes also called 'artificial human-level intelligence'.[19] General AI is the formless capacity embedded in general purpose systems that are comparable to

[16] See e.g., Techopedia.com, Artificial Intelligence, available at: https://www.techopedia.com/definition/190/artificial-intelligence-ai (accessed on January 31, 2018).

[17] The Economist, 2017, Battle of the brains, Google leads in the race to dominate artificial intelligence, December 7, 2017, available at: https://www.economist.com/news/business/21732125-tech-giants-are-investing-billions-transformative-technology-google-leads-race (accessed on January 31, 2018).

[18] See e.g. CNBC, 2017, Putin: Leader in artificial intelligence will rule the world, September 4, 2017, available at: https://www.cnbc.com/2017/09/04/putin-leader-in-artificial-intelligence-will-rule-world.html (accessed on February 7, 2018); Metz, Cade, 2017, Google is already late to China's AI revolution, February 2, 2017, Wired.com, available at: https://www.wired.com/2017/06/ai-revolution-bigger-google-facebook-microsoft/ (accessed on February 7, 2018), Armbruster, Alexander, 2017, Künstliche Intelligenz: Google-Manager Eric Schmidt warnt vor China, Frankfurter Allgemeine Zeitung Online, November 2, 2017, available at: http://www.faz.net/aktuell/wirtschaft/kuenstliche-intelligenz/kuenstliche-intelligenz-google-manager-eric-schmidt-warnt-vor-china-15273843.html (accessed on February 7, 2018).

[19] Müller, Vincent C., and Bostrom, Nick, 2016, Future Progress in Artificial Intelligence: A Survey of Expert Opinion, In: Müller, Vincent C., (ed.), Fundamental Issues of Artificial Intelligence, Synthese Library; Berlin: Springer, 553-571, 553.

that of the human mind.[20] If general AI was achieved, this might also lead to ‘artificial superintelligence’, which can be defined as ‘[...] any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.’[21]

3. Autonomous Technology (AT)

AT is a result of research in the fields of AI and robotics, but also draws on other disciplines such as mathematics, psychology and biology.[22] Currently, there exists no clear understanding and no universally valid definition of the term ‘autonomous’ or AT in the context of AI and robotics. However, there exist different attempts.

Sometimes a purely operational understanding of ‘autonomy’ is used. In this sense, the term ‘autonomous’ may refer to any outcome by a machine or software that is created without human intervention. This could include, e.g., a toaster’s ejection of a bread slice when it is warm. In this form, autonomy would be equivalent to automation[23] and would not be limited to digital technology but could be used in analog technology or mechanics as well.[24] Hence, this understanding does not locate AT exclusively within the research field of modern AI.

Some experts use a narrower understanding and limit the use of the attribute ‘autonomous’ to more complex technological processes. They argue that AT extends beyond conventional automation and can solve application problems by using materially different algorithms and software system architectures.[25] This perspective

[20] Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S., and Sowa, J. F., 2012, Mapping the landscape of human-level artificial general intelligence, AI Magazine, 33(1), 25-42.

[21] Bostrom, N., 2014, Superintelligence: Paths, dangers, strategies, Oxford: Oxford University Press, Ch. 2.

[22] Atkinson, David J., 2015, Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines, Foundation of Autonomy and Its (Cyber) Threats: From Individuals to Interdependence: Papers from the 2015 Spring Symposium, 6-13, 7.

[23] Christen, Markus, Burri, Thomas, Chapa, Joseph, Salvi, Raphael, Santoni de Sio, Filippo, and Sullins, John, 2017, An Evaluation Scheme for the Ethical Use of Autonomous Robotic Systems in Security Applications, Digital Society Initiative (DSI) of the University of Zurich, DSI White Paper Series, White Paper No. 1, 36.

[24] Helbing, 2018.

[25] Land mines are an often-cited example of an automated weapon, see e.g. Ibid., 46.

is more narrow and clearly locates the emergence of AT within the research of modern AI.

In this sense, the key benefit of AT is the ability of an autonomous system to '[…] explore the possibilities for action and decide 'what to do next' with little or no human involvement, and to do so in *unstructured situations* which may possess *significant uncertainty*. This process is, in practice, *indeterminate* in that we cannot foresee all possible relevant information […]. 'What to do next' may include […] a step in problem-solving, a change in attention, the creation or pursuit of a goal, and many other activities […].'[26] In other words, a system is 'autonomous' if it can change its behavior during operation in response to events that are *unanticipated*,[27] e.g. a self-driving car's reaction to traffic jam, a therapist chatbot's[28] answer to a person's lamenting about her disastrous day, or a missile defense system that intercepts an incoming hostile one, like Israel's Iron Dome.

The theoretical AI approach that is at the core of AT in its narrow understanding, and that enables technological systems to perform the above-mentioned actions without a human operator, is deep learning. Deep learning software tries to imitate the activity of layers of neurons in the human brain. Through improvements in mathematical formulas and continuously increasing computing power of computers, it is possible to model a huge number of layers of virtual neurons. Through an inflow of a vast amount of data, the software can recognize patterns in this data and 'learn' from it.[29] This is key for 'autonomous' systems' reaction to unanticipated changes: due to new data inflow, the software can recognize new patterns and adapt to a changing 'environment'. Thereby, an autonomous system can, e.g., modify its actions in order to follow its goal or agenda.

It is crucial to highlight that deep learning mechanisms are so complex that the human cannot comprehend why a technological process based on deep learning creates the

[26] Atkinson, 2015, 7, italics added. For further elaborations a limited use of the term 'autonomy' to its more complex forms, see Russell, Stuart J. and Norvig, Peter, 2014, Artificial intelligence: a modern approach, Third Edition, Pearson Education: Harlow; Van der Vyver, J.-J. et al., 2004, Towards genuine machine autonomy, in: Robotics and Autonomous Systems, Vol. 46, No. 3, 151-157.

[27] Watson, David P., and Scheidt, David H., 2005, Autonomous Systems, Johns Hopkins APL Technical Digest 26(4), 368-376, 368.

[28] See e.g. the 24/7 Woebot that chats in order to improve someone's mood, available at: https://woebot.io/ (accessed on February 14, 2018).

[29] Burkhalter, Patrick, 2018, Personal Interview, February 14, 2018.

outcome it does.[30] Hence, outputs of autonomous systems may not only come as a surprise due to their core capacity of choosing a course of action undetermined by a human operator, but also due to the impossibility of locating the technological 'trigger' for a certain output.

At this stage one may answer to a possible fear that software or machines could by themselves create something that may resemble, e.g., a free will. It is so that autonomous systems may perform actions that are both unanticipated and ex post untraceable. However, the first programming of the software with the potential for future 'autonomous behavior' is the engineer's and programmer's decision, and not an unavoidable fact. And it is up to humans to discuss and set standards that ensure the development of beneficial and safe technology.

Since there exists no agreement whether automated system (e.g. a toaster) should already be regarded as autonomous (no human operator controls the ejection of the warm bread), some experts see it useful to think of 'autonomy as a continuum'[31] or of 'degrees of autonomy'.[32] They would characterize automated processes or semi-autonomous processes as 'autonomous', however to a lower degree than 'fully autonomous' systems.[33] This takes into account a blurring of definitional boarders and reflects the lack of a clear definition of 'autonomy' in AI and Robotics, but does not fill this gap.

There exists also no agreement whether or not a system could be classified as 'autonomous' if only certain aspects of its capacities function without human

[30] Ibid. See also Knight, Will, 2017a, The Dark Secret at the Heart of AI, MIT Technology Review, April 11, 2017, available at: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (accessed on February 16, 2018).

[31] Asaro, Peter, 2009, How just could a robot war be?, Proceedings of the 2008 Conference on Current Issues in Computing and Philosophy, 50-64, 51; Nicholas Marsh, Defining the Scope of Autonomy: Issues for the Campaign to Stop Killer Robots 2 (2014), available at: http://file.prio.no/Publication_files/Prio/Marsh%20(2014)%20-%20Defining%20the%20Scope%20of%20Autonomy,%20PRIO%20Policy%20Brief%202-2014.pdf (accessed on February 14, 2018); Michael Biontino, Summary of Technical Issues: CCW Expert Meeting on Lethal Autonomous Weapons Systems 1 (2014), available at http://www.unog.ch/80256EDD006B8954/%28httpAssets%29/6035B96DE2BE0C59C1257CDA00553F03/$file/Germany_LAWS_Technical_Summary_2014.pdf (accessed on February 14, 2018).

[32] Christen et al., 2017, 10.

[33] Schörrig, Niklas, 2017, Automatisierung in der Militär- und Waffentechnik, 27. ETH-Arbeitstagung zur Sicherheitspolitik, Autonome Waffensysteme und ihre Folgen für die Sicherheitspolitik, February 3, 2017.

intervention. Some experts argue that, e.g., a system that can function independently from external energy sources (autarkic), or one that can adapt its programming behavior based on previous data acquired ('learning'), could already be regarded as 'autonomous'.[34]

Some experts also claim that the attribute 'autonomous' is used for a technological artefact when it becomes (nearly) impossible for a human being to intervene in a technological process. In this sense, 'autonomy' is no term that covers a set of clearly defined characteristics (e.g. an artificial agent's capacity to 'learn', to be autarkic, to function independently from human control), but one that describes the *result of a technological process for which the human cannot or does not want to bear responsibility*.[35]

This view is influenced by the highly important and thus not negligible fact that the term 'autonomy' has a rich philosophical history, and refers to an unquantifiable attribute intrinsic to human personhood. There are two distinct but interrelated understandings of 'autonomy' as a human attribute.

'Personal autonomy', on the one hand, refers to self-governance or the capacity to decide for oneself and follow a course of action in one's life, independent of moral content.[36] This necessarily leads to personal responsibility for the course of action taken.

On the other hand, 'moral autonomy,' usually traced back to Immanuel Kant, can be understood as the capacity of an individual human to deliberate, understand and give oneself the moral law. For Kant, it is by virtue of our autonomy that we are moral beings that can take on moral responsibility. At the same time, we are moral to the extent that we are autonomous.[37]

Having in mind this second classical understand of *moral* autonomy, connected with the fact that the term 'autonomy' is used when referring to software and machines, may have prematurely supported the idea of and fueled discussions about 'autonomous' robots that may also behave *morally* and *ethically*.[38] Both a precise technological

[34] Christen et al., 10.

[35] Helbing, 2018.

[36] Dryden, Jane, Internet Encyclopedia of Philosophy, Autonomy, available at: https://www.iep.utm.edu/autonomy/ (accessed on February 1, 2018).

[37] Kant, Immanuel, 1998 (1785), Groundwork for the Metaphysics of Morals, Cambridge: Cambridge University Press.

[38] Arkin, Ronald, 2009, Ethical Robots in Warfare, IEEE Technology and Society Magazine 28(1),

understanding as well as a careful linguistic usage may minimize or eliminate the risk of a (potentially unconscious) terminological confusion.[39]

However, it is barely possible to completely strip off a term from its 'classical' meaning. And the fact that 'autonomy', when used to characterize technological processes, does so when the latter create outcomes for which humans have a hard time taking control – in other words, when they actually *do* give away the capacity to decide for an action that leads to a technological process' outcome – there clearly exists an overlap of the 'classical' understanding of *personal* autonomy and the technological use of the term.

Due to this common contentual denominator of *personal* autonomy and 'autonomy' for technological artefacts, one could argue that the international debate about a definition of 'autonomy' for artifacts clearly distinguished from *personal* autonomy is misguided. The reason is that the technological use of the term 'autonomy' precisely uses this term in order to highlight a notion of 'self-governance' of an artifact. And whether or not this 'self-governance' is in fact technologically possible, one must *not* ignore that research endeavors to create 'autonomous' systems bear an immense risk of going hand in hand with losing human control over outputs (deep learning) and relinquishing human responsibility for outcomes. And this risk is independent of the term itself. In other words, a distinct definition of 'autonomy' for artefacts, measurable and potentially existing to degrees, obfuscates the fact that humans are creating technological instruments that may lose their instrumental character because we gradually give away responsibility for the outcomes of their usage.

Consequently, agreeing that 'autonomy' for artefacts is a term willingly borrowed from its 'classical' usage of personal self-governance, and intrinsically linked to responsibility, would shed a different light on the creation of autonomous artifacts

30-33; Arkin, Ronald, 2010, The case for ethical autonomy in unmanned systems, Journal of Military Ethics 9(4), 332-341; Arkin, Ronald, 2017, A roboticist's perspective on lethal autonomous weapon systems, UNODA Occasional Papers No. 30, New York: United Nations, 35-37; Anderson, M., Anderson, S., and Armen, C., 2004, Towards Machine Ethics, AAAI-04 Workshop on Agent Organizations: Theory and Practice, San Jose, CA; Anderson, M., Anderson, S., and Berenz, V., 2016, Ensuring Ethical Behavior from Autonomous Systems, Proc. AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments, available at: http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12555 (accessed on February 4, 2018); Moor, J., 2006, The Nature, Importance, and Difficulty of Machine Ethics, IEEE Intelligent Systems, July/August, 18-21; McLaren, B., 2005, Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning, 2005 AAAI Fall Symposium on Machine Ethics, AAAI Technical Report FS-05-06.

[39] See 6.2.

and thus lead to a different question: Why are we aiming at limiting the space for responsible human action instead of increasing it? It is highly important not to lose oneself in technological definitions of ‘autonomy’. ‘Autonomy’ for artifacts is a term that could function as an excuse for relinquished human responsibility for ‘ugly’ and potentially immoral outcomes, i.a., the killing of human beings in the case of LAWS.

4. Lethal Autonomous Weapons Systems (LAWS)

AT can supplant the human being from the decision-making process in a certain area. This can have an enormous potential for good (e.g. autonomously driving cars for visually impaired people, surgical robots[40]). However, besides promising applications of AT, autonomous software can be (and arguably already are) integrated into robots that can select and engage a (military) target (e.g. infrastructure and potentially also combatants) without a human override. Often-called Lethal Autonomous Weapons Systems (LAWS), as yet, there exists no agreed definition of LAWS. One reason for this lack of definition is that there exists, as highlighted above, no general understanding of the term ‘autonomy’ in AI and robotics.

The general idea is that a LAWS, once activated, would, with the help of sensors and computationally very intense algorithms, identify, search, select, and attack targets without further human intervention. Whether the human being can still overpower or veto an autonomous weapon’s ‘decision’ in order for it to be called a LAWS, is also debated.[41] However, military operational necessity precisely seem

[40] See e.g., Strickland, Eliza, 2017, In Flesh-Cutting Task, Autonomous Robot Surgeon Beats Human Surgeons, IEEE Spectrum, October 13, 2017, available at: https://spectrum.ieee.org/the-human-os/biomedical/devices/in-fleshcutting-task-autonomous-robot-surgeon-beats-human-surgeons (accessed on February 1, 2018).

[41] The US Department of Defense defines a weapons system as autonomous if it ‘[...] can select and engage targets without further intervention by a human operator.’ Department of Defense, Directive 3000.09, November 21, 2012, 13-14; The UN Special Rapporteur on extrajudicial, summary or arbitrary executions adds the element of choice: ‘The important element is that the robot has an autonomous “choice” regarding selection of a target and the use of lethal force.’ Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christoph Heyns, UN doc. A/HRC/23/47, § 38; Human Rights Watch (HRW) distinguishes level of autonomy in weapons systems and contrasts the terms ‘human-out-of-the-loop’ and ‘human-on-the-loop’. A ‘human-out-of-the-loop’ weapon is ‘[...] capable of selecting targets and delivering force without any human input or interaction [...].’ In other words, a ‘human-out-of-the-loop’ weapon’s decision cannot be vetoed by a human being. On the other hand side, a ‘human-on-the-loop’ weapon can ‘[...] select targets and deliver force

to require weapons systems which can function once human communication links break down.[42] Furthermore, the state-of-the-art research on AI is currently creating software which can 'learn' entirely on its own[43] and even 'learn' to 'learn' on its own.[44] Hence, (precursor) technologies for creating fully 'human-out-of-the-loop'[45] weapons systems already exist.

