Briefing Paper No.88
8 November 2021

Contact: Dr. Ian Davis
Email: idavis@natowatch.org
www.natowatch.org

NATO's new AI strategy: lacking in substance and lacking in leadership

By Peter Burt

The October 2021 meeting of NATO's Defence Ministers in Brussels (see NATO Watch Briefing no.87) saw Ministers agreeing to adopt NATO's new strategy for Artificial Intelligence (AI). The strategy is a first for NATO and is designed to set out how NATO will apply AI in its security role "in a protected and ethical way".

NATO has already started work on the adoption of AI with its Military Uses of Artificial Intelligence, Automation, and Robotics project and establishment of the NATO Data Science Centre, whilst various member states are rapidly taking forward their own programmes to develop military AI systems. Earlier this year NATO Defence Ministers endorsed a strategy on emerging and disruptive technologies – including AI – which will guide NATO's development of such technologies. The clearly stated purpose of developing these technologies is to maintain military superiority and a technological advantage.

The centrepiece of NATO's AI strategy is a set of principles intended ensure that AI is used responsibly by NATO and its allies, and in accordance with international law and the alliance's values. The strategy also discusses threats posed by the use of AI by NATO's adversaries, and how to improve co-operation with the tech sector on AI development.

What the strategy says

NATO's new AI strategy sets out four aims:

- encouraging the development and use of AI in a responsible manner for defence and security purposes;
- accelerating and mainstreaming the adoption of AI in NATO and its forces;
- protecting NATO's AI technologies and ability to innovate; and
- safeguarding against the malicious use of AI.

To meet these aims in an ethical manner, the strategy sets out six principles of responsible use for AI in defence – in shorthand, lawfulness, responsibility, explainability and traceability, reliability, governability, and bias mitigation (see box). These principles are to be applied across the life-cycle of AI applications, and the strategy commits NATO to conducting "appropriate" risk and/or impact assessments prior to deploying AI capabilities.

Unsurprisingly, the strategy acknowledges that NATO's adversaries are likely to exploit defects or limitations within AI technologies and that countermeasures will be necessary to prevent interference, manipulation, and sabotage. It also accepts that AI can be used to impact upon critical infrastructure and create disinformation opportunities. Whilst stating that NATO allies "will seek to prevent and counter any such efforts within the context of the Principles of Responsible Use", the strategy stops short of giving a commitment that NATO will never use AI for these purposes, and likewise does not prohibit NATO and its allies from developing weaponised AI systems that could delegate life and death decisions to artificial intelligence systems.

NATO's principles of responsible use for AI in defence

Lawfulness: AI applications will be developed and used in accordance with national and international law, including international humanitarian law and human rights law, as applicable.

Responsibility and Accountability: AI applications will be developed and used with appropriate levels of judgment and care; clear human responsibility shall apply in order to ensure accountability.

Explainability and Traceability: AI applications will be appropriately understandable and transparent, including through the use of review methodologies, sources, and procedures. This includes verification, assessment and validation mechanisms at either a NATO and/or national level.

Reliability: AI applications will have explicit, well-defined use cases. The safety, security, and robustness of such capabilities will be subject to testing and assurance within those use cases across their entire life cycle, including through established NATO and/or national certification procedures.

Governability: AI applications will be developed and used according to their intended functions and will allow for: appropriate human-machine interaction; the ability to detect and avoid unintended consequences; and the ability to take steps, such as disengagement or deactivation of systems, when such systems demonstrate unintended behaviour.

Bias Mitigation: Proactive steps will be taken to minimise any unintended bias in the development and use of AI applications and in data sets.

Running parallel to the AI strategy is NATO's new Data Exploitation Framework Policy, approved by Defence Ministers at the same meeting (but not yet published). The data exploitation framework sets out measures to treat data as a strategic asset which will be managed, analysed, and stored to provide the 'fuel' needed for AI to operate.

Positives

The new strategy sets out in general terms NATO's future approach to the use of AI, and the ethical principles it highlights are intended to encourage the responsible use of the technology in line with NATO's liberal values. This will be seen by many as a good starting point.

