Council of the European Union

Brussels, 26 October 2022
(OR. en)

14095/22
ADD 1

LIMITE

CORLX 995
CFSP/PESC 1432
CONUN 254
CODUN 48
CONOP 112
COTER 256
COARM 217

PROPOSAL

From:	High Representative of the Union for Foreign Affairs and Security Policy, signed by Mr Stefano SANNINO, Secretary-General
date of receipt:	26 October 2022
To:	Ms Thérèse BLANCHET, Secretary-General of the Council of the European Union
Subject:	Proposal of the High Representative of the Union for Foreign Affairs and Security Policy to the Council for a Council Decision in support of the implementation of a project "Promoting Responsible Innovation in Artificial Intelligence for Peace and Security"

Delegations will find attached document HR(2022) 238.

Encl.: HR(2022) 238

EUROPEAN EXTERNAL ACTION SERVICE

PROJECT DOCUMENT
Promoting Responsible Innovation in AI for Peace and Security
HR(2022)238

1. BACKGROUND

Recent advances in artificial intelligence (AI) have unlocked new possibilities to support and sustain peace and security, for instance, through technological improvements in areas such as conflict early warning, and arms and export control verification. On the other hand, these advances have enabled new means to generate—or aggravate—tensions, conflicts and insecurity between and within states. The risks posed by certain applications of AI, such as lethal autonomous weapons systems, have emerged as a major issue of concern for the arms control community. One risk pathway that deserves greater attention, and which current arms control and diplomatic efforts may be insufficient in responding to, is the diversion and misuse of civilian AI research and innovation by irresponsible actors, including malicious non-state actors, such as the misuse of Generative Adversarial Networks (GAN) to produce 'deepfakes' for disinformation campaigns.

AI is an enabling technology with great general-use potential. AI research and innovation that are developed for civilian applications could be (relatively easily) accessed and repurposed by certain actors for harmful or disruptive uses that could have implications for international peace and security. The diversion and misuse of civilian technology are not new phenomena nor are they unique to AI. In the related field of robotics, this was recently exemplified by the weaponization and use of recreational drones by Daesh/ISIS in Syria. But in the case of AI, the problem is complicated by multiple factors: the intangible and fast-changing nature of AI algorithms and data, which makes the transfer/proliferation of these difficult to control; the leading role of the private sector in the research, development and innovation ecosystem, and its consequent need to protect proprietary algorithms; and the global availability of the human expertise and material resources capable of repurposing AI technologies. Meanwhile, those working in AI in the civilian sector remain too often unaware of the potential implications that the diversion and misuse of their work could have for international peace and security or are hesitant to take part in the existing discussions on AI risks in arms control and non-proliferation circles.

There is a need to support greater engagement of the civilian AI community in understanding and mitigating the peace and security risks associated with the diversion and misuse of civilian AI technology by irresponsible actors. For the Stockholm International Peace Research Institute (SIPRI) and the United Nations Office for Disarmament Affairs (UNODA), this could be achieved through greater promotion of responsible innovation in the global civilian AI community. Past work by SIPRI and UNODA has shown that responsible innovation, as a self-governance mechanism, could provide the global civilian AI community with practical tools and methods to identify as well as help prevent and mitigate the risks that the diversion and misuse of civilian AI research and innovation could pose to peace and security. SIPRI's and UNODA's work also identified methodologies and several ongoing civilian-focused initiatives on responsible AI that could be built on to sensitize the civilian AI community to arms control and non-proliferation issues, expert debates and state positions on responsible development, diffusion, and use of AI, as well as lessons to be learned from defence sector responsibility work.[1] Critically, this prior work has clearly identified engagement with Science, Technology, Engineering and Mathematics (STEM) students, who are still engaging with AI in an educational format, as central to any effective responsible innovation effort.

2. OBJECTIVES

These projects aim to support greater engagement of the civilian AI community in mitigating the risks that the diversion and misuse of civilian AI research and innovation by irresponsible actors may pose to international peace and security. They aim to do so by, firstly, generating greater understanding of how decisions in the development and diffusion of AI research and innovation can impact the risks of diversion and misuse, and in turn generate risk or opportunities for peace and security, and secondly, by promoting responsible innovation processes, methods and tools which can help ensure the peaceful application of civilian innovations and the responsible dissemination of AI knowledge. To this end, they support capacity-building, research and engagement activities that will *i*) enhance the capacity within the global civilian AI community to include and address the peace and security risks presented by the diversion and misuse of civilian AI by irresponsible actors through responsible innovation processes; and *ii*) strengthen the connection between risk mitigation efforts in responsible AI in the civilian sphere with those already ongoing in the disarmament, arms control and non-proliferation community at an intergovernmental level. Crucially, they do not intend to establish any new standards, principles, or regulation, or otherwise step into areas within the competence of States. Instead, they intend to develop civilian responsible innovation efforts to include peace and security risks presented by the diversion and misuse of civilian AI by irresponsible actors, and provide education on existing relevant intergovernmental efforts.