From a military perspective, LAWS have many advantage over classical automated ore remotely controlled systems: LAWS would not depend on communication links; they could operate at increased range for extended periods; fewer humans would be needed to support military operations; their higher processing speeds would suit the also increasing pace of combat;[46] by replacing human soldiers, they will spare lives; and with the absence of emotions such as self-interest, fear or vengeance, their 'objective' 'decision-making' could lead to overall outcomes that are less harmful.[47]

However, the use of LAWS may also generate substantial threats: Generally, LAWS may change how humans exercise control over the use of force and also its consequences.

under the oversight of a human operator who can override the robots' actions [...]'. According to HRW, both types can be considered 'fully autonomous weapons' when supervision is so limited that the weapon can be considered 'out-of-the-loop.' Docherty, B., 2012, Losing Humanity: The Case Against Killer Robots, Human Rights Watch, November 2012, available at: https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots (accessed on February 1, 2018); The ICRC defines autonomous weapons systems as '[...] (a)ny weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention.' ICRC, 2016, Convention on Certain Conventional Weapons, Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), April 11 – 15, 2016, Geneva, Switzerland, 1.

[42] Adams, T., 2002, Future Warfare and the Decline of Human Decision making, Parameters, U.S. Army War College Quarterly, Winter 2001-02, 57-71.

[43] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A, Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. and Hassabis, D., 2017, Mastering the game of Go without human knowledge, Nature vol. 550, 354-359.

[44] See e.g., Finn, Chelsea, 2017, Learning to Learn, Berkeley Artificial Intelligence Research, available at: http://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/ (accessed February 2, 2018).

[45] Docherty, B., 2012.

[46] Thurnher, J., 2014, Examining Autonomous Weapons Systems from a Law of Armed Conflict Perspective, in: Nasu, H., and McLaughlin, R. (eds.), New Technologies and the Law of Armed Conflict, TMS Asser Press, 213-218.

[47] ICRC, 2011, International Humanitarian Law and the Challenges of Contemporary Armed Conflicts, Official Working Document of the 31st International Conference of the Red Cross

Further, humans may also not be able any more to predict who or what is made the target of an attack, or even explain why a particular target was chosen by a LAWS. This fact raises serious legal, ethical, humanitarian and security concerns.[48] From a humanitarian and ethical point of view, e.g., LAWS could be regarded as diminishing the value of human life as a machine and not a human being 'decides' to kill.[49] Also, the physical and emotional distance between the programmer or engineer of a LAWS and the targeted person may generate an indifference or even a 'Gameboy Mentality' on the side of the former.[50] From a security perspective, LAWS could be dangerous because they may also be imperfect and malfunction.[51] Moreover, the farther technology advances, the more the level of autonomy of a LAWS increases. This, further, leads to an increased unpredictability of outcomes of LAWS and may enable the interaction of multiple LAWS as e.g. self-organizing swarms.[52]

The focus of scholarly inquiry of the legality of LAWS was mainly on IHL,[53] which presents significant challenges for both the development and the use of LAWS, since the latter would face problems to meet IHL's requirements of distinction,[54]

and the Red Crescent, November 28 – December 1, 2011.

[48] Geneva Academy, 2017, Autonomous Weapons Systems: Legality under International Humanitarian Law and Human Rights, https://www.geneva-academy.ch/news/detail/48-autonomous-weapon-systems-legality-under-international-humanitarian-law-and-human-rights (accessed on February 2, 2018).

[49] UN Doc. A/HRC/23/47, § 109.

[50] Sassòli, Marco, 2014, Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified, International Law Studies Vol. 90, 308-340, 317.

[51] ICRC, 2014, Expert Meeting on 'Autonomous weapons systems: technical, military, legal and humanitarian aspects', March 26 – 28, 2014, Report of November 1, 2014, available at: https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014# (accessed on February 1, 2018).

[52] ICRC, 2016, 3.

[53] The reason for this legal focus on LAWS based almost exclusively on IHL is the fact that the UN CCW is underpinned by IHL, see also 3.1.5. This fact appears in an even odder light when considering that the first international thematic reference on autonomy in weapons systems was expressed by UN Special Rapporteur on Rapporteur on extrajudicial, summary or arbitrary executions, Christoph Heyns, in UN doc. A/HRC/23/47, § 38, for the Office of the High Commissioner for Human Rights.

[54] Art. 48, 49 51 (2) and 52 (2) Protocol Additional to the Geneva Conventions of August 12, 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), June 8, 1977.

proportionality[55] and precaution.[56] [57] Moreover, the nature of autonomy in a weapons system means that the lines of responsibility for an attack by a LAWS may not always be clear. Therefore, LAWS also challenge the legal concept of accountability.[58]

Recently, LAWS have also been discussed in the light of International Human Rights Law (IHRL), whose benchmark for the legal use of force is higher than under IHL.[59] However, the emphasis on IHRL falls behind the strong focus on IHL with the forum of the UN CCW.

5. The debate at the United Nations Convention on Certain Conventional Weapons (UN CCW)

LAWS have been taken up as an issue by the international arms control community in the framework of the UN CCW in 2014.[60] After a series of annual informal discussions, a Group of Governmental Experts (GGE) has debated on the subject matter for the first time as a formal meeting during a 5-day-gathering in the CCW framework in Geneva in November 2017.

The main points of discussion of the GGE were LAWS's potential legality under IHL, questions of accountability and responsibility for the use of LAWS during armed conflict, potential (working) definitions of LAWS, as well as the need for emerging norms, since LAWS highly challenge both existing IHL as well as normative principles.

[55] Art. 51 (5) (b) and Art. 57 Protocol I.

[56] Art. 57 (1) Protocol I.

[57] Brehm, Maya, 2017, Defending the boundary: Constraints and requirements on the use of autonomous weapons systems under international humanitarian and human rights law, Geneva Academy of International Humanitarian Law and Human Rights, 22; see also Bolton, M. 'From Minefields to Minespace: An Archeology of the Changing Architecture of Autonomous Killing in US Army Field Manuals on Landmines, Booby Traps and IEDs', 46 Political Geography (2015) 41–53.

[58] See e.g. Davison, Neil, 2017, A legal perspective: Autonomous weapon systems under international humanitarian law, UNODA Occasional Papers No. 30, New York: United Nations, 12, 16.

[59] Brehm, 2017; Heyns, Christof, 2016, Human Rights and the use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement, Human Rights Quarterly 38, 350-378; Heyns, Christof, 2014, Autonomous Weapons Systems and Human Rights Law, Presentation made at the informal expert meeting organized by the state parties to the Convention on Certain Conventional Weapons, May 13-14, 2017, Geneva, Switzerland.

[60] CCW/MSP/2014/3.

However, this first GGE on LAWS brought no agreement on a political declaration and also no path toward a new regulatory international treaty. The only common denominator was the general will of states to continue conversations in 2018.[61]

The UN CCW's debate bears at least five severe challenges to a comprehensive understanding of the risks of LAWS and AT.

(1) To date, states have neither agreed on a definition of LAWS nor of the concept of autonomy, nor on the fact whether increasingly autonomous weapons systems or precursor technologies already exist. Moreover, national as well as international policy debates on LAWS have lacked precise terminology.[62]

Bearing in mind the above-described thoughts on the technological concept of 'autonomy', this is no surprise. However, it is claimed that definitions will most likely play a key role in the international deliberation on the issue of LAWS.[63] This is true because, for one thing, in order to comprehensively discuss over a topic, it is crucial to base the debate on a common understanding of the issue. On the other hand, there exists a not negligible striving of some states and NGOs to ban LAWS.[64]

However, since AT and the concept of 'autonomy' for technological artefacts may be a proxy term for an ongoing trend in human's technological endeavours to give away control to technological agents and thereby relinquishing human responsibility for outcomes of autonomous systems, a premature agreement on a definition of 'autonomy' in weapons systems by the GGE on LAWS would most probably hide this trend. Therefore, instead of pressing for a definition of LAWS and 'autonomy' within the GGE, it would be advisable to locate these challenges within a bigger picture of the general relationship between humans and technology, and focus on the question whether we want to continue to regard technology as a controllable tool. In this sense, the GGE framework could be deemed as unfitting. Surely, principles for responsible AI research are both a first reflection of this underlying and ongoing paradigm change,

[61] CCW/GGE.1/2017/CRP.1, 4, 5.

[62] Ibid., 13. See above on Autonomous Technology.

[63] Nakamitsu, Izumi, 2017, Foreword to the Perspectives on Lethal Autonomous Weapons Systems, UNODA Occasional Papers No. 30, New York: United Nations, V.

[64] See the Campaign to Stop Killer Robots, https://www.stopkillerrobots.org/ (accessed on February 15, 2018). Currently, 22 states are backing this position: Algeria, Argentina, Bolivia, Brazil, Chile, Costa Rica, Cuba, Ecuador, Egypt, Ghana, Guatemala, Holy See, Iraq, Mexico, Nicaragua, Pakistan, Panama, Peru, State of Palestine, Uganda, Venezuela, Zimbabwe, Campaign to Stop Killer Robots, 2017, Country Views on Killer Robots, November 16, 2017.

as well as a first step in the direction of responsibly addressing the seriousness of this risk. A list of existing principles is found in the ANNEX of this paper.

(2) States are generally unwilling to share information on their capacity to develop LAWS. However, in order to gain a better understanding of the lessons learned from already existing weapons with certain levels of autonomy, the sharing of information is vital.[65]

(3) The GGE's mandate comprises the discussion of '[...] emerging technologies in the area of lethal autonomous weapons systems (LAWS) in the context of the objectives and the purposes of the convention [...]'.[66] However, the misuse of technology, e.g., by non-state actors, does not fall within the scope of this mandate.[67] Certainly, though, a holistic analysis and discussion of the peace and security implications of AT and new technologies requires the international community to address also the use of such by non-state actors.[68]

(4) LAWS represent a new category of weapons, in that their novelty lies in a formless technological capacity of recognizing patterns from a continuous inflow of data. The difference between a currently existing remotely controlled drone and a 'fully autonomous' drone does not lie in the casing, but in the fact that the latter is controlled by a software with autonomous capacities. The UN CCW, established in 1983, seeks to prohibit the use of certain conventional weapons. Its protocols currently prohibit

[65] ICRC, 2016, Autonomous weapons systems: Profound implications for future warfare, May 6, 2016, available at: https://www.icrc.org/en/document/autonomous-weapons-systems-profound-implications-future-warfare (accessed on February 4, 2018).

[66] CCW/CONF.V/10,10.

[67] Ambassador Amandeep Singh, 2017 GGE on LAWS, Geneva, November 13-17, 2017, Plenary Session of November 14, 2017.

[68] See e.g., the attack on Russian military facilities by a swarm of more than a dozen autonomous drones. Russia accused Turkish-backed rebel forces to be behind the attack. See e.g. Satherley, Dan, 2018, Wooden drone swarm attacks Russian forces in Syria, Newshub.com, available at: http://www.newshub.co.nz/home/world/2018/01/wooden-drone-swarm-attacks-russian-forces-in-syria.html (accessed on February 4, 2018); Embury-Dennis, Tom, 2018, Russia says mysterious armed drones are attacking its military base in Syria – and they don't know who's sending them, January 10, 2018, Independent.co.uk, available at: https://www.independent.co.uk/news/world/middle-east/russia-military-bases-drones-syria-armed-attacks-tartus-uavs-latakia-a8151066.html (accessed on February 4, 2018); Focus, 2018, Mit schwer bewaffnetem Drohnenschwarm: Terroristen greifen russischen Stützpunkt an, January 14, 2018, Focus.de, available at: https://www.focus.de/politik/ausland/drohnenschwarm-is-griff-russischen-stuetzpunkt-an-nun-naehrt-sich-ein-besorgniserregender-verdacht_id_8296804.html (accessed on January 15, 2017).

the use of weapons whose primary effect is to injure by fragments that, once within the human body, escape X-Ray detection, as well as the use of mines, booby-traps and incendiary weapons against civilians.[69] One may argue that the UN CCW's GGE on LAWS is not capable to the necessary degree to fully understand the technological complexity of current (not to mention future) AT.

(5) In addition, the CCW is a framework underpinned by IHL, which narrows the debate's focus on weapons and their use during *armed conflict*.[70] However, increasingly autonomous weapons systems can be and are used during peace time in law enforcement operations (e.g. crowd control, hostage situations), [71] where IHRL represents the legal benchmark.

Compared to IHL, IHRL is much more restrictive on the use of force. Military technology often finds its way into law enforcement. One may assume that once the advantages of increasingly autonomous systems have been proven in the military context, they might be considered for use during domestic law enforcement, although IHRL, regulating the latter, would prohibit their use.[72] Therefore, the CCW's/ GGE's approach could be criticized as not being legally comprehensive enough due to its limited focus on the use of a weapons during times of war.

[69] Protocol I to Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects as amended on 21 December 2001 (CCW) on Non-Detectable Fragments, Protocol II to the CCW on Prohibitions or Restrictions on the Use of Mines, Booby Traps and Other Devices, and Protocol III to the CCW on Prohibitions or Restrictions on the Use of Incendiary Weapons.

[70] Art. 1 and 2 CCW.

[71] See e.g. Opall-Rome, Barbara, 2016, Introducing: Israeli 12-Kilo Killer Robot, DefenseNews.com, May 8, 2016, available at: https://www.defensenews.com/global/mideast-africa/2016/05/08/introducing-israeli-12-kilo-killer-robot/ (accessed on February 4, 2018); Hurst, Luke, 2015, Indian Police Buy Pepper Spraying Drones To Control 'Unruly Mobs', Newsweek.com, April 7, 2015, available at: http://www.newsweek.com/pepper-spraying-drones-control-unruly-mobs-say-police-india-320189 (accessed on February 4, 2018). The 'Mozzy Wildlife Darting Copter' is promoted for wildlife capture, Desert Wolf: Leaders in Technology and Innovation, available at: http://www.desert-wolf.com/dw/products/unmanned-aerial-systems/mozzy-wildlife-darting-copter.html (accessed on February 4, 2018).

[72] Heyns, Christof, 2016, Human Rights and the use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement, Human Rights Quarterly 38, 350-378; Heyns, Christof, 2014, Autonomous Weapons Systems and Human Rights Law, Presentation made at the informal expert meeting organized by the state parties to the Convention on Certain Conventional Weapons, May 13-14, 2017, Geneva, Switzerland.

6. Further ways to weaponize AT

Furthermore, the CCW's discussion on LAWS has focused on conventional (physical/ robotic) systems which interact in a 3D reality with other machines or humans.[73] However, there exist further ways to weaponize AT.

(1) First, software with autonomous capacities can be used to act and interact entirely in the cyberspace. Those sometimes-called autonomous intelligent agents[74] are of tremendous military interest for similar reasons as for 'conventional' military operations: Autonomous intelligent agents acting in the cyberspace can support the decision-making process, they can identify an adversary's vulnerability and they can enable an ever-greater speed of response.[75] Hence, the use of autonomy for intangible cyber operations[76] (defensive or offensive) could be decisive and much more economic in current/future warfare.[77]

Five UN GGE discussions on cyber security have taken place since 2004/ 2005, and have confirmed that international law applies to the cyber space. Moreover, those GGEs have decided on a variety of confidence building measures (CBMs), and recommended norms for responsible State behavior in the domain.[78] Since both autonomous cyber weapons as well as LAWS are characterized by AT, both the GGE

[73] See also, UNIDIR, 2017, 1.