Leaving aside the issue of whether the reality matches the ideal when it comes to NATO's values, a values-based set of principles sets out NATO's position and acts as a clear statement of intent for others to follow, both among NATO member states and NATO's potential rivals. By publicly declaring these principles NATO has enabled others to hold the alliance to account against them in the future. Credit is due to NATO for taking this position. Encouragingly, the AI strategy also recognises that ethical principles need to be applied throughout the life-cycle of an AI product. This is important, because the later ethical factors are considered in the product life-cycle, the harder it may be to ensure they are upheld effectively.

NATO also recognises that, as a technology, AI is not without risks. The strategy gives a welcome pledge to conduct appropriate risk and/or impact assessments before deploying AI capabilities. These will need to be robust and wide-ranging, rather than the box-ticking exercise that such assessments can sometimes become. The specific risk of bias is given special attention: bias mitigation efforts will be adopted with the aim of minimising discrimination against traits such as gender, ethnicity or personal attributes. However, the strategy does not say how bias will be tackled - which requires structural changes which go well beyond the use of appropriate training data—and NATO may find that this is easier said than done.

The strategy also recognises that in due course AI technologies are likely to become widely available, and may be put to malicious uses by both state and non-state actors. NATO's strategy states that the alliance will aim to identify and safeguard against the threats from malicious use of AI, although again no detail is given on how this will be done. Past experience demonstrates that once new weapons are available, they will eventually find their way to non-state and criminal actors through irresponsible state exports, illegal transfers and diversion.

Questions

Whilst publication of the AI strategy is a step in the right direction for NATO, although perhaps not a particularly ambitious one, a number of questions about NATO's position on the use of AI remain unanswered. So far NATO has only published a 'summary' of its AI strategy, raising an obvious question: what does the rest of the document say? Does it expand upon resolving ethical and human rights concerns in more detail and include clear monitoring processes to do this, or does it set out a manifesto for tackling the realities of how NATO expects to take forward the use of AI for military purposes?

Running through the strategy is the mantra of interoperability – the desire for different systems to be able to work with each other across NATO's different forces and nations without any restrictions. An article in the NATO Review journal by two of the authors of the strategy, Zoe Stanley-Lockman and Edward Hunter Christie, makes it clear that "the aim of NATO's AI Strategy is to accelerate AI adoption", and NATO evidently intends AI to be adopted at all levels within the organisation and across all its roles. Despite the good intentions behind the ethical principles, a key thrust of the new strategy is to enable the uptake of AI for military purposes.

In the same article Stanley-Lockman and Christie also talk of AI as an enabler to "out-adapt competitors and adversaries", and the strategy itself states that collaboration and cooperation on AI among NATO allies is necessary "in order to maintain NATO's technological edge". NATO clearly sees the adoption of AI in terms of a zero-sum arms race against rivals such as China and Russia, which are also investing heavily in AI, and presumably believes it can win the race. This is problematic, as arms races have the potential to escalate, as has been the case with other weapons. Competition has no absolute end goal—merely the relative goal of staying ahead of the other competitors. Should one player cross an ethical line, such as developing and deploying autonomous weapon systems, others may be expected to follow suit with destabilising consequences.

NATO's ethical principles for the responsible use of AI, though welcome, raise several issues. Such statements of principle are now commonplace in the corporate sector and are increasingly being adopted by governments on both a unilateral and multilateral basis. NATO's principles are similar to principles adopted by the US Department of Defense for the ethical use of AI: indeed, in some places the wording is the same as that in the Department of Defense principles.

As with many such statements of principle, the NATO principles have no coherent means of implementation or enforcement. Their successful adoption will in many ways depend upon leadership and political and military culture, which is different in each of NATO's 30 member states. Will Turkey, for example, which has been a keen proponent of automated warfare and by some accounts has already developed AI-based loitering munitions with an autonomous capability to identify targets, be willing to follow the same rules as the USA? And would the USA under a second Trump administration follow the same rules as a Biden administration? In the absence of any binding enforcement mechanism NATO's principles may provide useful for public relations purposes but are likely to be less useful in preventing harm to humans, particularly those in the Global South who are already in situations of conflict, and historically marginalised groups.