To effectively reach and impact the civilian AI community, the projects deploy a three-pronged approach, seeking to

[1] Methodologies include for instance the Institute of Electrical and Electronics Engineers (IEEE) recommended practices for assessing the impact of autonomous and intelligent systems on human well-being (IEEE Std 7010-2020), The High-Level Expert Group on Artificial Intelligence' Assessment list for Trustworthy Artificial Intelligence (ALTAI). Initiatives include: The IEEE The global initiative on ethics for autonomous and intelligent systems; the Partnership on AI; the Global Partnership on AI.

(a) *engage with educators* – work with selected educators and developers of academic curricula on the development and promotion of educational materials that can be used to mainstream consideration of the peace and security risks that flow from the diversion and misuse of civilian AI research and innovation by irresponsible actors in the training of future AI practitioners (e.g. in courses on AI ethics and responsible innovation);
(b) *engage with students* – introduce selected Science, Technology, Engineering and Mathematics (STEM) students from around the world to how the peace and security risks posed by the diversion and misuse of civilian AI development by irresponsible actors may be identified, prevented or mitigated in the research and innovation process or through other governance processes; and
(c) *engage with the AI industry* – work with professional associations and standards bodies like the Institute of Electrical and Electronics Engineers (IEEE) to i) disseminate tailored education materials and engagement activities to technical professionals; ii) support positive uses of AI for peace and security; and iii) facilitate dialogue and information sharing between experts from academia, the private sector and government on how the risk of the diversion and misuse of civilian AI research and innovation by irresponsible actors can be mitigated.

Such an approach allows the projects to reach the AI community at all levels, including not only current practitioners but also future generations. It also enables engagement across academic, industry and other silos, and supports the sustainability of future efforts by establishing networks that cross these boundaries.

The projects also seek to employ the convening power and experience of SIPRI and UNODA to impact the AI community globally, not just EU stakeholders. SIPRI and UNODA are uniquely positioned to reach and facilitate engagement between AI actors from across Africa, Asia-Pacific, Europe, and North and South America. Both entities also have experience working in other fields of science and technology facing similar challenges of dual use and proliferation, including biotechnology. The projects also seek to take advantage of conditions present within the European Union, such as a) the existence of advanced multi-stakeholder processes on responsible AI; b) the high level of engagement in, and expertise on, disarmament, arms control and non-proliferation issues in the EU; c) the diversity of connections that academic, research and private sector organizations in the EU have with other regions, notably in the Global South, which will also be a major target for engagement; and d) the diversity of nationalities of students, educators, and engineers in universities, research institutions and the private sector.

Inclusion will be a core consideration for the conduct of the projects' activities. To effectively support the AI community, the projects recognize that the AI community consists of a diverse array of actors, and in particular that

(a) gender is a highly relevant factor. For this reason, gender will be mainstreamed in line with the UN system-wide gender mainstreaming and parity strategies. Participation of women in all activities under the project will be encouraged and required; and
(b) inclusion of persons with disabilities and the reasonable accommodation of needs will be carried out throughout. This will include the addressing of obstacles to the participation

of persons with disabilities as well as ensuring that steps are taken to engage with and facilitate the representation of the substantive views and experiences of persons with disabilities.

3. PROJECTS

The three projects described below are intended to be complementary and mutually supporting, with elements running throughout the 36 months.

Project 1 – Production of education and capacity-building material for the civilian AI community

3.1.Project purpose

Project 1 focuses on providing the knowledge and means for civilian AI actors to evaluate and mitigate the risks that the diversion and misuse of civilian AI research and innovation by irresponsible actors may pose to international peace and security. It aims to produce education and capacity-building material that will provide AI practitioners from all regions, levels and sectors (including AI-focused educators, curriculum developers, STEM students and AI engineers and researchers in academia and the private sector) with the information and tools necessary to

(a) understand how civilian AI research and innovation could be diverted and misused in ways that could present risks to international peace and security and how decisions in the development and diffusion of research and innovation can increase or decrease the risk of diversion and misuse;

(b) understand the efforts already engaged in by the disarmament, arms control and non-proliferation community to mitigate the risks of the diversion and misuse of civilian research and innovation; and

(c) practice responsible innovation in a way that mitigates the risk of diversion and misuse in the development and diffusion of research and innovation.

