Zurich, November 2-3 2022, Digital Society Initiative

Workshop "Controlling Autonomous Systems in the Security Sector – empirical, ethical and legal considerations"

In this workshop, recent results on studies concerning the control of autonomous systems used in the security sector will be discussed. Linda Eggert from the Institute for Ethics in AI (Oxford University) and Giacomo Persi Paoli, Head of the Security and Technology Programme at the United Nations Institute for Disarmament Research (Geneva) will serve as keynote speakers.

The workshop takes place at the Digital Society Initiative, Rämistrasse 69, 8001 Zurich, Switzerland.

If you want to participate, please use the registration link below. Here you can indicate in which parts of the workshop you want to participate.

https://ufspezurich.eu.qualtrics.com/jfe/form/SV_dmqKvCScnx1ZxAy

Program

Wednesday, November 2

12:30 – 13:00 Arrival of participants

13:00 – 13:30 Introduction

Welcome address DSI (Avi Bernstein, DSI Director)

Welcome address Armasuisse & Introduction into workshop theme (Pascal Vörös)

13:30 – 15:00 Block 1: Empirical studies (30' each; 20' talk & 10' Q&A)

Military expert and lay people opinions on security AI systems (Markus Christen)
In this talk, we will present findings from two surveys that address responsibility attribution and trust when using AI systems in security contexts and beyond. In a representative survey of the Swiss population, we find that context has a stronger impact on trust compared to responsibility attribution when either humans or AI systems are involved in decision tasks. Moral responsibility remains with human actors, irrespective whether the human actor has an operative or a supervision function; but the overall "responsibility load" for humans is increased when AI systems are involved. Those results are more pronounced when Swiss security actors are approached (second survey). In the second survey, we also present findings, what kind of AI system functions are considered more or less desirable for the Swiss Armed Forces.

Understanding human-AI-interaction (Abraham Bernstein)
In this talk, findings of two behavioral experiments are presented. In the first study, we investigate how the expert type (human vs. AI) and level of expert autonomy (adviser vs. decider) influence trust, perceived responsibility, and reliance. We find that participants consider humans to be more morally trustworthy but less capable than their AI equivalent. This shows in participants' reliance on AI: AI recommendations and decisions are accepted more often than the human expert's. However, AI team experts are perceived

to be less responsible than humans, while programmers and sellers of AI systems are deemed partially responsible instead. In the second study, the effect of breaking and rebuilding trust when interacting with an AI system has been investigated. We found evidence suggesting that trust development is a slow process that evolves over multiple sessions, and that first impressions of the intelligent system are highly influential.

Pitfalls when controlling AI systems (Serhiy Kandul)
Effective human control over AI relies on human ability to predict AI's behavior, e. g. to recognize when AI is likely to commit an error, which would allow humans to intervene and overrule the AI decision if necessary (Art. 14 of EU AI Act proposal). In this talk we present results of an experiment, where we asked participants to predict whether a specific operation conducted either by AI or by a human operator will be successful. We find that 1) AI's predictability is lower than human predictability and 2) participants are not aware of the difference.

15:00 – 15:30 Break

15:30 – 17:00 Block 2: Normative studies (30' each; 20' talk & 10' Q&A

Eight Recommendations for the Ethical and Legal Assessment of Robotic Systems Interacting with Humans (Thomas Burri)
In this talk, we present eight Recommendations with a view to improving the ethical and legal assessment of robotic systems in the security sector, i.e., defence, law enforcement, and disaster relief. We have drafted these Recommendations on the basis of our experience and research over the last years. The Recommendations should guide developers, lawyers, ethicists, and policymakers who are involved in one way or another with robotic systems in the security sector and have to navigate a highly technical and dynamic field. Taken together, our eight Recommendations boil down to the following: When technology runs away, normative discussions are in gridlock, and uncertainty prevails, adopt a pragmatic, practical, and dynamic form of applied ethics that eschews high politics and ultimately produces experience that can be fed back into high-level discussions and move them forward.

Command & Control in Mixed Human-Machine Teams: Future Human-Robot Interaction (Samuel Huber)
The advent of autonomous robots will have the potential to dramatically change human-machine interaction as we know it today. Up until now, robots have been used as tools, directly controlled and closely monitored by their human operators. Robots capable of planning and deciding on their own could be integrated in mixed human-robot team in the future. We asked ourselves which factors are important for the success of such mixed human-robot teams. We would like to introduce a Framework that integrates the most important of these factors.

The Ethical Assessment of Autonomous Systems in Practice (Daniel Trusilo, online)
There is a growing body of ethical AI principles for defense. Though the principles are an important step toward the responsible use of systems with autonomous capabilities, it is also essential to operationalize the principles for real-world systems. In this talk, I will discuss the real-world application of a tool developed to assess robotic systems designed for security applications that have a degree of autonomy.

17:00 – 17:30 Break

17:30 – 19:00 Keynotes, followed by panel (30' each talk)

Rethinking 'Meaningful Human Control' (Linda Eggert, University of Oxford)
Three principal concerns raised by autonomous weapons systems (AWS) are (1) that AWS may not be able to comply with the laws of war; (2) that delegating life-and-death decisions to algorithms violates human dignity; and (3) that ascribing responsibility for wrongful harms may become impossible. A recurrent response is that AWS must, like other weapons, remain under 'meaningful human control' (MHC). In this talk, I examine how far appeals to MHC can take us in addressing each of the three concerns. I suggest that the justificatory force of MHC is significantly more limited than typically assumed, and that we should rethink the role MHC plays in debates about the ethics of AWS.

Military applications of AI and Autonomy at the UN: the past, the present and the future (Giacomo Persi Paoli, UNIDIR)
The presentation will provide an overview of the discussions held within the UN, specifically in the GGE on Lethal Autonomous Weapon Systems. We will start by contextualizing the GGE LAWS discussion in the broader picture of current international efforts related to the establishment of shared principles for the use of AI. We will then move to look the GGE on LAWS and do an historical recap of its deliberations, focusing on the most important milestones, before analyzing the situation as it is today. We will then conclude by reflecting on possible avenues for the future for both the GGE on LAWS and wider UN efforts related to the use of AI in the security sector.

Panel with Linda Eggert and Giacomo Persi Paoli, moderated by Markus Christen

19:30 – 22:00 Dinner (invited only)

Thursday November 3

08:30 – 09:00 Arrival

09:00 – 12:00 Internal workshop: What should we know next? (flexible break)

12:30 Lunch