[74] Guarino, Alessandro, 2013, Autonomous Intelligent Agents in Cyber Offence, in: Podins, K., Stinissen, J., and Maybaum, M. (eds.), 5th International Conference on Cyber Conflict, NATO CCD COE Publications, 2013.

[75] Ibid, 4.

[76] There exists no standard terminology yet in this field.

[77] Meissner, Christopher, 2016, The Most Military Decisive Use of Autonomy You Won't See, DefenseOne, November 7, 2016, available at https://www.defenseone.com/ideas/2016/11/most-militarily-decisive-use-autonomy-you-wont-see-cyberspace-ops/132964/ (accessed on November 25, 2017). See e.g. the United States' cyberwarfare program MonsterMind. This software could constantly be on the lookout for traffic patterns indicating known or suspected cyberattacks. When it detected an attack, it would automatically block it from entering the country. This is regarded as a "kill" in cyber terminology. See e.g. Zetter, Kim, 2014, Meet Monstermind, The NSA Bot That Could Wage Cyberwar Autonomously, Wired, August 13, 2014, available at https://www.wired.com/2014/08/nsa-monstermind-cyberwarfare/ (accessed on November 28, 2017).

[78] UN Doc. A/70/174, 7-10.

on LAWS as well as those on cyber security share the thematic technological basis. Nevertheless, those international policy discussions have nearly no overlap.[79]

(2) Second, it is highly necessary to consider potentially malicious linkages of AT and other emerging technologies. Theoretically, it may be possible to create autonomous systems that control processes with the core aim of harming humans, e.g. the malicious use of biotechnology, 5G radiation,[80] or products of molecular nanotechnology.[81] [82] Current examples of such linkages do not exist. However, it is crucial to raise this concern early enough in order to trigger both research in this field as well as a *comprehensive* debate of peace and security implications of both AT and other emerging technologies.

In order to highlight the risks of a potential malicious linkage of different emerging technologies, recent technological breakthroughs in biotechnology shall figure as an example:

The term 'biotechnology' refers to '[...] any technological application that uses biological systems, living organisms, or derivatives thereof, to make or modify products or processes for specific use.'[83] One specific use of biotechnology is the creation of biological weapons. Biological weapons are designed to spread disease among people, animals and plants by introducing microorganisms and toxins, such as bacteria and viruses.

Using so-called DNA synthesis, which enables the artificial creation of DNA molecules, it may soon be possible to synthesize any virus whose DNA sequence is known.[84] Similarly, it is possible to insert small bacterial DNA fragments into another bacteria's

[79] For a good discussion on the questions of interaction between the GGEs on LAWS and cyber space, see UNIDIR, 2017.

[80] For health risks of 5G radiation, see e.g. Puzzanghera, Jim, 2016, Is 5G technology dangerous? Early data shows a slight increase of tumors in male rats exposed to cellphone radiation, Los Angeles Times, August 8, 2016, available at: http://www.latimes.com/business/la-fi-cellphone-5g-health-20160808-snap-story.html (accessed on February 15, 2018).

[81] Helbing, 2018.

[82] For dangers of molecular nanotechnology and molecular manufacturing, see e.g. Dangers of Molecular Manufacturing, Center for Responsible Nanotechnology, http://www.crnano.org/dangers.htm (accessed on February 15, 2018).

[83] Art. 2, Convention on Biological Diversity, of Rio de Janeiro of June 5, 1992.

[84] Hessel, A., Goodman, M., Kotler, S., 2012, Hacking the President's DNA, The Atlantic, available at http://www.theatlantic.com/magazine/archive/2012/11/hacking-the-presidents-dna/309147/?single_page=true (accessed on February 6, 2017).

DNA in order to increase its virulence, which would create a so-called 'binary biological weapon'.[85] Moreover, biotechnology could be used to manipulate cellular mechanisms to cause a disease. An agent could, e.g., be designed to induce cells to multiply uncontrollably, as in cancer, or induce programmed cell death (apotosis).[86] Further, in coming years it might be possible to design a pathogenic agent that targets a specific person's genome. When spread through a population that generally shows no or only minimal symptoms, it could nevertheless be fatal for the targeted person.[87]

Recently, scientists have been able to transform the four DNA nucleotid's letters into binary code, which now makes genetic engineering a matter of electronic manipulation and decreases the technique's cost.[88] Moreover, as of today, the European Nucleotide Archive of the European Bioinformatics Institute published sequences of 17075 genomes, including the genomes of 3316 bacteria and 4026 viruses.[89]

It is argued that biowarfare using genetically engineered pathogens can be considered as a potential revolution of military affairs.[90] Moreover, the exponential increase in computational power, the growing accessibility of genetic information and biological tools for the general public as well as the lack of governmental regulations also raise concerns about the non-state use of biowarfare.[91]

It is potentially possible to link AT and bioweapons, in that pathogens could be spread by autonomous systems.[92]

[85] Ainscough, M., 2002, Next Generation Bioweapons: Genetic Engineering and Biowarfare, available at: http://www.au.af.mil/au/awc/awcgate/cpc-pubs/biostorm/ainscough.pdf (accessed on February 6, 2018), 256.

[86] Ibid., 273.

[87] Hessel et al, 2012.

[88] Ibid.

[89] European Bioinformatics Institute, Access to Completed Genomes, https://www.ebi.ac.uk/genomes/index.html (accessed on February 6, 2018).

[90] Ainscough, M., 2002.

[91] See e.g. Kay, D., 2003, Genetically Engineered Bioweapons, available at: https://www.aaas.org/sites/default/files/migrate/uploads/ch17.pdf, (accessed on February 6, 2018); Breakingnews.ie, 2011, Advances in Genetics Could Create Deadly Biological Weapons, Clinton Warns, July 7, 2011, available at: http://www.breakingnews.ie/world/advances-in-genetics-could-create-deadly-biological-weapons-clinton-warns-531347.html (accessed on February 6, 2018).

[92] Helbing, 2018.

(1) Moreover, based on the above-mentioned understanding of ‘autonomy’ for artefacts as any *result of a technological process for which the human cannot or does not want to bear responsibility,* any intentionally harmful use of a technology whose causes for harm cannot be traced back to a human ‘trigger’ may be deemed an autonomous weapon.[93]

7. Peace-time threats of not-weaponized AT

The risks of AT for society are not limited to its weaponized use during an armed conflict. AT also bears risks for global society during peace-time, when it is not weaponized.

8. Mass disinformation generated by intelligent technology

Both fake news (deliberate misinformation via traditional or online media with the intent to mislead the readers) and internet trolls (the posting of erroneous, extraneous and off-topic messages in order to manipulate public opinion) could potentially be generated by autonomous intelligent agents, which could lead to mass disinformation guided by AT. Not only news portals that deliberately and automatically spread fake information, but also social bots on twitter have an immense potential for mass manipulation. Moreover, bots that deceive us are currently already more numerous than those that tell us the truth, and they hardly cost anything.[94]

In addition to *general* mass manipulation through widely spread disinformation by bots, research on AI makes it possible to generate *individualized* information.[95] In this case, people do not share a common reference point for information anymore. The

[93] Ibid.

[94] Laukenmann, Joachim, Der Nutzen von Lügenbots überwiegt: Interview mit Wirtschaftsinformatiker Oliver Bendel, #12 – Die Story des Tages, available at: https://mobile2.12app.ch/articles/29735653 (accessed on February 15, 2018).

[95] Cambridge Analytica has made lucrative use of those technological developments, see e.g. Hall, Jessica, 2017, Meet the weaponized propaganda that knows you better than yourself, Extremetech.com, March 1, 2017, accessible at: https://www.extremetech.com/extreme/245014-meet-sneaky-facebook-powered-propaganda-ai-might-just-know-better-know (accessed on February 15, 2018).

boarders between reality and artificial creation with regards to knowledge through individual research would blur.

Further, AI research is able to create so-called 'generative adversarial networks' (GAN) that can currently generate fake images and videos whose quality is such that humans are incapable of telling that they are not real shots.[96] Moreover, it is said that GANs could soon generate speech, language and behavior.[97]

With Adobe's application 'Project Voco' it is nowadays also possible to rapidly alter an existing voice recording to include words and phrases that the original speaker has never said.[98] One may assume that an altering of a recording by a machine or software instead of a human may soon be possible too.

When real videos, images and voice recordings become indistinguishable from fake ones, fake news will become even more prevalent, and video, image, and voice evidence could become inadmissible in court.[99]

9. Autonomously generated profiles

Computerized pattern and correlation recognition in order to identify and represent people, for example during criminal investigations, could be performed by AT. The detection and capture of potential (pre-emptive profiling) and actual criminals (e.g.) could be outsourced to increasingly autonomous machine calculation based on Big Data – uncontrollable for humans. Already today, deep learning software allow for ever-more perfected facial recognition. Facial recognition technology is a computer

[96] See e.g., Leary, Kyree, 2017, An AI that makes fake videos may facilitate the end of reality as we know it, Futurism, December 8, 2017, available at: https://futurism.com/ai-makes-fake-videos-facilitate-end-reality-know-it/ (accessed on February 15, 2018).

[97] Karras, T., Aila, T., Laine, S., and Lehtinen, J., 2018, Progressive Growing of GANs for Improved Quality, Stability, and Variation, NVidia Rsearch, submitted to ICLR 2018, available at: http://research.nvidia.com/sites/default/files/publications/karras2017gan-paper-v2.pdf (accessed on February 3, 2018); Future of Life Institute, 2018, Podcast: Top AI Breakthroughs and Challenges of 2017 with Richard Mallah and Chelsea Finn, January 31, 2018, available at: https://futureoflife.org/2018/01/31/podcast-top-ai-breakthroughs-and-challenges-of-2017-with-richard-mallah-and-chelsea-finn/ (accessed on February 2, 2018).

[98] BBC News, 2016, Adobe Voco 'Photoshop-for-voice' causes concern, November 7, 2016, BBC News Technology, available at: http://www.bbc.com/news/technology-37899902 (accessed on February 15, 2018).

[99] Leary, 2017.

application capable of identifying and verifying a person from a digital image or video. It is currently installed in public surveillance cameras, i.a., in Russia and China and used in order to continuously track potential criminals or public dissidents.[100]

Through increasingly autonomous criminal profiling the border between a criminal and a legally innocent person would be drawn exclusively by an algorithm, and vulnerable to incorrect data due to bad sensor-technologies, incompleteness or noise. Furthermore, categorizing potential criminals based on computational inferences somehow turns the presumption of innocence upside down, assuming a general potential for criminal conduct.[101]

Often, AI systems are claimed to be more 'objective' in their 'behavior' than a human, because they are not influenced by human feelings and prejudices. However, as 'intelligent' software and machines need to be 'fed' by a huge amount of data in order to 'learn' (a trait that we deem 'intelligent'), there exists the risk that they learn human prejudices from biased data. And so-called machine biases constitute a danger for AI-controlled or autonomous systems that some experts regard as far more acute than LAWS.[102] Based on the data a bot is fed by in order to learn, it could learn, e.g., to discriminate people of color or minorities, or gain a strict political attitude.[103]

[100] See e.g. Chin, Josh, and Lin, Lisa, 2017, China's All-Seeing Surveillance State Is Reading Its Citizen's Faces, The Wall Street Journal, June 26, 2017, available at https://www.wsj.com/articles/the-all-seeing-surveillance-state-feared-in-the-west-is-a-reality-in-china-1498493020 (accessed on November 27, 2017); Fischer, Sophie-Charlotte, 2018, Künstliche Intelligenz: Chinas Hightech-Ambitionen, CSS Analysen zur Sicherheitspolitik 220, Zurich: CSS ETH Zurich, 4; Mezzofiore, Gianluca, 2017, Moscow's facial recognition CCTV network is the biggest example of surveillance society yet, Mashable, September 28, 2017, available at http://mashable.com/2017/09/28/moscow-facial-recognition-cctv-network-big-brother/#kF19SB72r8qA (accessed on November 27, 2017). See also the new Israeli business 'Faception', which provides real-time facial personality analytics and personal profiling also from offline datasets, https://www.faception.com/our-technology (accessed on February 15, 2018).

[101] Hildebrandt, Mireille, 2015, Smart Technologies and the End(s) of Law, Novel Entanglements of Law and Technology, Elgar Publishing, 97.

[102] Knight, Will, 2017b, Forget Killer Robots – Bias is the real AI danger, MIT Technology Review, October 3, 2017, available at: https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/ (accessed on February 15, 2018).

[103] See e.g. 'Tay', a bot created by Microsoft who should learn from humans and turned into a Nazi within 24 hours, in: Steiner, Anna, 2016, Zum Nazi und Sexisten in 24 Stunden, Frankfurter Allgemeine, March 24, 2016, available at: http://www.faz.net/aktuell/wirtschaft/netzwirtschaft/microsofts-bot-tay-wird-durch-nutzer-zum-nazi-und-sexist-14144019.html (accessed on February 15, 2018); or Google's search algorithm that spread false information with a right wing bias, in: Solon, Olivia, and Levin, Sam, 2016, How Google's search algorithm

10. Autonomous technology in light of emerging resource-scarcity on our planet

The current global social, economic (including financial and monetary) and environmental trends render the planet's resources scarce. This constitutes a risk to humanity and make our present global human coexistence potentially unsustainable. Hence, some experts ask the question: In an increasingly unsustainable society in critical times, what kind of citizens should be protected, and whose lives could be sacrificed? Should the worth of people's lives be weighed according to a certain benchmark, so that we can more easily decide who could stay alive? And does a human being have the guts to decide – or should we outsource this decision to autonomous software?

For example, autonomous intelligent agents could be integrated into health insurance systems e.g., and feeding from their patients' data, they could determine who receives a potential treatment and who does not. This may yet be a dystopian idea. However, the idea of a rating system for citizens is already tested in China with the so-called Citizen Score Card, which represents the value of an individual citizen from a governmental perspective. [104] A rating system like this could potentially become a reference point of informing decisions that aim at limiting population figures. [105]

Therefore, the emergence of AT can force us even more to evaluate our current economic, social and environmental systems and trends in order not to put society at

spreads false information with a right wing bias, The Guardian, December 16, 2016, available at: https://www.theguardian.com/technology/2016/dec/16/google-autocomplete-rightwing-bias-algorithm-political-propaganda (accessed on February 15, 2018).

[104] Storm, Darlene, 2015, ACLU: Orwellian Citizen Score, China's credit score system, is a warning for Americans, Computerworld, October 7, 2015, available at https://www.computerworld.com/article/2990203/security/aclu-orwellian-citizen-score-chinas-credit-score-system-is-a-warning-for-americans.html (accessed on November 25, 2017); see also India's mandatory biometric ID system 'Aadhar': Pahwa, Nikhil, 2017, How not to screw up your national ID, Medianama, November 21, 2017, available at https://www.medianama.com/2017/11/223-how-not-to-screw-up-your-national-id-india-aadhaar/ (accessed on November 27, 2017) and the British 'Karma Police', a GCHQ program by the British government that creates personality profiles of British citizens, Brandom, Russel, 2015, British 'Karma Police' program carries out mass surveillance of the web, TheVerge.com, September 25, 2015, available at: https://www.theverge.com/2015/9/25/9397119/gchq-karma-police-web-surveillance (accessed on February 7, 2018).

