Factsheet 4

Responsible AI for peace and security
What does responsible innovation mean?

How to engage in responsible innovation

Applying responsible innovation

Looking at a framework
Responsible innovation is a popular concept, but also one without a clear, consensus definition[1]. It is usually conceived as an anticipatory approach to technology governance which seeks to ensure that innovation can be used for good, and that possible negative impacts can be identified and mitigated in advance. It relies on principles, tools and processes that can help innovators act responsibly. These typically aim to provide guidance in the absence of certainty around both the impact of emerging science and technologies and the coverage of existing regulations. Responsible innovation is considered a useful approach to governing technological innovation because it is meant to be:

1. **Comprehensive:** it seeks to identify problems throughout the life cycle of a technology
2. **Inclusive:** it seeks to involve a diverse array of actors
3. **Technology specific:** it identifies and addresses risks in relation to specific technologies or areas of application
4. **Reflective and Preventive:** it seeks to identify and respond to problems before they occur
5. **Principle based:** by focusing on principles rather than formal regulation, it can help bypass the political deadlock of international technology governance

---

[1] There are a number of different approaches to responsible innovation. One example is, “Responsible Research and Innovation”, a model developed in the context of the European Union, that can be defined as "a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)."

[https://renovonschomberg.wordpress.com/definition-of-responsible-innovation/](https://renovonschomberg.wordpress.com/definition-of-responsible-innovation/) (RRI lead)
How to engage in responsible innovation

- There is no single recipe for responsible innovation. There are different sets of tools and processes that can be deployed at different stages of the research and technology lifecycles, which involve different constellations of stakeholders. These include ethical review processes, guidelines and principles, codes of conduct, industry standards, methodologies for impact assessment, and capacity-building & training mechanisms. There are, however, a few elements that are considered good practice:
  - **When** to engage in Responsible AI - Responsible innovation is intended to guide the entire life cycle of a technology: from conception, design and development to commercialisation and management, and even decommissioning. As such it is meant to be an iterative process, rather than a single action. Risk assessments, for instance, should be done at key junctures throughout the AI life cycle, not just at the beginning or end of development. Different stages in the life cycle may generate different concerns and risk management requires continual adaption and improvement.
  - **Who** should engage in Responsible AI - Responsible innovation "is a collective endeavor"[2]. It requires the involvement of a wide array of stakeholders, and not just the people involved in the conception, design, development and commercialisation of the technology (from engineers and product managers to those involved in ethical screening and internal compliance programmes). It is considered good practice for instance to involve external stakeholders. These may be social scientists (ethicists, lawyers), end-users or representatives of social groups that may be negatively impacted by the use of such technologies. External stakeholders are also key, as they can bring unique contextual knowledge that may not be apparent to technical or commercial teams. For example, when developing a facial recognition system, it would be important to reach out to communities poorly represented in the dataset, and those communities in which the system will be used – especially if they are different from the communities engaged in development. Engagement with external stakeholders can take a number of forms, including representation on review boards, public consultations, and other outreach.

Owen et al. p.46
How to assess and address risks. Existing methodologies and processes generally recommend three critical steps:

- **Baseline data collection and analysis** of the technology or decision that is to be risk assessed. This first step is intended to identify the inherent properties of the technology and the variables that are essential for carrying out the risk assessment.

- **Impact assessment.** This second step is intended to identify and characterize the spectrum of risks, along with the benefits associated with the development, diffusion and (mis)use of the technology. It usually requires the involvement of external stakeholders, who can help assess the negative impact that may arise from both the intended use as well unintended use of the technology.

- **Intervention.** This third and final step is intended to identify means of interventions. These may take the form of a) functional requirements for the design, development and diffusion of new research products or services; or b) means for engagement within/outside one workplace/community.

Image credit - canva text to image tool
There are already many responsible AI initiatives worth noting, some of which we provide here.