3.1.1. Project description

This project will produce three separate sets of education and capacity-building materials.

(a) *Handbook (1)* – The handbook will compile basic knowledge and means for AI actors to evaluate and mitigate, in the research and innovation process, the risks of the diversion and misuse of civilian AI technology by irresponsible actors. It will discuss why and how decisions around the development and diffusion of research and innovation can impact the risks of diversion and misuse, and in turn generate risks or opportunities for peace and security. It will also introduce relevant international law and export control obligations in addition to safety and security considerations under discussion in military as well as disarmament, arms control, and non-proliferation circles; and present example processes and tools to practice responsible innovation, such as technology impact assessment methodologies and risk-assessment templates.

(b) *Podcast series (~10)* – These podcasts will act as an accessible and engaging medium for AI actors to learn about why and how responsible AI innovation processes can support international peace and security through the mitigation of risks presented by diversion and misuse from irresponsible actors. The series will review important themes (e.g., the pattern of diversion and misuse of dual/general-use research and innovation; humanitarian, strategic, and political challenges associated with the potential misuse of civilian AI research innovation; challenges that disarmament, arms control, and non-proliferation circles face in risk-mitigation efforts; how to do responsible innovation through risk assessment; export control compliance; risk reduction by design; responsible publishing; knowing your customers; and experience of tabletop exercises) and will be structured around interviews that the project team will conduct with representatives from relevant communities.

(c) *Blog series (9-10)* – The team will develop a curated blog post series aimed at raising the profile of efforts that try to cross boundaries between the civilian-focused 'responsible AI' and arms control and non-proliferation communities. The blog series will provide a platform to disseminate insights, ideas and solutions regarding the identification and addressing of risks associated with the diversion, and misuse of civilian AI in the research and innovation process. The blog will seek to represent the diversity of thought and perspectives present in the AI sector.

These materials will be disseminated publicly through the websites of the implementing actors, their social media presence, and through direct communication with relevant academic entities, civilian AI professional associations and other appropriate groups.

3.1.2. Expected results of the project

This project is expected to establish a new set of materials by which civilian AI practitioners can become sensitized to a) how civilian AI research and innovation could be diverted and misused in ways that may present risks to international peace and security, b) how such risks are being addressed by the disarmament, arms control, and non-proliferation community, and c) how AI practitioners could further contribute to the mitigation of such risks through responsible innovation processes.

This is expected to advance the engagement of the civilian AI sector in mitigating the risks that the diversion and misuse of civilian AI may pose to international peace and security; improve the capacity of technical practitioners to engage with relevant processes in disarmament, arms control, and non-proliferation community as well as support the engagement of new audiences not traditionally included in disarmament and non-proliferation education efforts.

The material is also expected to support the implementation of the other projects and will serve as a basis for the educational and capacity-building activities in Project 2 as well as dialogue and engagement activities in Project 3. These activities are expected in turn to feed back into the production and refinement of the material. Such an iterative approach is expected to help address potential obstacles to their promotion, diffusion and use within the AI community, including

issues related to language, content, context and availability, which could preclude their impact at the global level, particularly in the Global South.

3.2. Project 2 – Education and capacity-building activities for future AI practitioners

3.2.1. Project purpose

The purpose of Project 2 is to support the integration of the problem of the diversion and misuse of civilian AI research by irresponsible actors in the education of future generations of AI practitioners. In the long term, this will ensure that the Science, Technology, Engineering and Mathematics (STEM) students shaping the future of AI are aware of the negative impacts that the diversion and misuse of their work by irresponsible actors could have on international peace and security and that they will have the basic tools necessary to identify and mitigate such a risk in the research and innovation process.

This project will conduct a series of educational and capacity-building workshops with educators and students in collaboration with selected international universities and industry actors. The project thereby seeks to develop capacity-building activities that educators and developers of academic curricula could use to include in the training of future AI practitioners (e.g. courses on AI ethics and responsible innovation) and considerations for the risks of diversion and misuse of civilian AI research and innovation by irresponsible actors and connect those to the larger peace and security context. Through these workshops, the project will also seek to identify a network of interested educators, curriculum developers and students who would support the dissemination and promotion of the project education material and capacity-building activities in the AI education community and AI practitioner community. This networking component seeks to ensure the sustainability of the projects beyond their immediate duration and to enable the building of stronger links in support of civilian technical engagement towards larger peace, security, disarmament and arms control goals.

