UNIDIR
UNITED NATIONS INSTITUTE
FOR DISARMAMENT RESEARCH

UNIDIR invites you to the side event

Autonomous Weapon Systems:

Understanding Bias in Machine Learning and Artificial Intelligence

5 October 2017, 13h15–14h45, Room 12, UN Headquarters

Algorithmic bias has become a topic of attention in the societal conversations on **machine learning and artificial intelligence**. In the past months, several instances of **algorithmic bias** have made the news — such as misclassification of humans by image recognition algorithms — and major expert-led initiatives, such as AI Now, have identified bias as an area requiring much deeper attention and understanding.

Increasing autonomy in weapon systems is the focus of growing international attention, as are the implications of artificial intelligence for international security. The algorithms that make increasing autonomy in weapon systems possible are not spared from bias. Therefore, it is critical to develop a better understanding of **how biases influence outcomes** in learning systems. What can we learn about bias from other fields where decisions with significant human impact are already made by learning algorithms? What do we already know about **detecting bias**—both unintentional and intentional? How could we know in which ways algorithms are biased? **Is all bias bad**? And are there specific issues concerning bias that we need to be mindful of in relation to discussions at the upcoming GGE on Lethal Autonomous Weapon Systems?

Featured panellists:

Cathy O'Neil on understanding algorithmic bias and its significance

David Danks on bias issues that are distinctive to autonomous weapon systems

Respondent: **Ambassador Amandeep Singh Gill**

Followed by a discussion moderated by **Kerstin Vignard**, Deputy to the Director, UNIDIR

About the experts

Cathy O'Neil earned a Ph.D. in math from Harvard, was a postdoc at the MIT math department, and a professor at Barnard College where she published a number of research papers in arithmetic algebraic geometry. She then switched over to the private sector, working as a quant for the hedge fund D.E. Shaw in the middle of the credit crisis, and then for RiskMetrics, a risk software company that assesses risk for the holdings of hedge funds and banks. She left finance in 2011 and started working as a data scientist in the New York start-up scene, building models that predicted people's purchases and clicks. She wrote *Doing Data Science* in 2013 and launched the Lede Program in Data Journalism at Columbia in 2014. Cathy is the author of *Weapons of Math Destruction: how big data increases inequality and threatens democracy*.

David Danks is the L. L. Thurstone Professor of Philosophy & Psychology, and Head of the Department of Philosophy, at Carnegie Mellon University. He works at the intersection of philosophy, cognitive science, and machine learning, integrating ideas, methods, and frameworks from each to advance our understanding of complex, cross-disciplinary problems. Most recently, Danks has used interdisciplinary approaches to address the human and social impacts when autonomous capabilities are introduced into technological systems, whether self-driving cars, autonomous weapons, or healthcare robots. His work is both theoretical and practical, including collaborations with industry groups and government agencies. His earlier work on computational cognitive science resulted in his book, *Unifying the Mind: Cognitive Representations as Graphical Models*, which developed an integrated cognitive model of complex human cognition.

Ambassador Amandeep Gill is Permanent Representative of India to the Conference on Disarmament and Chairman of the Group of Governmental Experts on Lethal Autonomous Weapons Systems. He has served abroad at the Indian Mission to the Conference on Disarmament in Geneva, the Indian Embassy in Tehran, the High Commission of India in Colombo and the Indian Mission to the UN in Geneva. At headquarters in New Delhi, he has served thrice in the Disarmament and International Security Affairs Division as well as in the United Nations Division. He served as the Head of the Disarmament and International Security Division in the Ministry of External Affairs from 2013 until recently. He has also served as an expert on the UN Secretary General's panels of experts on the Fissile Material Cut-off Treaty (FMCT), Small Arms and Light Weapons and on Missiles. Amandeep Gill trained as an electronics engineer; his Ph.D. from King's College London is on Learning in Multilateral Forums.

About the Project

In 2013 UNIDIR launched its work on the weaponization of increasingly autonomous technologies, bringing together expertise on conventional weapon systems, ICTs and cyber operations, and artificial intelligence. This project focuses on advancing the nascent multilateral security discussion by refining the areas of concern, identifying relevant linkages, and learning from approaches from other domains that may be of relevance to this topic. The project's primary aim is to provide insights and conceptual frameworks that will enable policy-makers to better think, discuss and make informed decisions about autonomy in weapon systems.

Currently in its third phase, the project is promoting practical understanding among policy-makers of the potential challenges raised by increasingly autonomous technological capabilities — and in particular learning systems — in the near to medium term. To achieve this objective, Phase III is built upon three inter-connected pillars: supporting the GGE discussions starting in CCW; developing tech-gaming scenarios that help policy makers think through the implications of near-term technological progress on the developmental trajectory of autonomous weapon systems; and cross-disciplinary expert group meetings on learning systems, biases, and artificial intelligence.

For more information, visit http://bit.ly/UNIDIR_Autonomy

Related work

Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies

This paper offers an introduction to unintentional risks related to increasing machine learning in weapon systems. Recent international attention on autonomous weapon systems has focused on the implications of what amounts to a 'responsibility gap' in machine targeting and attack in war. As important as this is, the full scope for accidents created by the development and deployment of such systems is not captured in this debate. It is necessary to reflect on the potential for AWS to fail in ways that are unanticipated and harmful to humans.

Download the paper at http://bit.ly/UNIDIR_Autonomy_Risk