- The IEEE Global initiative on Ethics of Autonomous and Intelligent Systems (IEEE ethics in action). This effort connects academia, industry, and civil society, with deliverables that include:
  - Ethically aligned design report
  - IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being
  - IEEE standard model process for addressing ethical concerns during system design
  - IEEE ethics certification program for autonomous and intelligent systems
  - A call for action for business using AI

- The Partnership on AI, a private sector led initiative.
- The Global Partnership on AI, an international initiative to promote responsible AI use that respects human rights and democratic values.
- The work of the Montreal AI Ethics Institute, which focuses on democratizing AI ethics literacy
- These initiatives provide a solid collection of principles, tools and processes, but they generally provide little detailed guidance on how to identify and address risks connected to peace and security. Addressing this gap is the core goal of this work by UNODA and SIPRI.
Looking at a framework

- If we want to look at how this gap can be addressed, we can start by considering the work of the High-level Expert Group on AI (AI HLEG), convened by the European Commission. In 2019 the Group put forward recommendations on trustworthy AI based on three components – that it should be legal, ethical and robust (i.e. reliable and safe from a technical standpoint). We can use these three components as a conceptual framework for considering the elements of responsible innovation as they relate to international peace and security[3].

1. AI technology should be developed and diffused in line with the limits and requirements of international law – international humanitarian law and human rights law – as well as any national legal obligations, and commitments made in arms control and export control fora.

2. AI technology should be developed while keeping in mind ethical considerations, which, unlike legal requirements, are not usually captured in black and white. There are different approaches to ethics (e.g. consequentialist versus deontological), and views about ethical behaviour with regard to armed conflict and international security may also differ from one country to another and may vary over time. However, some points of consensus have emerged, as reflected by the 11 guiding principles of the LAWS GGE (see also FS2 and FS3). One of the key principles is that humans should remain responsible for decisions on deployment of weapons and the use of force. As a result, technology should be designed in a way that both allows humans to exercise moral agency and oversight; and prevents the creation of a gap in accountability by diffusing responsibility—it should avoid causing a situation where it is difficult to assign responsibility when something goes wrong.

3. Safety (here in line with “robustness”) should be a central consideration to reduce the risk of misuse and unintended use of high-risk AI systems. High-risk systems include safety-critical (or life-critical) systems, such as cars, and systems that have a major impact on critical infrastructure like power grids and healthcare. This means ensuring that systems are robust in their design and do not have inherent flaws that might, for example, make them more vulnerable to cyberattacks or adversarial attacks by malicious actors, or even simply increase the risk of accidents.

Looking at a framework

- If we look at this through the lens of AI that should be legal, ethical and robust, we can think about how to operationalise these components:
  - For legal use, the actor can engage directly with a compliance office, and establish if, and in what contexts, use would be legal.
  - For ethical use, the actor can employ an ethics review board. This is a well-established model in other fields (especially healthcare and academia), that is also frequently employed for AI. This is a team of experts employed both to review proposals and to make risk-mitigation recommendations to researchers and product developers.
  - For robust use, the actor should engage in thorough Testing, Evaluation, Validation and Verification (TEVV).

  This is already a key concept for AI development, as well as in the development of non-AI military systems, but presents specific challenges for peace and security. TEVV is vital for ensuring that AI applications are deployed safely and robustly. A challenge in this regard is that it is not currently clear that there are any proven effective methods for testing nondeterministic (less predictable and more adaptable) systems[4]. This is therefore an issue that militaries, as well as the private sector, take very seriously, and an area that requires significant consideration, investment, and improvements in capacity.

  It is also important to identify and consider the spectrum of peace and security risks associated with the development, diffusion and (mis)use of the technology in question. This involves exploring the pathways through which the development, diffusion and (mis)use could generate negative effects; and what these negative effects would be (who/what would be impacted, when and where). Proper consideration of this kind requires exploring two different risk pathways independently – 1) effects that may flow from research and innovation’s intended use, and 2) consequences of intended and unintended misuse (see also FS1).
Made possible thanks to generous funding from the Republic of Korea