3.2.2. Project description

This project will conduct a series of educational and capacity-building workshops with educators and students from selected universities from around the world. These would consist of a mix of lectures and interactive activities that will provide educators and students with opportunities to reflect on how civilian AI research and innovation could be diverted and misused in ways that may present risks to international peace and security and how on such risks can be identified, prevented or mitigated in the research and innovation process, or through other governance processes. These activities will build on prior smaller-scale pilot work carried out by UNODA which experimented with methods to engage with and sensitize STEM students to the importance of considering the wider impact of their work as well as engage with expertise outside of their home fields. Concretely, these would consist of

(a) *regional capacity-building workshops for educators and students (4)* – the regional workshops will conduct and promote activities that educators can use to build STEM students' capacity in responsible AI innovation with a particular focus on how to evaluate

and mitigate the risks of the diversion and misuse of civilian AI technology by irresponsible actors. Each workshop will be organized with an EU-based university and a high-profile university from a different global region, thereby always connecting a diverse set of EU based participants with a diverse set based outside of the EU. The workshops will then cover Latin America and the Caribbean, North America, Africa and Asia & the Pacific. This will allow the participation of students (at the master's and PhD levels) from across the world, including from the Global South. The workshop would be conducted primarily in English, but where feasible, participants would be provided with the opportunity to engage in activities based on alternative language groupings; and

(b) *international workshop on sustainable capacity building (1)* – the workshop will draw on lessons learned from the regional workshops and facilitate the exchange of information and experiences between educators and selected students from the universities involved in the project. The workshop would discuss how to refine the activities and tools elaborated over the course of the project, and disseminate them beyond the group of participating universities. It will also discuss how to support the engagement of students in responsible AI that addresses diversion and misuse risks for international peace and security once they have entered the workforce.

SIPRI's and UNODA's networks and presence in Africa, Asia-Pacific, Europe, North and South America will be used to facilitate and support aspects of the activities as appropriate.

3.2.3. Expected results of the project

The project is expected to create models of capacity building and engagement activities that educators and developers of academic curricula could replicate to sensitize future AI practitioners to the problems of the diversion and misuse of civilian AI by irresponsible actors and how they can help mitigate these problems through responsible innovation processes. After completing the project activities, the participants (educators, but also STEM students) will be expected to be able to use and promote responsible innovation tools, methods and concepts to identify and mitigate the risks of diversion and misuse in the development and diffusion of civilian AI research and innovation.

The project activities are also expected to generate a network of educators, curriculum developers and students who would not only promote the project activities within the AI education and professional communities (e.g. during conferences of the IEEE Computational Intelligence Society) but also be available to contribute with technical capacity to state-led international governance processes (e.g. the Convention on Certain Conventional Weapons' process on emerging technologies in the area of Lethal Autonomous Weapons Systems).

The short- and long-term value of these activities will be demonstrated through pre and post activity surveys.

3.3. Project 3 – Facilitating the longer-term sustainable development, dissemination, and impact of responsible innovation in AI for peace and security

3.3.1. Project purpose

The purpose of Project 3 is to facilitate the longer-term sustainable development, dissemination and impact of responsible innovation in AI as a means of mitigating the risks that the diversion and misuse of civilian AI research and innovation may pose to peace and security. It aims to do so through roundtables with the AI industry, multi-stakeholder dialogues, the creation of a public report, and targeted dissemination activities. The project aims to ensure that the work generated, particularly the education, capacity-building and engagement activities, reaches and impacts the AI community at large, at all levels (from students to engineers and other AI professionals) and across geographical, sectorial, and other boundaries. To increase the possibility to make a broad and deep impact, it is essential to cooperate with professional organizations in this space, such as the IEEE, and conduct multi-dimensional engagements across academia, industry and other silos. Such efforts will give the opportunity to interested representatives from different AI communities to take ownership of the problem and provide their own views about how risk-mitigation efforts may be carried out and promoted sustainably within and across the global AI community. It is also important for the long-term value of the project to states, intergovernmental organizations and others that AI practitioners can learn from and engage with governmental experts engaged in risk mitigation in the disarmament, arms control and non-proliferation context. It is also critical for sustainability to ensure that the insights generated through the engagement activities are analysed, consolidated and disseminated appropriately.

