‘AI, Human Dignity and Inclusive Societies’

Panel Questions for Regina Surber
AI4Good, Geneva, Switzerland
29 May 2019

Risks and Opportunities of AI

1. *In what areas do you see AI as posing the greatest risk to human rights?*

I see three different areas as very problematic

The first one concerns the use of AI in our information ecosystem and human dignity.
In simple words, AI can be used for two distinct things with regards to our information ecosystem.
First, it can be used to analyze data that we as humans generate. And our data reflects our behavior and our psychology. Thereby, AI can be used to create psychological profiles of people.
And second, AI – together with other emerging technologies – allows to manipulate information, to manipulate information, to a degree that was just unimagined before. It allows the artificial creation, e.g., of text content, video and sound that we cannot distinguish anymore from real information, video and sound.
So, if we bring those two AI capacities together, we can summarize that AI can analyze our psychological disposition and tailor information to exploit this psychological disposition. Therefore, we can be manipulated and our behavior can be influenced.
So, this is in strong contrast to personal autonomy. Personal autonomy means, in very simple words, that we as humans must be able to act without influence of someone else. And personal autonomy is, according to traditional western philosophy, the reason or the basis for why persons have dignity. Therefore, AI-enabled technologies make individual and general psychological and behavioral manipulation much easier, widespread, scalable, targeted and efficient. This goes against our personal autonomy and therefore, in my view, helps to violate our human dignity.

The second concerns data ownership – which I believe is no unfamiliar term to many of you. I do not have to go in great detail here. But just to maybe highlight one thing: In order to talk about data ownership, we have to understand what data is. Data is information. It is knowledge. And it is not entirely legally clear to whom this knowledge belongs. So, if the case can be made that data that we as humans produce actually does belong to us individually, this might allow us to say that if our data is collected, stored, used, mis-used, by governments, companies, private individuals, our right to property has been violated.

And third, I would like to name the use of AI-supported technologies in classification of people: Nowadays, AI-tech assists authorities in creating predictive criminal profiling, that is they calculate the probability that a person can become a criminal in the future, based on behavioral data of that person. In other words, individuals can be categorized as potential criminals based

on computational inferences. This, in my view, can be regarded as a general assumption of potential criminal conduct by all. Which is the opposite of the presumption of innocence, a human right (Art. 11 UDHR).

2. *What practical steps should we be taking to mitigate these risks, and to ensure that human rights are protected as AI technology advances?*

First of all, it is important to understand what is currently happening.
Currently, major tech companies are setting up what they call 'ethical principles', principles like 'data privacy', 'transparency' or 'accountability'. Those are standards that safeguard basic or human rights, rights that are – in many nations of the world – guaranteed by national constitution. Hence, classically, if a law, e.g., risked to violate those basic rights, it needed to pass through parliament. However, with regards possible violations of our basic rights by new technologies, tech companies are arguably a step ahead of governments, because it is them who are starting to set 'rules', let's say. And those ethical standards set up by the private sector are arguably based on competitive thinking and developed under time pressure of global business. So, whether we are talking of 'ethical washing' or real added value, meaning ethics as an end in itself, is a valid question.

So, what does this show us with regards to the HR-framework?
The HR-framework includes the principle of transparency. This principle requires that people must know and understand how major decisions affecting their rights are made. So, with regards to new technologies, what is affecting our rights does nowadays hardly, if not at all, pass through parliament anymore (1) and, therefore, governments cannot explain the potential rights violation because they are not even the origin of what affects our rights in the first place (2). To me, this is a highly problematic development.

In my opinion, we should not apply too much and too many AI-enabled technologies. I believe it is time to pause, to observe what is happening, and to reflect upon what is happening. We need to create humility and vision. The risks to human rights we named today are only a few, and the speed of development and the complexity of AI-enabled technologies might not allow us to see the whole risk spectrum yet. We have to think about those risks more first, and be wise – let's say – before rushing into action.

3. *What can we learn from more security-centered conversations about autonomous weapons in informing the way we think about the human rights implications of all types of AI?*

First, the debates on LAWS focus on traditional territorial state security. However, risks of emerging technologies are inherently local and citizen-based (see Q 1.) Therefore, it is important that the individual human being comes back into the epicenter of security concerns arising from new technologies.

And second, the debates on LAWS focus on their use during war, where, in simple words, it is International Humanitarian Law that is the main legal framework. The use of LAWS during law enforcement operations (e.g. so e.g. hostage situations or crowd control, or during anti-terrorist operations) is not discussed. In those situations, Human Rights Law is the main legal

reference point. And, arguably, fully autonomous weapons systems might violate the right to life, to personal and physical security, as well as human dignity.

