Technology is not neutral
Pathways to negative outcomes
Impacts on peace and security

Made possible thanks to generous funding from the Republic of Korea
What are we talking about?

- A key question when we look at all new and emerging technologies and their impacts on peace and security is: **is technology neutral**? Really, we’re asking if impacts come only from the way a technology is used, or if they also stem from the technology itself.
- One view says that technology is fundamentally neutral. This is sometimes called the “Value Neutrality Thesis”. In this view, the negative impacts of technology come from the way people use it. A common example is a knife, which could equally be used to attack a person or to cut an apple[1]. Sometimes this view says that negative impacts can start in the design of technology, and sometimes only in its use.
- Most modern technology theorists **reject** the idea that technology is fundamentally neutral[2]. Instead, they generally recognize:

  1) the “intentionality” of technology: it is designed and deployed based on the values of the designer and deployer, then used according to the values of the user, and
  2) the “directionality” of technology: it adds choices or improves processes which point in a certain direction. For example, a gun can also be used as a paperweight, but everything in its design and development guides the user to a different use.

- With many technologies, including AI, specific innovations can create or exacerbate problems, **even without** any hostile intent. Algorithmic bias is an example of this issue. A translation tool that consistently assigns masculine gender to positive statements, and feminine gender to negative statements, is most likely demonstrating bias due to issues with the dataset it uses, not because of deliberate attempts by developers to produce a biased tool.
- The development of technology requires a degree of specialization in niche areas, that compartmentalizes the creators[3] and makes it difficult to grasp the consequences of a particular design decision (see also FS3 & FS4). **Responsible Innovation** (see also FS4) recognizes that these issues cannot be adequately addressed by individual actors and requires a multi-partner approach involving, for example, perspectives from the public, social scientists, lawyers & civil society organisations.

---

[3]This includes all those who work in the development phase
Pathways to negative outcomes

- To understand how AI can present risks for peace and security it is useful to look first at two of the primary pathways through which the use of AI can generate a negative impact in general.

  1. Consequences that come from intended use: these are the unplanned negative impacts that come from the regular use of an AI tool. The algorithmic bias example above, in which a translation tool demonstrates gender bias, illustrates this. The tool is being used in the way intended by its creators, and the production of gender-biased results is an unintended negative consequence.

  2. Consequences that come from unintended use: these are the negative impacts of an AI tool being used in a way its creators did not intend. A good example is the use of Generative Adversarial Networks (GANs) in the production of the “deepfake” for malicious purposes\[4\]. Use of GANs to support disinformation and cyber-crimes was clearly not the original intention of Ian Goodfellow and colleagues during their first development\[5\].

- There are multiple factors that can affect these risks. These include robustness, accessibility and vulnerability. The level of robustness of an AI system – that is the ability to maintain acceptable behaviour despite errors – impacts the risk of unintended consequences. The extent to which an AI is widely accessible (e.g. because it is open-source, low-cost, or a widely available commercial product) and vulnerable to hacking or adversarial attack can increase or decrease the risk of unintended use.

- When considering the consequences, it is useful to think of
  1. Both the immediate (intended and unintended) consequences of use, and the cascading second and third-order effects that stem from those consequences.
  2. The severity and reversibility of consequences. Are people’s well-being, fundamental rights and possibly lives at stake? Can the consequences be contained or mitigated, or would they have long-term and potentially irreversible impacts?

\[4\]https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/
Impacts on peace and security

- In the peace and security context, the two risk pathways mentioned may materialise in the following ways.
  - (1) **AI development intended for military use may have unintended consequences on peace and security.** For instance, bias in the targeting function of an autonomous weapons system (AWS) could cause the system to wrongfully attack civilians or civilian objects. Defensive and offensive AWS could interact in a way that could lead to rapid and unintended conflict escalation – possibly even a ‘flash war’. These are immediate consequences. The second or third-order effects are also important. For instance, a country that develops and adopts AWS to enhance its sense of security might generate insecurity on the part of its potential adversaries, who may react by developing their own capacities and establishing arms race conditions, which could be destabilizing for regional and/or global security.
  - (2) **AI development, whether intended for civilian/peaceful uses, or military uses, could be diverted and misused in ways that reinforce or create new threats to peace and security.** These include the use of AI tools for political disinformation, cyberattacks or terrorism.

- The first risk pathway is a concern primarily for the defence industry and the military. The second risk pathway, on the other hand, is an issue of concern for the AI community as a whole, given the dual-use potential of much AI development.
- Let’s, therefore, look more closely at the risks that flow from the diversion and misuse of AI.
- To understand how this risk might materialize and generate first, second and third-order effects on peace and security, it is helpful to consider a hypothetical example involving generative adversarial networks (GAN)

**Impacts on peace and security**

<table>
<thead>
<tr>
<th>Early academic research on autoencoders and generative adversarial networks (GAN)</th>
<th>Proliferation of knowledge and skills for development and use of autoencoders and GANs</th>
<th>Development and dissemination of first deepfakes without malicious intent and proliferation of access to tools</th>
</tr>
</thead>
</table>

- Reduced trust of news sources during an emergency leading to reduced civilian compliance with emergency procedures
- Skepticism to trust news sources in conflict settings, potential disorder in military command
- Anonymous development and dissemination on social media of deepfakes of a national leader during a conflict

- As we can see, in this example the second and third-order consequences of the development of a civilian AI technology and its dissemination include harms to peace and security. It is also important to think about the severity of that impact, who it has impacted, and the reversibility of the harm.
- If we think about the original developers of GANs, we can see AI can be used with hostile intent to harm peace and security but can also harm peace and security without specific hostile intent by the designer.
- Harm can simply come from the way an AI tool frames, guides or constrains behaviour, which brings us back to the ideas of “intentionality” and “directionality” of technology.
- In simple terms, choices in the development and diffusion of AI research and innovation can impact the risks of **diversion and misuse**, which in turn generate risks or opportunities for peace and security. These factsheets will go into detail in issues relating to the development of civilian AI and peace and security, including international governance and regulation, the political context, responsible innovation of AI, and standards, tools and practices.
Made possible thanks to generous funding from the Republic of Korea