Despite the plethora of corporate statements on ethical principles, those working in the tech sector are sceptical about the prospect that ethical AI design will be adopted as a norm over the next decade. A non-random poll conducted by the Pew Research Centre in early 2021 found that 68% of experts in the field thought that ethical principles focused primarily on the public good will not be employed in most AI systems by 2030. Their concerns recognised that the main developers and deployers of AI are focused on profit-seeking and social control, and that global competition will matter more to the development of AI than any ethical issues. This latter factor will very much influence NATO's future adoption of AI.

Although NATO's ethical AI principles are stated to have been developed on the basis of "Allied approaches and relevant work in applicable international fora", it is not clear whether they draw on views from the wide range of professional disciplines necessary to develop a representative and rounded view of the ethical pitfalls and risks associated for AI, or from a diversity of perspectives. There has certainly been no open consultation on their formulation,

and it is possible that the principles may only represent the perceptions of a relatively narrow range of predominantly white male technocrats and military planners drawn from within NATO, member governments (notably the USA), and the arms industry. Ordinary people, particularly marginalised groups and those in the Global South, will ultimately face the consequences and impacts of NATO's decisions on AI systems, yet the public have certainly not been involved in making decisions on this strategy, which will set the framework for NATO's future AI choices.

NATO's AI strategy does not discuss the development of AI-driven autonomous weapon systems – a significant omission, given the ethical issues that this application of AI would raise and the challenges to human rights that such weapons would pose. As a bare minimum, the strategy could—and should—have endorsed the 'Guiding Principles' on emerging technologies in the area of lethal autonomous weapons which have recently been adopted by consensus by High Contracting Parties to the Convention on Certain Conventional Weapons (CCW).

Despite increasing pressure for a ban on the development and use of lethal autonomous weapons, key NATO states including the USA and UK have been lukewarm about the need for a ban, arguing that current laws of war are adequate to regulate any such weapons. This seems optimistic, given the rapid development of AI technology and the push by some states over recent years to redefine the rules that govern use of armed force to suit their own purposes. To date no NATO member states have supported the call for a ban on lethal autonomous weapons. If NATO members were serious about ensuring that the military use of AI adheres to international humanitarian law and human rights law, they would call for and engage in negotiations on a legally binding instrument on autonomous weapons systems at the CCW; which is expected to take a historic decision on this issue at its Review Conference in December 2021.

Investment and innovation in artificial intelligence is being led by the private sector, and not by the world's militaries. Recognising this, NATO's AI strategy places a premium on engaging with "start-ups, innovative small and medium enterprises, and academic researchers that either have not considered working on defence and security solutions, or simply find the adoption pathways too slow or restrictive for their business models". NATO wants to "make defence and security a more attractive sector for civilian innovators to partner with". Aiming to capture the civilian tech sector in this way risks increasing the influence of the military-industrial-political complex to an even greater extent than is already the case. Will this really help advance NATO's liberal and democratic values?

Conclusion

Although NATO's AI strategy contains a few worthwhile nuggets, on balance it is tame and unambitious. It is depressingly clear that NATO sees AI as basically another way of using technology to wage war more effectively, and is not willing to show any real leadership to mitigate the risks that AI poses to human rights and dignity. War is, after all, the highest area of risk when it comes to the potential for human rights abuses, yet NATO's strategy says nothing about measures to effectively govern military AI and autonomous weapon systems.

AI has the potential to help humanity tackle intractable 'wicked problems' such as climate change and unsustainable development, and tackle the course of dangerous behaviour which is threatening the survival of our species. But this will only happen if it is employed under wise and decisive human leadership, otherwise it may dramatically compound the problems we face, threatening international security and human rights even more. NATO, unfortunately, seems unwilling to provide that leadership.

Acknowledgement: The author would like to thank Wanda Muñoz for advice and comments on an earlier draft of this paper.

Author: Peter Burt is a researcher at Drone Wars UK working on issues relating to artificial intelligence and autonomy and their role in the future development of drones. Before moving to Drone Wars UK Peter was Director for the Nuclear Information Service for six years, and he is also a Trustee the Nuclear Education Trust.

Disclosure: Drone Wars UK is a member of the Campaign to Stop Killer Robots.