[105] Helbing, Dirk, Nagler, Jan, and Van den Hoven, Jeroen, 2017, Ethics for Times of Crisis: How not to use autonomous systems in an unsustainable world, available at https://www.

risk of being kept in quantitative borders set by algorithms and based on utilitarian calculations.

11. Arguments for shaping an international interdisciplinary debate

11.1 The polity of the cyberspace

Code is the regulator of the cyberspace, the way a constitution can be regarded as a regulator of society. Code enables the exchange of data among networks, which is currently still generally neutral regarding the content of the data and ignorant about the user. This feature of codes makes regulating behavior in the cyberspace difficult. However, code is not fixed, but the architecture of the cyberspace can be changed by the ones who code. The fact that it is hard to know who someone is in the Net and what the character of the content is that is delivered, can be changed. New architecture can facilitate identification and rate data content. This architecture can either be privacy-enhancing or not. This depends on the incentives that those who set it up are facing.

In other words, there exists a choice whether to influence the 'regulability' of the cyberspace as well as a choice on how this regulation should look like. Moreover, the way a constitution represents the normative values of a society through codifying them by law, code can be said to reflect a choice of values that should guide actions and inactions in the cyberspace. If code represents the law of cyberspace, and computer software potentially interferes with citizens' privacy and maybe physical integrity (LAWS), should their use be restricted and regulated by a democratic process?[106]

This argument for a value-sensitive design of any software code that does or could interfere with citizen's privacy and physical integrity approved by a democratic

researchgate.net/publication/320740872_Ethics_for_Times_of_Crisis_How_not_to_use_autonomous_systems_in_an_unsustainable_world (accessed on November 25, 2017).

[106] Lessing, Lawrence, 2000, Code Is Law, On Liberty in Cyberspace, Harvard Magazine, January-February 2000, available at http://socialmachines.media.mit.edu/wp-content/uploads/sites/27/2015/03/Code-is-Law-Harvard-Magazine-Jan-Feb-2000.pdf (accessed on November 25, 2017); see also Van den Hoven, Jeroen, Vermaas, Home Pieter, and Van de Poel, Ibo (Eds.), 2015, Handbooks of Ethics, Values and Technological Design: Sources, Theory, Values and Application Domains, Springer.

political process would, as a first step, require a constant and very strong interaction between technological experts and both national and international policy-makers. Only thereby could the current policy discussions on technology lose their theoretical aspect and become more practical, which is crucial in order to potentially introduce the necessary aspects into a legislative process. Creating a fixed national and international policy-technology interface would require an architectural change of national and international political institutions, similar to the United Arab Emirates new state minister for AI.[107]

Furthermore, source codes of AT and AI-controlled systems needed to be open source in order to be accessible for a political discussion and potential introduction into a legislative process. This condition will require deeply considered answers on the question of property rights of source codes of autonomous systems.

11.2 The subtle linguistics and the human-machine analogy

The international debate on AT and LAWS contains the unexamined assumption that humans and artificially intelligent systems are different only to a degree, and that human qualities can be reproduced in a machine. This underlying belief is the reason why the international debate uses anthropomorphic language – machine 'decision-making', machine 'learning', let alone machine 'intelligence' or 'autonomy' – to describe current technological artefacts.

On this subject it is crucial to highlight two points: First, the human-machine analogy grew out of the initial wish and claim of AI research to understand the human brain by modelling it. However, this analogy still has a mere hypothetical character. Science could not yet fully reveal what happens in the human brain when, e.g., a decision is taken,[108] or how and if 'consciousness' can be linked to a physical process.[109]

[107] Galeon, Dom, 2017, An Inside Look at the First Nation With a State Minister for Artificial Intelligence, Futurism, December 11, 2017, available at: https://futurism.com/uae-minister-artificial-intelligence/ (accessed on February 16, 2018).

[108] See e.g. Holdgraf, Chris, 2015, Decisions in the Brain, Berkeley Neuroscience News, June 15, 2015, available at: http://neuroscience.berkeley.edu/decisions-in-the-brain/ (accessed on December 9, 2017); Neue Zürcher Zeitung, Schaltzentrale Hirn,

[109] Kesser, Eduard, 2017, Das leer Gehirn, Neue Zürcher Zeitung, November 17, 2017, available at: https://www.nzz.ch/wissenschaft/das-leere-gehirn-ld.1329199 (accessed on February 16, 2018).

And second, a software is usually named by its purpose, and not by its structure. If the purpose of, e.g., an 'autonomous' software is to supplant the human in an area where the latter used to take a 'human decision' in no way implies that the software 'takes a decision' as well.[110] Hence, by comparing humans and machines or software at a common reference point (e.g. capacity to 'decide', 'learn' or 'behave morally') we may risk falling into a linguistic trap and prematurely overestimate technological artifacts and underestimate human capacities, let alone human language.

Language frames the way we think, understand and compare. Using the same language for machines and software as for humans could lead us to make potentially false comparisons – 'machines decide *better* than humans'.[111] Keeping in mind also the above-discussed risk for terminological confusion through the term 'autonomy' or 'intelligence', the question whether we need a new language for technological artefacts may be legitimate.

11.3 A moral argument for a sustainable environment

We are on the threshold of a paradigm shift where the human being will not be the only existing 'intelligent system' on the planet with the capacity for autonomous action anymore. Depending on the features that are encoded in increasingly autonomous systems and the existing risks of unpredictable outcomes[112] and vulnerabilities to hacking (e.g.), these systems may challenge the structure of current human society and might even become a risk for humanity as a species. Some experts also argue that organic human life is merely a short precursor in the evolutionary history of intelligent 'life' in the universe, which might soon be represented by inorganic machines with a far more powerful intellect than humans.[113]

Some already prepare for a potential emergence of general AI through the enhancement of human brain power through AI itself: Elon Musk's recently launched

[110] See McDermott, Drew, 1981, Artificial Intelligence meets Natural Stupidity, in: Haugeland, John (ed.), Mind Design – Philosophy, Psychology, Artificial Intelligence, Cambridge MA: The MIT Press.

[111] Müller, Jürg, 2017, 'Oft entscheiden Menschen sehr schlecht', Neue Zürcher Zeitung, November 1, 2017, available at: https://www.nzz.ch/wirtschaft/oft-entscheiden-menschen-sehr-schlecht-ld.1325428 (accessed on February 16, 2018).

[112] Knight, Will, 2017a.

[113] Rees, Martin, 2015, What do you think about machines that can think?, Edge.org, available at: https://www.edge.org/response-detail/26160 (accessed February 16, 2018).

company ‘Neuralink’ is exploring ‘neural lace’ technology – the implanting of tiny electrodes into the human brain to give us direct computing capabilities.[114] He argues that a ‘[...] merger of biological intelligence and machine intelligence [...]’ would be necessary for humans to stay economically valuable in a future of general AI.[115] Another way to keep up with AI and AT systems in a potential future world could also be paved by a genetic upgrade of humans through gene editing, which can nowadays already be used to alter the DNA of embryos.[116] In other words, research is focused on technology that would not only help us to *do*, but that has the potential to help us *be*.[117]

A recent survey with the aim at clarifying expert opinions on the possibility and risks of human-like machine intelligence, based on 550 AI expert opinions, revealed a view among experts that AI systems will probably (over 50%) reach overall human ability by 2040-2050, and very likely (with 90% probability) by 2075. From reaching human-level-intelligence, experts assume that artificial superintelligence will be reached within 30 years after with a probability of 75%. Moreover, the respondents say that the probability that this development may be ‘bad’ or ‘extremely bad’ for humanity is 31%.[118]

In this light, some experts claim that there exists a moral duty to pre-emptively decide not to create an invasive artificial species of autonomous agents that could endanger the lives of human beings on the planet.[119]

[114] See https://www.neuralink.com/ (accessed on February 16, 2018).

[115] The Guardian, 2017, Elon Musk wants to connect brains to computers with new company, March 28, 2017, available at: https://www.theguardian.com/technology/2017/mar/28/elon-musk-merge-brains-computers-neuralink (accessed on February 16, 2018).

[116] Helbing, 2018; see also Levitt, Mairi, 2015, Would you edit your unborn child’s genes so they were successful?, The Guardian, November 3, 2015, available at: https://www.theguardian.com/sustainable-business/2015/nov/03/designer-baby-pgd-would-you-edit-your-unborn-child-genes-more-successful (accessed on February 16, 2018); for a list of gene editing companies, see e.g. https://www.nanalyze.com/2015/04/7-gene-editing-companies-investors-should-watch/ (accessed on February 16, 2018).

[117] Prabhakar, Arati, 2017, The merging of humans and machines is happening now, Wired, January 27, 2017, available at: http://www.wired.co.uk/article/darpa-arati-prabhakar-humans-machines (accessed on February 17, 2018).

[118] Müller, Vincent C., and Bostrom, Nick, 2016.

[119] Helbing, Dirk, 2017, Open Discussion on Presentation on Lethal Autonomous Weapons Systems, November 13, 2017, ETH Zurich, Switzerland; see also Cellan-Jones, Rory, 2014, Stephen Hawkings warns artificial intelligence could end mankind, BBC Online, December 2, 2014, available at http://www.bbc.com/news/technology-30290540 (accessed on November

12. Conclusion

The international community should not get lost in attempts to define the term 'autonomy' for technological artefacts. Years of research and four years of discussions of LAWS within the UN CCW have not lead to terminological clarification, but opinions on the scope and content of the term 'autonomy' or AT have become more diverse.

If the term 'autonomy' for technological artefacts was defined to include a set of clearly delineated characteristics (e.g. 'learning', 'creating or pursuing of a goal', 'independent of human control, operation or intervention'), future technological research might reveal further potential characteristics which then would be excluded from this definition.

A fixed definition of 'autonomy' for technological artefacts, yet, could lead to a clear definition of LAWS within the GGE. On the one hand, this could encourage a potential outcome of the UN discussions (e.g. Code of Conduct or norms for responsible State behaviour). On the other hand, again, new and yet unknown technological developments interesting for military use might be beyond the scope of this definition of LAWS. Hence, the pressure of defining 'autonomy' in order to proceed with the GGE debate would most possibly lead to a definition that reflects the current and maybe also a conceivable future's technological potentials. However, the exponential pace with which AI research advances must alert us to yet unknown potentials and risks.

Consequently, the endeavour to minimize risks of AI and AT must not focus on definitional questions regarding LAWS but concentrate on binding principles for responsible AI research. This alternative track would take into account the fact that 'autonomy' for technological artefacts, e.g. LAWS, can and should be regarded as a proxy term for the loss of human control and responsibility for outcomes of technological processes. Principles guiding AI research could require programmers and engineers only to develop technological artefacts whose outcomes will stay controllable for humans, and for which the latter would, hence, always bear responsibility. Initiatives of professional organizations as well as representatives of the private sector have led to several lists of principles for responsible/ ethical research on AI and autonomy (ANNEX). It would be advisable to bundle those principles and create an international body that would supervise compliance.

27, 2017).

Consequently, an open discussion on whether or not humanity accepts the fact that technology is already crossing a threshold after which its creations might not be controllable for humans anymore, must be encouraged. Luckily, the UN CCW's debate on LAWS has brought this crucial moment into the public spotlight. Yet, for a purposeful discussion of this broader question, both the architecture of CCW forum as well as its limited mandate of LAWS are unsuitable.

As this paper has attempted to show, AT is much more than just its representation in LAWS. If we trustfully want to look into the future of humanity, it is a prerequisite to gain a holistic understanding of all the peace and security implications of AT and emerging technologies.

Autonomous cyber weapons and autonomous weapons during law enforcement operations are excluded from the CCW discussion, yet they reflect the seriousness of the risk of weaponized AT to the same as or even to a higher degree than LAWS. Hence, if the international community shall prove its serious commitment to the issue of emerging technologies, autonomous cyber weapons and autonomous weapons during law enforcement must be included in international discussions immediately.

Second, a holistic understanding of all the peace and security implications of AT must include peace-time threats of not-weaponized AT, such as mass dis- and misinformation as well as autonomous profiling and citizen control. A fixed body of experts at the UN level should take on committed discussions of peace and security implications of not-weaponized AT during peace-time.

Third, a holistic understanding of all peace and security implications of emerging technologies is necessary. This includes, i.a., AI, biotechnology, 5G radiation, and molecular nanotechnology. This paper had a limited focus on AT. However, threats for humanity stem from many more technological endeavours, whose risks are yet to be analysed. A fixed body of experts at the UN level should take on discussions of the peace and security implications of all emerging technologies.

Moreover, this paper has pointed out that the international debate on LAWS contains the unexamined assumption of a human-machine analogy. However, the view that human qualities can be reproduced in a machine should not be accepted unconditionally. As long as science cannot fully reveal the physical representation of human intelligence, consciousness, and decision-making processes in the human brain, self-protection should force us to acknowledge human distinctiveness. The fact that 'being human' is unquantifiable for science must not mean that human distinctiveness does not exist. We have a duty to preserve an assumption of this distinctiveness by limiting potential technologies that could challenge it or even wipe it out.

One way of preserving an understanding of the distinctiveness of ‘being human’ is by a careful use of language. Software or machine ‘autonomy’, ‘intelligence’ or ‘agency’ are terms that are very problematic in this sense. A premature heroization of technology could be prevented by introducing distinct terms. By using a term such as, e.g., ‘artefact with *cognitive functions*’ instead of ‘intelligent agent’, the fact that the machine is performing a *function* would be highlighted. This would set a clear boundary to being ‘human and intelligent’, as humans are never only performing a function, but are always an end in themselves. Moreover, the term ‘artefact’ would point out its objective character as opposed to ‘agent’.

In addition, this paper has challenged the view of the inevitability of AT and LAWS, which, unfortunately, reigns the minds of some commentators.[120] We argued that the use of any software that could potentially interfere with a citizen’s privacy or physical integrity could and should be regulated by a democratic process, in the same way as laws with the same quality are. This is a high demand. However, since it is possible that the future is far closer as we might think, it is highly important to start thinking and planning for this future today.

An introduction of software codes into a legislative process would require a creation of a constant policy-technology interface through, e.g., fixed state departments for technology/ AI. A constant dialogue between tech experts and policy-makers through an institutional integration could limit the risk that both programmers of (potentially) harming AT and policy-makers palm off the responsibility of ‘immoral’ outcomes to each other. Further, such an idea would require source codes to be publicly accessible, for which deeply considered answers on the question of property rights of source codes of autonomous and other systems are a prerequisite.

Humanity is striding into a future where machines and software will have an unprecedented role in almost all aspects of our lives. Moreover, future technology may have an immense potential in order for humans to define what they want to

[120] See e.g. ‘Warfare will continue and autonomous robots will ultimately be deployed in its conduct’, Arkin, Ronald, 2009, Governing Lethal Behavior in Autonomous Robots, CRC Press, 29; or ‘Autonomous weapons systems are the next logical and seemingly inevitable step in the continuing evolution of military technologies’, Beard, Jack M., 2014, Autonomous Weapons and Human Responsibilities, Georgetown Journal of International Law 45, 617-681, 620.

become. If we want to wisely navigate through a future that we might share with artefacts with cognitive abilities, we need to discuss some serious questions on ‘autonomy’, ‘responsibility,’ ‘privacy’ and ‘identity’– and we have to do it now. This paper represents a small contribution to those profound challenges. More will be needed.