3.3.2. Project description

This project consists of these key strands:

(a) *multi-stakeholder dialogues on 'responsible AI innovation for peace and security' (Up to 9)* – this series of virtual dialogue meetings would bring together experts from academia, research, the private sector, and traditional arms control from the EU and beyond to discuss

i. technological trends that may generate diffusion, diversion and misuse risks with impacts for international peace and security;
ii. how to engage in risk mitigation through responsible innovation processes, methods and means, and opportunities and challenges for dialogue and knowledge sharing between stakeholder communities, including those operating in other sectors such as the biological and chemical industries; and
iii. the potential value, purpose, and format of a self-sustaining network of experts and dialogue activities. The expert group will meet several times per year and work towards the organization of two public events for the wider community.

Of the nine virtual meetings, two are intended to be open to the public, in order to facilitate broader consultation.

(b) *private sector roundtables (Up to 6)* – this series of virtual roundtables will initiate a dialogue with actors working with responsible AI innovation processes in the private sector (e.g. Partnership on AI) as to how they can contribute to minimizing the risks of the diversion and misuse of civilian AI technologies by irresponsible actors, as well as explore possible incentives within private sector development for doing so. Topics will include

i. the relevance of the international security and disarmament context for the private sector;
ii. the legal environment(s) in which AI development, deployment and operation exists around the world;
iii. how to build on risk-assessment mechanisms and other measures that are part of, or could be integrated into, responsible innovation processes and corporate compliance programmes; and
iv. lessons to be learned from other industries, processes and frameworks relating to arms control (e.g. biological, and chemical industries).

(c) *report on AI community perspectives on arms control and risk mitigation in AI, targeted at AI and arms control communities (1)* – the development of this report will capture and consolidate the findings and recommendations of the project into a single reference document aimed at both the civilian-responsible AI and the arms control communities. The report would discuss how international peace and security risks associated with the diversion and misuse of civilian AI research and innovation can be identified, evaluated and addressed.

(d) *dissemination events targeted at consultation and engagement with AI and arms control communities (tbd)* – the team will seek opportunities to communicate the work and its findings and recommendations throughout the project. The format of the events and content of the presentations would be tailored to the needs of the target groups. These may include meetings of CONOP, the European AI Alliance Assembly; the Group of Governmental Experts on emerging technologies in the area of Lethal Autonomous Weapon Systems; the Inter-Agency Working Group on AI (IAWG-AI); ITU's initiative AI for Good; UNIDIR's Annual Innovation Dialogue; and the Institute of Electrical and Electronics Engineers. The team would also seek to engage bilaterally with relevant stakeholders from government, academia and the private sector.

3.3.3. Expected results of the project

This project is expected to set the foundations for the sustainable development, dissemination and impact of the responsible innovation of AI processes addressing diffusion, diversion and misuse risks and their implications for peace and security beyond the immediate duration of the council decision.

The multi-stakeholder dialogue is expected to provide a model for information sharing and collaboration on risk mitigation not only within the global AI community but also between the civilian-responsible AI community and the disarmament, arms control, and non-proliferation

communities. Such a model could be used to familiarize policy makers with key technological and scientific advances relevant to the responsible innovation of AI and also familiarize technical audiences with the environment in which policy makers are currently engaging. The project is expected to facilitate sustainable relationships and engagement between interested actors within and across these different communities. Such heterogenous network effects are expected to enable the greater development and widespread promotion of the responsible innovation of AI for peace and security beyond the timeframe of the project.

The private sector dialogue is expected to enable greater and deeper engagement of the AI private sector in the identification, prevention and mitigation of the peace and security risks stemming from the diversion and misuse of civilian AI research and innovation. The project is expected to give key actors in private sector processes greater understanding and ownership of the problems it seeks to address. In addition, it aims to facilitate the (wider) adoption and implementation of responsible innovation processes, methods and means in existing corporate risk management mechanisms and procedures.

The multi-stakeholder dialogue and the private sector roundtable are also expected to generate insights on a number of substantial issues, including a) how responsible innovation methods and means can be further refined and deployed to identify, prevent and mitigate risks posed by the diversion and misuse of civilian AI research and innovation; b) how AI research and innovation can be used positively to support peace and security objectives (e.g. applications for conflict early warning and humanitarian assistance); and c) how to facilitate greater dialogue and information sharing between risk mitigation efforts undertaken in the civilian-responsible AI community (i.e. various IEEE-led initiatives) with those already ongoing in the disarmament, arms control and non-proliferation community at an intergovernmental level.

The report and dissemination activities will analyse, consolidate and disseminate insights generated through Projects 1, 2 and 3, and thus support the promotion of the findings and recommendations of the project activities within the global AI community, as well as within the policy community. They are also expected to help ensure sustainability of impact beyond the timeframe of the projects.

4. DURATION

The total estimated duration of the projects' implementation is 36 months.