4. *What are some of the potential risks posed by AI in the global south that may be underdiscussed or ignored in conversations which tend to center on large, economically successful democracies in the global north?*

Here I want to say two things:

Generally, as AI-enabled technologies pay off very, very well, political and economic interest in them is very, very high. So, as AI needs data, human data can be leveraged for economic and political interests, and arguably also for humanitarian purposes. E.g. when a government tries to attract a tech company that invests in AI by offering its citizens' medical data. Or when a great power trains its AI algorithms in the global south just to have a more diversified dataset. Or when refugees only receive humanitarian aid when they give away their biometric data. In other words, data geopolitics can lead to the exploitation of vulnerable communities – such as those from the global south. This can increase inequality – something that the international community obviously wants to reduce (SDG 10.)

More specifically, the use of AI in health care in the global south can be very risky. Why is that? AI-systems in health care, need to be trained on a large amount of medical data. And this data must represent the beneficiaries – the patients – very well. However, medical data in resource-poor areas of our world (global south), is often either very scarce, not available, very difficult to collect, or it is challenging to digitalize the data. This fact bears two different risks:

First, when AI systems are trained on faulty or 'bad' data, this might affect the quality and applicability of the system in reality. In the medical field, this might therefore affect patient care.

And second, as data in resource-poor environments, so the global south, is scarce, AI systems in health care are often trained by data generated by populations from the global north. However, it is scientifically well established that certain illnesses, e.g., just happen more often in one 'ethnic' group compared to another, naturally. Therefore, if AI systems are trained on health data that is not representative of the target group, in this case the global south, then they recreate the bias in the data, and have no or (highly) undesirable effects on the patients in the global south.

5. *Where do you see the greatest opportunities for AI to have a positive impact on people's lives, and what are the risks associated with those advancements?*

I think the greatest opportunity of the current and potential future AI development is the light it can potentially throw on ourselves. So, what do I mean by that? AI-enabled technologies allow us to outsource certain – classically inherently human – tasks to software. This allows us to give away control, creative action, and responsibility to technology. Think about LAWS: the control and responsibility for the most powerful and the most destructive human capacity – to kill another human being – can be outsourced to software. This may limit the space for human responsibility in the world.

So, new technologies offer us a moment to think about our fantastic human ability of creation, choice, decision, and responsibility for our creations, choices, and decisions – precisely because they risk us to (un)knowingly give them away.
So, if we are able to bring this very discussion out into the general public, it can give us the space and the power to reflect upon who we are as human beings with our creative abilities and choices. And – in a second step – to either choose to keep and safeguard those human abilities where we can – or not.

6. *What role, if any, do you believe that AI should play in governance processes themselves? What safeguards need to be put in place to ensure AI is not misused in these contexts?*

Here I would like to say two things:

With regards to safeguards against abuse and misuse: Any technology can be misused. Technology in itself is neither good nor bad. It is a human choice what to use it for. Therefore, if we criminalize or limit certain technologies, this could also limit their potential for Good.
So, I believe securing technological systems or criminalizing them will never be a practicable long-term solution. Because humans – or states – are either too curious or too afraid so that they still harm others, or too bored not to harm others. It is a problem of this strange cultivated nature of competition, power and harming. Therefore, the only long-term solution, in my view, is to work precisely on this inner attitude.

With regards to the role AI-supported systems should play in government: Governments – and arguably very few people in the world, with the exception of tech experts – do not know what AI really is or could be, and what its effects on the world are or could be. Therefore, it would be wise not to integrate AI in governance processes too soon, or not at all. Sometimes, it is better to pause and try to create general and public understanding, instead of rushing into action.

Human Rights as a Framework for AI Governance

1. *What do you see as the greatest advantage of the human rights framework as the foundation for governance of AI, especially in comparison to ethical frameworks?*

For the HR-framework, international and national institutional understanding is already there. This makes it a potentially useful concept for analysis in my opinion. Also, the tech community, especially research, has a long-standing reluctance, almost a resistance to 'being restrained' by ethical thinking. Maybe the HR-framework offers a more practicable reference point for analysis and policy shaping.

2. *What do you see as the greatest challenge in applying the human rights framework to the governance of AI?*

The main principles of the HR-framework, e.g. participation, accountability, transparency, require any governance instrument to be value-driven, human-centered & human-controlled.

This is, arguably, a strong contrast to towards where emerging technologies are currently developing.