Based on this paper’s conclusions, ICT4Peace would welcome:

1. A creation of an UN level body for technology, with the tasks of ensuring responsible technological research and discussing peace and security implications of emerging technologies, i.a. AI and AT, biotechnology, 5G, molecular nanotechnology. This body would also set principles for responsible research in the above-mentioned scientific fields and ensure compliance. An adequate functioning of this body would make the UN CCW’s discussion on LAWS redundant. Hence, point (2) would be temporary.

2. An inclusion of autonomous cyber weapons and autonomous weapons during law enforcement into international discussions. The former could be integrated into the GGE on LAWS, and the latter could be taken up by the Human Rights Council.

3. A combined UN position of all the peace and security implications of emerging technologies.

4. A public questioning of the human-machine analogy, and a potential introduction of new terms, replacing ‘AI’ and ‘autonomy’. Examples are ‘artefact’ instead of ‘agent’ or ‘…with cognitive functions/ capabilities’ instead of ‘intelligent’.

5. A creation of a constant national policy-technology interface through, e.g., fixed state ministers for technology/ AI.

6. An engaged debate on property rights on source codes of AI and AT software.

7. An increased engagement of civil society, including the private sector and academia, on the questions of human control of and responsibility for technological outcomes.

8. A constant dialogue between tech experts and civil society. Therefore, technologists must learn to transfer their expert knowledge in a practical way. This could be enhanced if such practice was included in university curricula.

ANNEX: Existing guidelines on responsible AI, AT and Robotics research

This annex contains six lists of guidelines for ethical/ responsible AI and AT research. Since the research field of robotics is highly linked to the research on AI and AT, and many endeavors have already lead to lists of principles in robotics, the annex also includes four lists of guidelines for ethical/ responsible robotics research.

A. GUIDELINES ON RESPONSIBLE AI RESEARCH:

Future of Life Institute (FLI)

The FLI is a volunteer-run research and outreach organization in the Boston area that works to mitigate existential risks facing humanity, particularly existential risk from advanced artificial intelligence (AI). Its founders include MIT cosmologist Max Tegmark, Skype co-founder Jaan Tallinn, and its board of advisors includes cosmologist Stephen Hawking and entrepreneur Elon Musk.

https://futureoflife.org/ai-principles/

‘Asilomar AI Principles of 2017

Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.

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.'

Association for Computing Machinery (ACM)

ACM, the world's largest educational and scientific computing society, delivers resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources. The ACM Code of Ethics and Professional Conduct includes, i.a., four principles relating to ethical and responsible research. Due to its length, only those four are included in this annex.

https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct#CONTENTS

'ACM Code of Ethics and Professional Conduct Adopted by ACM Council 10/16/92. [...]

General Moral Imperatives

As an ACM member I will ....

1.1 Contribute to society and human well-being.

This principle concerning the quality of life of all people affirms an obligation to protect fundamental human rights and to respect the diversity of all cultures. An essential aim of computing professionals is to minimize negative consequences of computing systems, including threats to health and safety. When designing or implementing systems, computing professionals must attempt to ensure that the products of their efforts will be used in socially responsible ways, will meet social needs, and will avoid harmful effects to health and welfare.

In addition to a safe social environment, human well-being includes a safe natural environment. Therefore, computing professionals who design and develop systems must be alert to, and make others aware of, any potential damage to the local or global environment.

Avoid harm to others

"Harm" means injury or negative consequences, such as undesirable loss of information, loss of property, property damage, or unwanted environmental impacts. This principle prohibits use of computing technology in ways that result in harm to any of the following: users, the general public, employees, employers. Harmful actions include intentional destruction or modification of files and programs leading to serious loss of resources or unnecessary expenditure of human resources such as the time and effort required to purge systems of "computer viruses."

Well-intended actions, including those that accomplish assigned duties, may lead to harm unexpectedly. In such an event the responsible person or persons are obligated to undo or mitigate the negative consequences as much as possible. One way to avoid unintentional harm is to carefully consider potential impacts on all those affected by decisions made during design and implementation.

To minimize the possibility of indirectly harming others, computing professionals must minimize malfunctions by following generally accepted standards for system design and testing. Furthermore, it is often necessary to assess the social consequences of systems to project the likelihood of any serious harm to others. If system features are misrepresented to users, coworkers, or supervisors, the individual computing professional is responsible for any resulting injury.

In the work environment the computing professional has the additional obligation to report any signs of system dangers that might result in serious personal or social damage. If one's superiors do not act to curtail or mitigate such dangers, it may be necessary to "blow the whistle" to help correct the problem or reduce the risk. However, capricious or misguided reporting of violations can, itself, be harmful. Before reporting violations, all relevant aspects of the incident must be thoroughly assessed. In particular, the assessment of risk and responsibility must be credible. It is suggested that advice be sought from other computing professionals. See principle 2.5 regarding thorough evaluations.

[...]

1.7 Respect the privacy of others

Computing and communication technology enables the collection and exchange of personal information on a scale unprecedented in the history of civilization. Thus there is increased potential for violating the privacy of individuals and groups. It is the responsibility of professionals to maintain the privacy and integrity of data describing individuals. This includes taking precautions to ensure the accuracy of data, as well as protecting it from unauthorized access or accidental disclosure to inappropriate individuals. Furthermore, procedures must be established to allow individuals to review their records and correct inaccuracies.

This imperative implies that only the necessary amount of personal information be collected in a system, that retention and disposal periods for that information be clearly defined and enforced, and that personal information gathered for a specific purpose not be used for other purposes without consent of the individual(s). These principles apply to electronic communications, including electronic mail, and prohibit procedures that capture or monitor electronic user data, including messages, without the permission of users or bona fide authorization related to system operation and maintenance. User data observed during the normal duties of system operation and maintenance must be treated with strictest confidentiality, except in cases where it is evidence for the violation of law, organizational regulations, or this Code. In these cases, the nature or contents of that information must be disclosed only to proper authorities.

[...]

3.5 Articulate and support policies that protect the dignity of users and others affected by a computing system.

Designing or implementing systems that deliberately or inadvertently demean individuals or groups is ethically unacceptable. Computer professionals who are in decision making positions should verify that systems are designed and implemented to protect personal privacy and enhance personal dignity.

[...]’

Institute of Electric and Electronical Engineers (IEEE)

IEEE is the world’s largest technical professional organization dedicated to advancing technology for the benefit of humanity. IEEE and its members inspire a global community to innovate for a better tomorrow through its more than 420,000 members in over 160 countries, and its highly cited publications, conferences, technology standards, and professional and educational activities. IEEE is the trusted “voice” for engineering, computing, and technology information around the globe.

The IEEE created the **Global Initiative on Ethics of Autonomous and Intelligent Systems**, an incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies. It aims at ensuring that every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity. The latest version of the book can be downloaded here: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

IBM

https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/

'Purpose: The purpose of AI and cognitive systems developed and applied by the IBM company is to augment human intelligence. Our technology, products, services and policies will be designed to enhance and extend human capability, expertise and potential. Our position is based not only on principle but also on science. Cognitive systems will not realistically attain consciousness or independent agency. Rather, they will increasingly be embedded in the processes, systems, products and services by which business and society function—all of which will and should remain within human control.

Transparency: For cognitive systems to fulfill their world-changing potential, it is vital that people have confidence in their recommendations, judgments and uses. Therefore, the IBM company will make clear:

When and for what purposes AI is being applied in the cognitive solutions we develop and deploy.

The major sources of data and expertise that inform the insights of cognitive solutions, as well as the methods used to train those systems and solutions.

The principle that clients own their own business models and intellectual property and that they can use AI and cognitive systems to enhance the advantages they have built, often through years of experience. We will work with our clients to protect their data and insights, and will encourage our clients, partners and industry colleagues to adopt similar practices.

Skills: The economic and societal benefits of this new era will not be realized if the human side of the equation is not supported. This is uniquely important with cognitive technology, which augments human intelligence and expertise and works collaboratively with humans. Therefore, the IBM company will work to help students, workers and citizens acquire the skills and knowledge to engage safely, securely and effectively in a relationship with cognitive systems, and to perform the new kinds of work and jobs that will emerge in a cognitive economy.'

DeepMind

DeepMind has created the DeepMind Ethics & Society, a research unit that aims to explore the key ethical challenges facing the field of AI, through interdisciplinary work that brings together the technical insights of it DeepMind team and the diverse range of people who will be affected by it.

https://deepmind.com/applied/deepmind-ethics-society/principles/

'DeepMind Ethics & Society Principles

Social benefit: We believe AI should be developed in ways that serve the global social and environmental good, helping to build fairer and more equal societies. Our research will focus directly on ways in which AI can be used to improve people's lives, placing their rights and well-being at its very heart.

Rigorous and evidence-based: Our technical research has long conformed to the highest academic standards, and we're committed to maintaining these standards when studying the impact of AI on society. We will conduct intellectually rigorous, evidence-based research that explores the opportunities and challenges posed by these technologies. The academic tradition of peer review opens up research to critical feedback and is crucial for this kind of work.

Transparent and open: We will always be open about who we work with and what projects we fund. All of our research grants will be unrestricted and we will never attempt to influence or pre-determine the outcome of studies we commission. When we collaborate or co-publish with external researchers, we will disclose whether they have received funding from us. Any published academic papers produced by the Ethics & Society team will be made available through open access schemes.

Diverse and interdisciplinary: We will strive to involve the broadest possible range of voices in our work, bringing different disciplines together so as to include diverse viewpoints. We recognize that questions raised by AI extend well beyond the technical domain, and can only be answered if we make deliberate efforts to involve different sources of expertise and knowledge.

Collaborative and inclusive: We believe a technology that has the potential to impact all of society must be shaped by and accountable to all of society. We are therefore committed to supporting a range of public and academic dialogues about AI. By establishing ongoing collaboration between our researchers and the people affected by these new technologies, we seek to ensure that AI works for the benefit of all.'

Microsoft

https://www.microsoft.com/en-us/ai/our-approach-to-ai

'Microsoft AI Principles

Fairness: AI must maximize efficiencies without destroying dignity and guard against bias

Accountability: AI must have algorithmic accountability

Transparency: AI must be transparent

Ethics: AI must assist humanity and be designed for intelligent privacy'

B. GUIDELINES ON RESPONSIBLE ROBOTICS RESEARCH:

Engineering and Physical Science Research Council (EPSRC)

EPSRC is the main UK government agency for funding research and training in engineering and the physical sciences, investing more than £800 million a year in a broad range of subjects - from mathematics to materials science, and from information technology to structural engineering. Its mission is to promote and support, by any means, high quality basic, strategic and applied research and related postgraduate training in engineering and the physical sciences; to advance knowledge and technology (including the promotion and support of the exploitation of research outcomes), and provide trained scientists and engineers, which meet the needs of users and beneficiaries (including the chemical, communications, construction, electrical, electronic, energy, engineering, information technology, pharmaceutical, process and other industries).

https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

‘Principles of Robotics

Note: The rules are presented in a semi-legal version; a more loose, but easier to express, version that captures the sense for a non-specialist audience and a commentary of the issues being addressed and why the rule is important.

Principles for designers, builders and users of robots:

I.

Legal

Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.

General Audience

Robots should not be designed as weapons, except for national security reasons.

Commentary

Tools have more than one use. We allow guns to be designed which farmers use to kill pests and vermin but killing human beings with them (outside warfare) is clearly wrong. Knives can be used to spread butter or to stab people. In most societies, neither guns nor knives are banned but controls may be imposed if necessary (e.g. gun laws) to secure public safety. Robots also have multiple uses. Although a creative end-user could probably use any robot for violent ends, just as with a blunt instrument, we are saying that robots should never be designed solely or even principally, to be used as weapons with deadly or other offensive capability. This law, if adopted, limits

the commercial capacities of robots, but we view it as an essential principle for their acceptance as safe in civil society.

II.

Legal

Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.

General Audience

Robots should be designed and operated to comply with existing law, including privacy.

Commentary

We can make sure that robot actions are designed to obey the laws humans have made.

There are two important points here. First, of course no one is likely deliberately set out to build a robot which breaks the law. But designers are not lawyers and need to be reminded that building robots which do their tasks as well as possible will sometimes need to be balanced against protective laws and accepted human rights standards. Privacy is a particularly difficult issue, which is why it is mentioned. For example, a robot used in the care of a vulnerable individual may well be usefully designed to collect information about that person 24/7 and transmit it to hospitals for medical purposes. But the benefit of this must be balanced against that person's right to privacy and to control their own life e.g. refusing treatment. Data collected should only be kept for a limited time; again the law puts certain safeguards in place. Robot designers have to think about how laws like these can be respected during the design process (e.g. by providing off-switches).

Secondly, this law is designed to make it clear that robots are just tools, designed to achieve goals and desires that humans specify. Users and owners have responsibilities as well as designers and manufacturers. Sometimes it is up to designers to think ahead because robots may have the ability to learn and adapt their behaviour. But users may also make robots do things their designers did not foresee. Sometimes it is the owner's job to supervise the user (e.g. if a parent bought a robot to play with a child). But if a robot's actions do turn out to break the law, it will always be the responsibility, legal and moral, of one or more human beings, not of the robot (We consider how to find out who is responsible in law 5, below).

III.

Legal

Robots are products. They should be designed using processes which assure their safety and security.

General Audience

Robots are products: as with other products, they should be designed to be safe and secure.

Commentary

Robots are simply not people. They are pieces of technology their owners may certainly want to protect (just as we have alarms for our houses and cars, and security guards for our factories) but we will always value human safety over that of machines. Our principle aim here, was to make sure that the safety and security of robots in society would be assured, so that people can trust and have confidence in them.

This is not a new problem in technology. We already have rules and processes that guarantee that, e.g. household appliances and children's toys are safe to buy and use. There are well worked out existing consumer safety regimes to assure this: e.g. industry kite-marks, British and international standards, testing methodologies for software to make sure the bugs are out, etc. We are also aware that the public knows that software and computers can be "hacked" by outsiders, and processes also need to be developed to show that robots are secure as far as possible from such attacks. We think that such rules, standards and tests should be publicly adopted or developed for the robotics industry as soon as possible to assure the public that every safeguard has been taken before a robot is ever released to market. Such a process will also clarify for industry exactly what they have to do.

This still leaves a debate open about how far those who own or operate robots should be allowed to protect them from e.g. theft or vandalism, say by built-in taser shocks. The group chose to delete a phrase that had ensured the right of manufacturers or owners to include "self defence" capability into a robot. In other words we do not think a robot should ever be "armed" to protect itself. This actually goes further than existing law, where the general question would be whether the owner of the appliance had committed a criminal act like assault without reasonable excuse.

IV.

Legal

Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.

General Audience

Robots are manufactured artefacts: the illusion of emotions and intent should not be used to exploit vulnerable users.

Commentary

One of the great promises of robotics is that robot toys may give pleasure, comfort and even a form of companionship to people who are not able to care for pets, whether due to rules of their homes, physical capacity, time or money. However, once a user becomes attached to such a toy, it would be possible for manufacturers to claim the robot has needs or desires that could unfairly cost the owners or their families more money. The legal version of this rule was designed to say that although it is permissible and even sometimes desirable for a robot to sometimes give the impression of real intelligence, anyone who owns or interacts with a robot should be able to find out what it really is and perhaps what it was really manufactured to do. Robot intelligence is artificial, and we thought that the best way to protect consumers was to remind them of that by guaranteeing a way for them to “lift the curtain” (to use the metaphor from The Wizard of Oz).

This was the most difficult law to express clearly and we spent a great deal of time debating the phrasing used. Achieving it in practice will need still more thought. Should all robots have visible bar-codes or similar? Should the user or owner (e.g. a parent who buys a robot for a child) always be able to look up a database or register where the robot’s functionality is specified? See also rule 5 below.saying that robots should never be designed solely or even principally, to be used as weapons with deadly or other offensive capability. This law, if adopted, limits the commercial capacities of robots, but we view it as an essential principle for their acceptance as safe in civil society.

V.

Legal

The person with legal responsibility for a robot should be attributed.

General Audience

It should be possible to find out who is responsible for any robot.

Commentary

In this rule we try to provide a practical framework for what all the rules above already implicitly depend on: a robot is never legally responsible for anything. It is a tool. If it malfunctions and causes damage, a human will be to blame. Finding out who the responsible person is may not however be easy. In the UK, a register of who is responsible for a car (the “registered keeper”) is held by DVLA; by contrast no one needs to register as the official owner of a dog or cat. We felt the first model was more appropriate for robots, as there will be an interest not just to stop a robot whose actions are causing harm, but people affected may also wish to seek financial compensation from the person responsible.

Responsibility might be practically addressed in a number of ways. For example, one way forward would be a licence and register (just as there is for cars) that records who is responsible for any robot. This might apply to all or only operate where that ownership is not obvious (e.g. for a robot that might roam outside a house or operate in a public institution such as a school or hospital). Alternately, every robot could be released with a searchable online licence which records the name of the designer / manufacturer and the responsible human who acquired it (such a licence could also specify the details we talked about in rule 4 above). There is clearly more debate and consultation required.

Importantly, it should still remain possible for legal liability to be shared or transferred e.g. both designer and user might share fault where a robot malfunctions during use due to a mixture of design problems and user modifications. In such circumstances, legal rules already exist to allocate liability (although we might wish to clarify these, or require insurance). But a register would always allow an aggrieved person a place to start, by finding out who was, on first principles, responsible for the robot in question.

Seven High-Level Messages

In addition to the above principles the group also developed an overarching set of messages designed to encourage responsibility within the robotics research and industrial community, and thereby gain trust in the work it does. The spirit of responsible innovation is, for the most part, already out there but we felt it worthwhile to make this explicit. The following commentary explains the principles.

I.

Principle

We believe robots have the potential to provide immense positive impact to society. We want to encourage responsible robot research.

Commentary

This was originally the “0th” rule, which we came up with midway through. But we want to emphasize that the entire point of this exercise is positive, though some of the rules can be seen as negative, restricting or even fear-mongering. We think fear-mongering has already happened, and further that there are legitimate concerns about the use of robots. We think the work here is the best way to ensure the potential of robotics for all is realised while avoiding the pitfalls.

II.

Principle

Bad practice hurts us all.

Commentary

It’s easy to overlook the work of people who seem determined to be extremist or irresponsible, but doing this could easily put us in the position that GM scientists are in now, where nothing they say in the press has any consequence. We need to engage with the public and take responsibility for our public image.

III.

Principle

Addressing obvious public concerns will help us all make progress.

Commentary

The previous note applies also to concerns raised by the general public and science fiction writers, not only our colleagues.

IV.

Principle

It is important to demonstrate that we, as roboticists, are committed to the best possible standards of practice.

Commentary

As previous

V.

Principle

To understand the context and consequences of our research we should work with experts from other disciplines including: social sciences, law, philosophy and the arts.

Commentary

We should understand how others perceive our work, what the legal and social consequences of our work may be. We must figure out how to best integrate our robots into the social, legal and cultural framework of our society. We need to figure out how to engage in conversation about the real abilities of our research with people from a variety of cultural backgrounds who will be looking at our work with a wide range of assumptions, myths and narratives behind them.

VI.

Principle

We should consider the ethics of transparency: are there limits to what should be openly available

Commentary

This point was illustrated by an interesting discussion about open-source software and operating systems in the context where the systems that can exploit this software have the additional capacities that robots have. What do you get when you give “script kiddies” robots? We were all very much in favour of the open source movement, but we think we should get help thinking about this particular issue and the broader issues around open science generally.

VII.

Principle

When we see erroneous accounts in the press, we commit to take the time to contact the reporting journalists.

Commentary

Many people are frustrated when they see outrageous claims in the press. But in fact science reporters do not really want to be made fools of, and in general such claims can be corrected and sources discredited by a quiet & simple word to the reporters on the byline. A campaign like this was already run successfully once in the late 1990s.

RoboLaw Project

The main goal of the RoboLaw project is to achieve a comprehensive study of the various facets of robotics and law and lay the groundwork for a framework of "Robolaw" in Europe. The RoboLaw project aims at understanding the legal and ethical implications of emerging robotic technologies and of uncovering (1) whether existing legal frameworks are adequate and workable in light of the advent and rapid proliferation of robotics technologies, and (2) in which ways developments in the field of robotics affect norms, values and social processes we hold dear.

The problem of regulating new technologies has been tackled in Europe almost by every legal system: Therefore, it is possible to rely on a background which includes a large amount of studies on the relationship between law and science and between law and technology. Nevertheless, the RoboLaw project is focused on the extreme frontiers of technological advance, confronting the legal "status" of robotics, nanotechnologies, neuroprostheses, brain-computer interfaces, areas in which very little work has been done so far.

This project is the first in-depth investigation into the requirements and regulatory framework(s) of "robolaw" in the age of the actualization of advanced robotics, and the first study to combine the many different legal themes that have been investigated in isolation before. Moreover, it is the first research to delve into the legal and ethical consequences of developments in robotics within specific legal systems within the EU and to compare these with the US and the Far East, Japan in particular.

The complete book of **Guidelines on Regulating Robotics** can be found here: http://www.robolaw.eu/

European Robotics Research Network (EURON)

EURON is a network of excellence in robotics, that is aimed at coordination and promotion of robotics research in Europe. The network is sponsored by the European Commission through the Future and Emerging Technologies Programme.

Its **Roboethics Roadmap** can be found here:

http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf

European Civil Law Rules in Robotics

The European Parliament’s Legal Affairs Committee commissioned this study to evaluate and analyse, from a legal and ethical perspective, a number of future European civil law rules in robotics.

The full report can be found here: http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf

A CALL FOR MORE ENGAGED TECHNOLOGISTS, AND DIALOGUE INSTEAD OF MONOLOGUES, SUMMARY REPORT ON THE 2ND SESSION OF THE GROUP OF GOVERNMENTAL EXPERTS ON LAWS AT THE UN IN GENEVA

27 – 31 August 2018

Paper published in August 2018 by the ICT4Peace Foundation, Geneva[1]

From 27-31 August 2018, the Convention on Certain Conventional Weapons (CCW)[2] completed its sixth year of discussions on Lethal Autonomous Weapons Systems (LAWS). Representatives of more than 82 countries convened at the United Nations in Geneva as a so-called Group of Governmental Experts (GGE). It was the second and last meeting of the GGE in 2018.

Four agenda items were debated during the one-week session: (1) the potential military applications of emerging technologies in the field of LAWS, (2) the characteristics of LAWS, (3) if and to what degree a human element should and could be secured in the use of lethal force, and (4) possible options to address the humanitarian and international security challenges posed by LAWS.

Inputs on potential military applications of related technologies (1) have mainly been channeled through expert members of national delegations,[3] and a panel put together at the invitation of Chairman Amandeep Singh Gill on Monday, 27 August.[4] In this opening panel, Dr. Dörmann and Lieutenant Colonel (LK) Korpela

[1] https://ict4peace.org/wp-content/uploads/2019/08/ICT4Peace-2018-AI-AT-LAWS-Peace-Time-Threats.pdf

[2] The CCW is properly referred to as the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects.

[3] e.g. Sweden on 'Sensor-fuzed munition: An autonomous weapon?'

[4] Dr. Lydia Kostopoulos, Digital Society Institute; Prof. Anthony Gillespie, UCL; Dr. Knut

presented some of the main perspectives that emerged during the rest of the week-long debate. Dr. Dörmann argued that, if machines can self-initiate an attack, this necessarily introduces uncertainty regarding location, timing, and nature of this attack. Consequently, this would imply a significant risk that the machine will not be able to comply with international humanitarian law (IHL), especially the principles of distinction, proportionality, or precaution. LC Korpela, on the other hand, argued that the idea of LAWS is really to help militaries to adapt to uncertain environments, allowing them to take more accurate decisions than humans in less time.

With regards to the characterization of LAWS (2), many states argued that it is not in their interest to develop fully autonomous weapons systems, as commanders always want to retain a certain amount of control over the use of force. Therefore, many states outlined their interest in ensuring human control/judgment in attack decisions, in the ability to cancel an attack, as well as within the accountability of operators, in order to guarantee that the use of force remains in human hands rather than within machine algorithms. This perspective was reflected by a majority of governments. They regard a LAWS as a weapons system for which a certain degree of human control is ensured. Therefore, agenda items (2), the characterization of LAWS, and (3), the degree of the human element, had fluid boarders during the debate. During the discussion on the outcome document, it was suggested that both items should be merged under one sub-chapter. This proposal was rejected. The degree of human involvement (3) was the main 'dividing point' regarding the type of outcome of the CCW's debate on LAWS (4). Some states argued that meaningful human control over, e.g., targeting, selection and execution of force must always be guaranteed. Most states favoured a negotiation of a legally binding instrument, in order to ensure that everyone abides by the same rules,[5] or at least a non-binding political declaration.[6] Other states argued that the human element still needs to be better understood: what does human control really mean and how much and where in the targeting cycle must it be ensured? Those states opposed immediate further legal or political restraint and preferred discussions to continue as is. Some of those states argued that the lack of a common

Dörmann, ICRC; LC Chris Korpela, DOD US; Gautam Shroff, Tata Consultancy Services

[5] This is the uncompromising demand of Algeria, Argentina, Austria, Bolivia, Brazil, Chile, China (with regards to 'fully' autonomous weapons), Colombia, Costa Rica, Cuba, Djibouti, Ecuador, Egypt, Ghana, Guatemala, Holy See, Iraq, Mexico, Nicaragua, Pakistan, Panama, Peru, State of Palestine, Uganda, Venezuela and Zimbabwe. Colombia, Iraq, Pakistan, Panama, the non-aligned movement (NAM) group of states and others called for an immediate preemptive ban of LAWS.

[6] Switzerland, France, Germany.

understanding of either human control or LAWS, or both, required discussions to continue until more clarity is achieved.[7] Others insisted on their own definition of LAWS, which complicated overall acceptance.[8]

The parties of the CCW that met formally as the GGE were tasked to make a recommendation on future work to the CCW annual meeting in November 2018. Although many states called for a legally binding instrument or a political declaration, the GGE, after eight extra-hours of discussions until 1:10 am Saturday, 1 September, rejected those options and decided to continue with its current mandate. The draft final report with possible guiding principles can be found here.

General observations:

1. Representatives of the tech sector seem to be underrepresented in the debate. On Thursday, Conscious Coders held a side event on the technical aspects and risks of AI, which was highly welcomed by state representatives. The latter argued that such a clear technological overview had been lacking within the debate.[9]

2. NGOs, especially the Campaign to Stop Killer Robots, which argues for a ban, usually base their arguments on ethical considerations. They state that death by a machine is unethical, as a machine lacks basic human characteristics such as compassion, empathy, dignity, and the understanding of human life and of the 'taking' of human life. They use this argument in order to create a distinction between two situations: the one where a machine kills an enemy combatant, and the other where a combatant kills an enemy combatant – arguing that the former is unethical. Yet, one must argue that a situation where a combatant kills another enemy combatant is not a situation where characteristics like compassion, empathy, dignity and the understanding of human life are at the forefront. Consequently, basing anethical argument with the view to distinguish those situations on the above-mentioned human concepts forces one to ask what remains of ethics.

[7] Australia, Israel, United States of America, Republic of Korea.

[8] Russia.

[9] For the latest report on the misuse of AI, see https://www.thinktech.ngo/wp-content/uploads/2018/06/ConsciousCoders_issue_01_web.pdf (accessed on 3 September, 2018).

3. The discussion does not include any reference on narrow and general artificial intelligence (AI), which could be helpful to understand the difference between current and future LAWS.[10]

The discussion also does not distinguish between autonomy on land, underwater and in air. The autonomous technologies as well as military endeavors to use them vary greatly for those different war scenarios. More in-depth analysis of the matter at hand is needed. This can only be created by a dialogue between different groups. At present, it seems that all participants, both the group that favor a ban as well as those that want more discussion, are engaged in monologues that do not intersect.

[10] See e.g. Lewis, Lawrence, 2018, AI and Autonomy in War: Understanding and Mitigating Risks, available at: https://www.cna.org/CNA_files/PDF/DOP-2018-U-018296-Final.pdf (accessed on 3 September, 2018).

OP-EDS

AUSLAGERUNG DES GRUNDRECHTSSCHUTZES VON DER POLITIK AUF FIRMEN

GASTKOMMENTAR (NZZ)

Künstliche Intelligenz gefährdet unsere Privatsphäre. Die Politik ist überfordert, Technologiekonzerne versuchen durch Selbstregulierung Standards zu setzen. Es ist ein Umdenken erforderlich.

Regina Surber, Neue Zürcher Zeitung, 24. April.2019[1]

Wir spüren es zwar im Alltag nicht immer, aber wir wissen es: Künstliche Intelligenz (KI) dringt mehr und mehr stark in unsere Privatsphäre ein. Privatsphäre definieren wir dabei als die Möglichkeit für das einzelne Individuum, sich zurückzuziehen und private Informationen zurückzuhalten, wenn wir das so wollen. Sie galt einst als Vorbedingung für die Ausübung gewisser Menschenrechte, etwa des Rechts auf freie Meinungsäusserung oder der Wahl - oder Versammlungsfreiheit. Dieses Recht auf Privatsphäre verlangt auch im Informationszeitalter, dass wir selber kontrollieren können, wie unsere Daten gespeichert, verändert und ausgetauscht werden.

Neue Risiken

Mit dem Aufkommen von immer neuen Datendurchsuchungs- und Datenerhebungsmethoden (Data-Mining) wird dieses Recht zunehmend infrage gestellt: Regierungsbehörden und Unternehmen können heute den einzelnen Bürger leicht identifizieren und Profile über ihn erstellen. Die rasch wachsende Rechenkapazität beschleunigt, vergrössert und automatisiert diese Möglichkeiten,

[1] https://www.nzz.ch/meinung/daten-grundrechtsschutz-und-politik-ld.1470731

Informationen zu sammeln und zu verarbeiten. Dies stellt die liberale Gesellschaft vor eine grosse gesellschaftliche Herausforderung.

KI gefährdet den Schutz unserer Privatsphäre dabei auf unterschiedliche Weise: Smartphones und Computersoftware generieren konstant Daten, aufgrund deren wir identifiziert, verfolgt und überwacht werden können, egal, ob wir uns zu Hause oder am Arbeitsplatz befinden. Selbst an sich anonyme persönliche Daten können durch KI leicht deanonymisiert werden. KI wird künftig auch immer genauer Stimmen und Gesichter identifizieren können. Strafverfolgungsbehörden können solcherart Individuen auch aufspüren, ohne dass klare Verdachtsmomente vorliegen und ohne dass rechtsstaatliche Voraussetzungen einhalten werden müssen.

KI kann durch ausgetüftelte Algorithmen auch sensitive persönliche Informationen aus nichtsensitiven Daten ableiten: Gefühlszustände, politische Einstellungen, Gesundheit oder sexuelle Orientierung. An sich harmlose Ortungs- und Log-in-Daten ermöglichen dabei erstaunliche Rückschlüsse auf das einzelne Individuum. Und so kann KI Personen auch klassifizieren und beurteilen, ohne dass dafür die Zustimmung des Einzelnen eingeholt werden muss. Chinas soziales Kreditsystem ist ein Beispiel, wie solche persönlichen Informationen verwendet werden können, um einzelne Individuen oder bestimmte soziale Gruppen vom Zugang zu Krediten, Anstellungen, Mietobjekten oder sozialen Dienstleistungen auszuschliessen.

Diese neuen Risiken für unser Grundrecht auf Privatsphäre verlangen eine öffentliche und politische Debatte. Längst sind grosse Technologiekonzerne eingesprungen und füllen hinsichtlich des Datenschutzes das rechtspolitische Vakuum durch Selbstregulierung: Microsoft, IBM, Google und Co. haben sich selber Regeln auferlegt, oft «ethische Standards» genannt, mittels deren sie garantieren wollen, dass ihre KI-unterstützten Technologien die Privatsphäre schätzen und schützen. Unternehmen beobachten, analysieren und bewerten also die Bedürfnisse der Bevölkerung nach Privatsphäre, dies jedoch stets vor einem wettbewerbsorientierten Hintergrund und unter hohem Zeitdruck, denn der technologische Fortschritt kennt keine Geduld.

Die Politik hinkt hinterher

Dieser Selbstregulierung fehlt allerdings die demokratische Legitimation und Kontrolle. Schranken für Grundrechtsüberschreitungen oder - verletzungen werden nicht mehr von der Politik, sondern von der Privatwirtschaft definiert. Dies betrifft derzeit primär noch die Technologie-Grosskonzerne, künftig werden aber fast alle Firmen weit stärker auf die KI zurückgreifen.

Die Politik hinkt beim Thema Grundrechtsschutz der Realität hinterher und überlässt bedeutsame Aufgaben weitgehend der Privatwirtschaft. Dies aus zwei Gründen: Erstens mahlen die Mühlen der Politik im Vergleich zum rasanten technologischen Fortschritt viel zu langsam, und zweitens fehlt es in der Politik klar an Know-how im Zusammenhang mit den neuen Technologien.

Deshalb braucht es ein Umdenken. Traditionell sucht man in der Schweizer Politik immer nach perfekten Lösungen. Dadurch setzt man sich oft allzu hohe Hürden: Man strebt abschliessende, fixfertige Politiklösungen an und verliert dabei vor lauter Bäumen den Blick für den Wald, den es eigentlich rasch zu bändigen gilt. Gefragt ist künftig ein fokussierter und konstanter Austausch mit Technologie-Experten: So kann man sich von Baum zu Baum hangeln und auf diese Weise versuchen, die politische Hoheit über den Grundrechtsschutz zurückzugewinnen. Dies setzt einen konstruktiven Dialog mit den privatwirtschaftlichen Vorreitern voraus.

AUTONOME INTELLIGENZ IST NICHT NUR IN KRIEGSROBOTERN RISKANT

GASTKOMMENTAR (NZZ)

Die Diskussion über die Gefährlichkeit autonomer Waffensysteme ist wichtig, verläuft aber einseitig. Künstliche Intelligenz bietet sich auch für hybride Kriegsführung an.

Regina Surber Neue Zürcher Zeitung, 20. Februar. 2018[1]

Kürzlich machte eine neue Anwendung von künstlicher Intelligenz in den nationalen und internationalen Medien Furore: Autonome Waffensysteme, Kriegsroboter oder sogenannte LAWS (lethal autonomous weapons systems) fanden durch News-Beiträge und vor allem durch den Kurzfilm «Slaughterbots» Eingang in unsere zerebralen Angstzentren. Dieser mediale Fokus ist gerechtfertigt, weil Waffensysteme, die ein Ziel ohne menschliches Vetorecht identifizieren, aussuchen, verfolgen und attackieren können, heute existieren und stetig verbesserte Algorithmen zu immer ausgefeilteren Nachfolgern führen können. Zudem zwingen uns LAWS, grundlegende Fragen zu stellen und den Ist-Zustand der Welt sowie existierende und potenzielle technische Zukunftsformer zu hinterfragen.

Diese Fragen sind einerseits normativ: Darf ein Mensch einen Algorithmus kreieren und verwenden, welcher den Tod eines anderen berechnen und herbeiführen kann? Andererseits sind es Fragen, die das grundlegende Selbstverständnis des Menschen betreffen: Können Leistungen von Mensch und Maschine am selben Massstab gemessen werden, und darf man Mensch und Maschine vergleichen? Dass eine breitere Öffentlichkeit als ein Uno-Gremium, welches gegenwärtig über die kriegsvölkerrechtlichen Aspekte von LAWS diskutiert, sich der Kontroversen neuer Technologien bewusst wird, ist aufgrund der Tragweite dieser Fragen notwendig.

[1] https://www.nzz.ch/meinung/autonome-intelligenz-ist-nicht-nur-in-kriegsrobotern-riskant-ld.1351011

Auch der zivile Bereich betroffen

Das Thema LAWS bringt aber das Risiko mit sich, dass die öffentliche Debatte bald verstummt. Denn die Kontroverse ist auf der höchstmöglichen Diskussionsebene bei der Uno in Genf vielleicht ganz gut versorgt. Und warum sollte sich z. B. gerade die schweizerische Öffentlichkeit mit Kriegsrobotern auseinandersetzen?

Die stark vernetzte Schweiz drängt sich als Diskussionsschauplatz für zukunftsformende Ideen geradezu auf.

Erstens diskutiert das Uno-Gremium lediglich über den Gebrauch von LAWS zu Kriegszeiten. Autonome Waffensysteme können aber auch während nationaler Polizeioperationen – beispielsweise Geiselsituationen und Massenkontrollen – verwendet werden und werden für solche Szenarien von Firmen wie Desert Wolf schon entwickelt. Auch ignoriert die Uno-Debatte die wirtschaftlich und strategisch lukrative Verwendung von autonomer Technologie in Cyberoperationen – z. B. beim NSA-Programm «MonsterMind». Zudem blendet das Forum mögliche Risiken anderer neuer Technologien – wie zum Beispiel 5G und Biotechnologie – sowie eventuelle Verknüpfungen derselben mit autonomer Technologie aus.

Zweitens birgt autonome Technologie nicht nur Risiken, wenn sie absichtlich als Waffe entwickelt und verwendet wird. Sie könnte etwa Fake-News generieren und die Masse falsch informieren. Oder sie kann genutzt werden, um eigens Täterprofile zu generieren und die Grenze zwischen einem Kriminellen und einem rechtlich Unschuldigen basierend auf Big Data selber zu kalkulieren und zu ziehen. Überwachungskameras in Moskau und China sind vermehrt mit Gesichtserkennungstechnologie versehen und generieren kontinuierlich Daten von Gesichtern und dem Verhalten der entsprechenden Personen.

Auch müssen wir uns die Frage stellen, ob wir in Zukunft gezwungen sein werden, die Weltpopulation künstlich in Grenzen zu halten, weil das heutige globale Finanz- und Wirtschaftssystem die Ressourcen nicht für alle zufriedenstellend verteilt. Entscheidungen über Leben und Tod eines Weltenbürgers z. B. einer autonomen Software, versteckt im Gesundheitssystem, zu überlassen, würde uns die moralische Schwere der Entscheidung oberflächlich betrachtet vorerst abnehmen. Gekoppelt mit einem Bewertungssystem für Bürger – ein Prototyp wird China mit dem «Citizen Score» per 2020 landesweit einführen –, könnten die Populationszahlen basierend auf utilitaristischen Kalkulationen durch autonome Technologien begrenzt werden.

Risiken minimieren

Diese Gedankengänge zwingen uns, unser zeitgenössisches Finanz- und Wirtschaftssystem sowie unsere gesellschaftlichen Wertesysteme zu hinterfragen. Murphy’s Law besagt, dass alles, was schiefgehen kann, irgendwann schiefgehen wird. Deswegen haben wir die moralische Pflicht, Risiken für gefährliche Szenarien so stark zu minimieren wie möglich. Es braucht also, drittens, eine ethische Debatte auf einem ganz anderen Niveau.

LAWS dürfen deshalb nicht nur als das Problem der Kriegsroboter verstanden werden, welches an der Uno in Schach zu halten versucht wird. Wir müssen LAWS als Vorboten einer globalen Entwicklung verstehen hin zu einer Welt, in welcher der Mensch nicht mehr das einzige «intelligente System» mit der Fähigkeit zu autonomem Handeln darstellen könnte. Für solche Fragen braucht es die aktive Mitwirkung von Zivilgesellschaft, Akademie und Privatwirtschaft. Die Schweiz und vor allem die Stadt Zürich als Sitz der Tech-Grössen Google, IBM und Disney sowie der renommierten ETH drängt sich als Diskussionsschauplatz für zukunftsformende Ideen geradezu auf.

VIER FORDERUNGEN ZUR REGULIERUNG KÜNSTLICHER INTELLIGENZ

GASTKOMMENTAR (NZZ)

In den USA wie in China wird derzeit massiv in die Entwicklung von künstlicher Intelligenz (KI) investiert. Wie die Versuche mit selbstfahrenden Autos zeigen, sind deren Möglichkeiten noch limitiert. Doch das kann sich schnell ändern. Es ist dringlich, dass wir uns früh genug darauf vorbereiten.

Regina Surber und Daniel Stauffacher Neue Zürcher Zeitung, 19. September, 2018[1]

Die aus der Forschung zur künstlichen Intelligenz (KI) entstehenden Technologien helfen Banken bei der Digitalisierung, lösen Justizfälle, schwärmen als koordinierte Drohnen aus, sind der Schlüssel der intelligenten Netzwerkstruktur von jedem Internetprovider oder sitzen als Roboterhunde auf unserem Schoss. KI-unterstützte Technologien sind also stille Basis der Gesellschaft geworden, was den Hype um die zwei Buchstaben rechtfertigt. Allerdings reden einige über KI, ohne zu wissen, worum es sich im Kern handelt und wie gross die daraus resultierenden Potenziale und Risiken für Mensch und Gesellschaft sind. Die Risiken verlangen dringend nach entsprechenden Regierungsmassnahmen. KI muss von Menschen kontrolliert und in die richtigen Bahnen geleitet werden.

Die Forschung zu KI geht einerseits dahin, Soft- und Hardware zu kreieren, welche Merkmale menschlicher Intelligenz wie zum Beispiel die Fähigkeit zur Problemlösung oder zum Lernen aufweisen sollen. Andererseits bezeichnet KI das formlose Können einer Soft- oder einer Hardware, welches die obgenannten intelligenten Merkmale erzeugt, wie die Fähigkeit einer Software, autonom ein Auto zu fahren. KI kann sowohl als kommerzialisierbare Ressource wie auch als gestaltlose Grundlage für Wohlstand behandelt werden. Es wohnt ihr beträchtliches politischen Gewicht inne.

[1] https://www.nzz.ch/meinung/vier-forderungen-zur-regulierung-kuenstlicher-intelligenz-ld.1407898

Risikoreiche Transformationen

Heutige KI im zweiten Sinne bezeichnet man als «schwach», weil sie nur eine einzige Aufgabe gut lösen kann, wie etwa Gesichtserkennung. Eine «starke» KI wiederum würde eine dem Menschen vergleichbare Intelligenz demonstrieren. «Künstliche Superintelligenz» bezeichnet eine dem Menschen überlegene Intelligenz. Gewisse Experten glauben, dass starke KI innerhalb der nächsten 75 Jahre hergestellt werden kann, andere tun das als Science-Fiction ab.

KI muss von Menschen kontrolliert und in die richtigen Bahnen geleitet werden

KI ist ein Treiber risikoreicher gesellschaftlicher Transformationen: Autonome Waffen können auf Insektengrösse verkleinert werden und in grosser Zahl zu sehr billigen intelligenten Massenvernichtungswaffen werden. Kriegsführung ist dann nicht mehr Kampf zwischen Soldaten, sondern Systemkonfrontation auf elektromagnetischer Ebene und im Cyberspace, wo autonome Cyberwaffen eine Hauptrolle spielen. Auch kann intelligente Software künstliche Krankheitserreger kreieren.

Zudem führen verzerrte Daten zu verzerrten KI-Software-Resultaten, was schon in rassistischen Justizentscheidungen in den USA resultierte. So werden soziale Stigmata mittels Technologien reproduziert, deren Entscheidungen im Einzelfall nicht zurückzuverfolgen und schwer anfechtbar sind. Ausserdem führt Massen-, Fehl- und Falschinformation zum Verlust eines gesellschaftlichen Wahrheitsklimas, was die Frage aufwirft, ob wir ein Recht auf wahrheitsgetreue Informationen haben.

Jetzt handeln

Solche leisen Veränderungen verlangen ein unverzügliches Engagement von Politik, Akademie und Zivilgesellschaft: Erstens ist eine fundierte öffentliche Diskussion über soziale Auswirkungen von KI-unterstützten Technologien zwingend. Zweitens muss KI-Forschung ethisch eingebettet werden, weshalb universitäre Lehrstühle für Ethik und Technologie geschaffen werden müssen, was an der ETH Zürich momentan diskutiert wird. Hier muss die Privatwirtschaft als heutiger Hauptinvestor in KI mit einbezogen werden.

Drittens muss sich die Infrastruktur unserer nationalen Politik raschestmöglich dem Paradigmenwechsel anpassen, bevor es zu spät oder technisch zu komplex wird. Diese Funktion könnte in einem ersten Schritt ein hochrangiger Delegierter des Bundesrates für Technologiefragen wahrnehmen. Und viertens gilt es, Klarheit darüber zu erlangen, ob Algorithmen, welche die Privatsphäre oder gar die körperliche Integrität – Stichwort autonome Waffensysteme – von Bürgern, sprich unsere in der Verfassung verankerten Grundrechte zu verletzen in der Lage sind, vom Parlament diskutiert und allenfalls verboten werden sollten.

GEFÄHRLICHES SPIEL OHNE REGELN

GASTKOMMENTAR (NZZ)

Cyberangriffe auf kritische Infrastrukturen häufen sich und scheinen vermehrt staatlichen Ursprungs. Der Cyberspace ist eine völkerrechtliche Grauzone, die rechtlich verbindliche Konturen braucht.

Regina Surber Neue Zürcher Zeitung, 22. Februar 2017[1]

Es scheint evident, dass **viele der jüngsten Cyberattacken** auf kritische Infrastrukturen nicht mehr nichtstaatlichen Cyberkriminellen angelastet werden können, sondern von Staaten angeordnet wurden. Dabei gibt es noch immer keine allgemein anerkannten, geschweige denn verbindlichen Normen, die staatliche Cyberangriffe beschränken. Zwischenstaatliche Konflikte in den Cyberspace auszulagern, bietet sich deswegen an. **Tatsache ist, dass für staatliche Cyberoperationen bereits enorme finanzielle Mittel aufgewendet werden.** Gemäss dem United Nations Institute for Disarmament Research (Unidir) besitzen mehr als 47 Staaten Cyber-Security-Programme, bei denen den nationalen Streitkräften eine beachtliche Rolle zugeschrieben wird.

Potenziell verheerende Konsequenzen

Ohne Einschränkung der staatlichen Handlungsspielräume würde der Cyberspace als Free-Fire-Zone belassen – mit potenziell verheerenden Konsequenzen. Moderne Technologien im Cyberspace bringen **eine neue Generation von unsichtbaren und ungreifbaren Offensivwaffen** zum Einsatz. Cyberangriffe sind unmittelbar und schwierig zu erkennen. Zudem ist es einfacher, zu attackieren, als sich zu verteidigen. Dies senkt die Hürde für Präventivschläge. Ferner ist der **Grat zwischen Spionageaktionen und Angriffen** in der virtuellen Welt viel schmaler als in der realen: Wenn man in ein Computernetzwerk eindringen kann, so kann man das System ebenso einfach manipulieren wie zerstören. Cyberangriffe auf kritische Infrastrukturen sind rechtlich nicht zwingend verboten. Deshalb steht einem Wettrüsten mit Cyberwaffen wenig im Weg.

[1] https://www.nzz.ch/meinung/staatliche-cyberattacken-gefaehrliches-spiel-ohne-regeln-ld.146933

Wir benötigen internationale Normen, die den Cyberspace mittels klarer Definitionen vor zulässigem und unzulässigem staatlichem Handeln schützen und so als globales Gemeingut bewahren. Sowohl die Uno-Charta von 1945 als auch die Genfer Konventionen von 1949 und 1977, die Grundpfeiler des bestehenden Kriegsvölkerrechts, schweigen zum Thema Cyberkonflikte. Ob und wie das bestehende Recht auf diese angewandt werden könnte, ist umstritten.

Die Nato wagte sich in ihrem Tallinn-Manual, einem Handbuch zu Cyberkrieg-Regeln, 2013 an eine erste, jedoch nichtbindende Definition von Cyberangriffen. Das Dokument wurde grösstenteils von westlichen Militärs und militärnahen Juristen ausgearbeitet, weswegen es ihm an globalem Geltungsanspruch fehlt.

Die richtigen Weichen stellen

Auf internationaler Ebene führt die Uno-Expertengruppe zu Information und Telekommunikation im Kontext der internationalen Sicherheit die rechtliche Diskussion und potenzielle Entwicklung eines Normensystems an. 2013 entschied sie, dass obengenanntes geltendes Recht, speziell die Uno-Charta, per se auf staatliche Cyberaktivitäten anwendbar ist. Wie genau, liess sie aber im Dunkeln. 2015 identifizierten die Experten zwar konkretere Normen – etwa, dass Staaten nicht wissentlich und absichtlich kritische Infrastrukturen angreifen dürfen –, unterstrichen jedoch den freiwilligen Charakter dieser Vorschriften.

Diese Woche trifft sich die Gruppe, in der auch die Schweiz Mitglied ist, zu weiteren Gesprächen in Genf. In der gegenwärtigen Uno-Expertengruppe sitzen – nebst den fünf permanenten Mitgliedern des Sicherheitsrats – nur zwanzig Staaten. Von einem globalen Engagement kann trotz Uno-Attribut auch hier nicht wirklich die Rede sein.

Die genannten Initiativen stellen zwar die richtigen Weichen. Wenn die Chance auf Akzeptanz und Einhaltung von neuen Normen für verantwortungsbewusstes staatliches Verhalten im Cyberspace aber maximiert werden will, muss eine Mehrheit der Staaten in die Diskussion ihrer Schaffung mit einbezogen werden. Ausserdem sind zentrale Fragen noch immer unbeantwortet: Wie soll bestehendes Recht auf den Cyberspace angewandt werden? Muss neues geschaffen werden, und, wenn ja, welche Lücken hat es zu füllen? Und vor allem: Soll die Einhaltung dieser Normen freiwillig bleiben? Es scheint offensichtlich, dass in Anbetracht der jüngsten staatlichen Cyberangriffe ein paar unverbindliche Prinzipien nicht genügen. Ohne rechtlich verbindliche Verbote sind obengenannte Risiken wie ein Wettrüsten mit Cyberwaffen nicht gebannt.

TERRORISMUS, EINE VIRTUELLE TATSACHE

NZZ - Ausland

Terroristen nutzen das Internet raffiniert, um Anschläge zu koordinieren und Gewalt zu propagieren. Sie scheinen Regierungen und Internetfirmen stets einen Schritt voraus. Wie lässt sich das ändern?

Daniel Stauffacher, Regina Surber, Neue Zürcher Zeitung, 5. October 2016[1]

Eigentlich hätten wir es wissen müssen. Bereits 1990 prognostizierten Experten der Vereinten Nationen, dass der technische Fortschritt, den wir ungehemmt fördern, Phänomene unterstützen könnte, die wir bekämpfen möchten: Terrorismus, innerstaatliche Gewalt, ethnische und religiöse Intoleranz. 25 Jahre später ist allen klar, dass sich diese Warnungen bewahrheitet haben: Viele der jüngst aufgetauchten Terrororganisationen haben sich zu versierten Nutzern des Internets gemausert, insbesondere von Social-Media-Plattformen.

Die Liaison von IT und IS

Die Fachkenntnis und das Raffinement, mit welchen diese Gruppierungen die Informationstechnologien nutzen, haben viele überrascht. Beispiele gibt es genügend: Die Terrormiliz Islamischer Staat (IS) verteilt Anleitungen an ihre Mitglieder, in denen sie bestimmte Online-Plattformen empfiehlt und erklärt, wie man das Risiko minimiert, abgehört zu werden. Ferner benutzt der IS das Internet unmittelbar nach seinen Anschlägen, wenn das Potenzial, Neumitglieder zu rekrutieren, besonders gross ist. Al-Kaida verteilte über ihr Online-Magazin «Inspire» Anleitungen zum Bombenbau, was als eine der Inspirationsquellen für den Anschlag am Boston-Marathon 2013 gilt. Die der IS-Ideologie verwandte Gruppierung «Wafa Media Foundation» kündigte im Juni Anschläge in Spanien an und ermutigte private Einzelkämpfer, Spanier zu entführen. Auch pflegen unzählige radikalisierte Personen gekonnt Social-Media-Konten.

Das Internet ist für Terrororganisationen also Kapital. Es vereinfacht die Kommunikation, die Propaganda, die Aufforderung zu Gewalt, die Rekrutierung von Mitgliedern, den Wissenstransfer sowie die finanzielle Abwicklung von Anschlägen. Kombiniert,

[1] https://www.nzz.ch/international/nahost-und-afrika/das-netz-als-waffe-terrorismus-eine-virtuelle-tatsache-ld.120312

verstärken diese Funktionen die Wirkungsmacht von Terrorgruppen enorm. Wenn Rekrutierung und Kommunikation nicht mehr nur physisch, sondern auch im Netz stattfinden, nützen klassische militärische Ansätze wenig. Das bekräftigt auch der Uno-Generalsekretär in seinem jüngsten Bericht: Die gegenwärtigen militärischen und wirtschaftlichen Massnahmen gegen den IS haben nicht geholfen, seine Nutzung des Cyberspace zu reduzieren.

Was entgegnet man dieser neuen Form der Terrororganisation – insbesondere angesichts des exponentiell wachsenden technischen Fortschritts? Wie können Regierungen, die im heutigen internationalen System das Gewalt- und Sicherheitsmonopol besitzen, wirkungsvoll gegen einen Gegner vorgehen, der sich Dienstleistungen bedient, die hauptsächlich von privatwirtschaftlichen Akteuren angeboten werden?

Ideen aus dem Kalten Krieg

Auf nationaler Ebene fokussieren sich die Reaktionen erstens auf die Deradikalisierung und die Entkräftung von ideologischen Botschaften im Internet. Verwendet werden oft Propagandastrategien, welche aus den Zeiten des Kalten Krieges stammen. Beispiele sind die sogenannte Counter-Initiative des Vereinigten Königreichs oder das «Madison Valley Wood»-Projekt in den USA**.**

Zweitens verpflichten Staaten Internetfirmen dazu, bedrohliche Inhalte entweder von vorneherein zu blockieren oder vor der Veröffentlichung herauszufiltern. Diese Regulierungsmassnahmen basieren allgemein auf rechtlichen Grundlagen. Allerdings stützen einige Länder diese Weisungen nicht auf offizielle Rechtstexte, sondern auf die Nutzungsbedingungen der IT-Unternehmen ab. Die Regulierungsmassnahmen sind in diesem Falle aussergesetzlich. Ein Beispiel hierfür ist die United Kingdom Counter-Terrorism Internet Referral Unit (CTIRU), durch deren Aufforderung seit 2010 mehr als 163 000 Online-Inhalte auf diversen Websites gelöscht worden sind.

Auf der inter- und supranationalen Ebene konzentrieren sich die Reaktionen ebenfalls auf die Verbreitung von Gegennarrativen sowie die Filterung und Überwachung von Inhalten. Das Counter-Terrorism Executive Directorate erarbeitet zurzeit einen Vorschlag für eine Rahmenvereinbarung, um den von IS und al-Kaida verwendeten Narrativen entgegenzuwirken. Die Internet Referral Unit der EU ist eine der CTIRU ähnliche Institution, welche terroristisch motivierte Inhalte im Netz identifiziert und den EU-Mitgliedstaaten meldet.

Bei staatlichen Regulierungsmassnahmen ist das Einbeziehen der Privatwirtschaft – vor allem im Bereich der Informationstechnologie – essenziell. Leistungsträger wie Twitter, Facebook und Microsoft besitzen eine enorme Macht im Cyberspace. Viele dieser Unternehmen, besonders aus dem Bereich der Social Media, sahen sich bisher jedoch gezwungen, selbständig Massnahmen gegen die terroristische Nutzung ihrer Produkte zu ergreifen. Sie löschen deshalb vermehrt eigenhändig Inhalte von ihren Websites. Twitter hat etwa innert sieben Monaten 125 000 Benutzerkonten mit Verbindungen zu Terroristen von seiner Plattform entfernt. Viele Firmen ändern auch die Nutzungsbedingungen und verbieten die Veröffentlichung von «terroristischen Inhalten» auf ihren Websites. Das Problem ist, dass der Terminus nicht einheitlich definiert ist. Microsoft stützt sich deshalb auf eine Liste des Uno-Sicherheitsrates: Jegliches Material, welches mit den darauf aufgeführten Organisationen in Verbindung steht, stuft Microsoft als «terroristisch» ein und entfernt es.

Unsicherheit bleibt

Terroristische Inhalte im Netz können also gelöscht werden – aber an anderen Orten im Internet genauso rasch wieder auftauchen. Um dies zu verhindern, investieren Firmen neuerdings in Technologien, die Inhalte erkennen und entfernen, auch nachdem sie schon von einer Website gelöscht worden sind. Dies entlastet vor allem kleinere Firmen, die keine Ressourcen für derartige Kontrollmechanismen aufbringen können.

Die entscheidende Frage lautet allerdings: Sind diese Massnahmen effektiv? Es ist schlicht zu früh, um das zu beantworten. Auch ist ungewiss, wie Regierungen und Firmen den Erfolg dieser Ansätze messen können. Ferner ist offen, wie sich Staaten an die Herausforderungen, die mit der technischen Entwicklung einhergehen, anpassen können.

Eine wiederkehrende Frage ist auch, ob man gar die terroristische Nutzung des Internets unterstützen statt unterdrücken soll. Wenn gelöschte terroristische Inhalte an anderen Orten im Netz sofort wieder auftauchen, ist das Filtern nur eine kurzfristige Lösung, die enorme Ressourcen verschlingt. Erlaubt man hingegen terroristische

Inhalte, können Strafverfolgungsbehörden die Urheber einfacher überwachen und allenfalls Anschläge verhindern.

Des Weiteren werfen diese Entwicklungen komplexe normative Fragen auf: Kann man die staatliche Sicherheitsverantwortung mit den Anforderungen des Rechts auf Meinungs- und Informationsfreiheit in Einklang bringen? Welche Verantwortung tragen private Akteure in der Bekämpfung der terroristischen Nutzung von Informationstechnologien, und worauf basiert diese Verantwortung (Menschenrechte, Nutzungsbedingungen, Vertragsvereinbarungen)? Dürfen Regierungen die Durchsetzung ihrer Regulierungsmassnahmen vollständig an private Firmen auslagern?

Wenn sich Regierungen immer stärker auf technisch ausgerichtete Lösungen verlassen, ignorieren sie strukturelle Faktoren, die für das Entstehen des Terrorismus ursprünglich verantwortlich waren. Hart erarbeitete Prinzipien wie Mitsprache, Transparenz und Verantwortlichkeit in der Entscheidungsfindung werden so auf den zweiten Platz verwiesen. Es liegt auf der Hand, dass so viele verschiedene Akteure wie möglich in die Diskussionen über Herausforderungen, Lösungen, die Beurteilung der Effektivität und der sozialen Auswirkungen der Gegenmassnahmen involviert werden müssen. Internationale Initiativen wie der ICT Sector Guide on Implementing the Business and Human Rights Principles der EU, die Principles on Freedom and Privacy der Global-Network-Initiative, sowie die ICT4Peace-UNCTED-Initiative bieten Diskussionsplattformen und integrieren Akteure aus Politik und Privatwirtschaft.

Gleichzeitig müssen wir mit der rasanten technologischen Entwicklung Schritt halten. Künftige Technologien werden sicherlich unser heutiges Vorstellungsvermögen sprengen, deswegen müssen wir über bisherige Grenzen hinausdenken. Überwachung, Filterung und Gegennarrative mögen helfen, aber genügen wahrscheinlich kaum. Innovationen, gepaart mit Pragmatismus und extremer Schnelligkeit, sind unerlässlich.

RECORDED TALKS

Lecture on “Corona, Technology and Human Rights” https://ict4peace.org/activities/regina-surber-of-ict4peace-on-corona-technology-and-humanrights/

Lecture on “Autonomous Weapons Wai Talk”, HWZ Zurich, audio-recorded, 20 February 2020 https://ict4peace.org/activities/ict4peace-and-zhet-at-waitalk-the-dark-side-of-ai/

Lecture on “LAWS and the UN GGE Process” Swiss Federal Institute of Technology (ETH), audio-recorded, 29 July 2019 https://www.youtube.com/watch?v=T9U0xhUSR7g

Panel contribution on ‘AI: Civilian, Transdisciplinary, International Perspectives,’ video-recorded, side event of the UN GGE debate in Geneva, 2019. https://www.youtube.com/watch?v=1n_HtLdWJ5I

Presentation on “AI: LAWS and Peace-Time Threats’ at Swiss Cognitive Zurich”, video-recorded, 16 January 2018 https://www.youtube.com/watch?v=1n_HtLdWJ5I

Lecture on “AI: LAWS and Peace-Time Threats, Swiss Federal Institute of Technology (ETH)”, audio-recorded, November 2017 https://ethicsandtechnology.org/artificial-intelligence-lethal-autonomous-weapons-systems-and-peace-time-threats/

About ICT4Peace Foundation

ICT4Peace is a policy and action-oriented international Foundation. The purpose is to save lives and protect human dignity through Information and Communication Technology. Since 2003 ICT4Peace explores and champions the use of ICTs and new media for peaceful purposes, including for peacebuilding, crisis management and humanitarian operations. Since 2007 ICT4Peace promotes cybersecurity and a peaceful cyberspace through inter alia international negotiations with governments, international organisations, companies and non-state actors.

The ICT4Peace project was launched with the support of the Swiss Government in 2003 with the publication of a book by the UN ICT Task Force on the practice and theory of ICT in the conflict cycle and peace building in 2005 and the approval of para 36 of the Tunis Commitment of the UN World Summit on the Information Society (WSIS) in 2005.

ICT4Peace Website: www.ict4peace.org

ICT4Peace Academy - www.academy.ict4peace.org

ICT4Peace Publications: www.ict4peace.org/publications

ICT4Peace on Twitter - www.twitter.com/ict4peace

ICT4Peace on Facebook - www.facebook.com/ict4peace

ICT for peace foundation