Centre for
Humanitarian
Dialogue

Mediation for peace

Code of Conduct on Artificial Intelligence in Military Systems

Background on the Code of Conduct

This draft Code of Conduct for AI-enabled military systems is the product of a two-year consultation process among Chinese, American, and international experts convened in person and online by the Centre for Humanitarian Dialogue (HD). The goal of the consultation process was to determine whether certain principles and limitations might be agreed regarding weapons and related military systems with significant AI components, especially among those international actors whose technology and deployment in this area is most advanced. Participants in the dialogue included current academics and former officials with military, diplomatic, intelligence, weapons design and legal backgrounds from the United States, China and an international delegation from Europe and Latin America. While some experts participated in previous UN conferences related to limitations on advanced weaponry – such as the UN CCW's Lethal Autonomous Weapons Systems (LAWS) Group of Governmental Experts – the purpose of the consultation was to break away from public positions and to see if a discreet process could find common ground on limitations before AI-enabled military systems become so commonly used as to make future limitation impractical.

Rather than define LAWS or limit consideration to weapon systems, this Code considers the impact of AI on weapons and on related intelligence and targeting systems as they would be used in real-world conflict. In the consultation process, experts participated in scenarios and simulations that highlighted:

- the challenges of proving, when an accident occurs, the AI-enabled system operated in accidental or unexpected ways;
- the risk of misinterpreting actions of AI-enabled military systems and the differences in signaling between manned and unmanned assets;
- entrenched bias against providing to an adversary any transparency related to the testing and evaluation process of advanced weapons systems, even where that transparency would facilitate mutually agreed goals such as safety, security or accountability; and
- the need to maintain human control over weapons to prevent mistakes or accidents in targeting or launch, particularly in sensitive conflict settings and in the nuclear domain.

In sharing this Code with officials, HD Centre hopes that the concepts might motivate thinking among state actors about what agreements are possible related to the future design, deployment and assessment of military systems with significant AI capabilities. Common interests were identified in safety, security, non-proliferation and other areas. We hope to facilitate official consideration of possible limitations on AI-related systems before they become so widespread as to make limitation impractical. Similarly, we hope that official consideration of limitations could be made easier by the fact that the elements of this Code have already been considered and agreed by defense and security-conscious experts from key countries, mindful of the competitive international environment and national security challenges present in today's world.

Preamble

1. The purpose of this draft Code of Conduct is to articulate a set of principles which key states could consider adopting for the design, deployment, use, and assessment of military systems which contain one or more significant components produced through artificial intelligence (AI), especially machine learning (ML). It is desirable that other states in the international community and, to the extent relevant, non-state actors could also agree to these principles.

2. This is not intended to be a comprehensive or legally binding Code of Conduct but could be used as a starting point for developing such a Code of Conduct in the future.

3. This Code of Conduct looks at AI-enabled military systems and components from a risk-based perspective, considering in particular the phases of design/testing, deployment/use, and post-use assessment. In identifying elements for inclusion, this Code of Conduct focuses on topics of significant interest or value to policymakers on which leading States involved in the design and deployment of AI-enabled weapons systems may be able to reach consensus.

4. This Code of Conduct is drafted mindful of the goals of:

    a. Adherence to international law, particularly international humanitarian law (IHL);

    b. Respect for the principles of the United Nations Charter;

    c. Preventing loss of life and avoiding unintentional conflict or escalation of conflict;

    d. Protection of national security, sovereignty, and defense;

    e. Encouraging the continued development of basic science and productive civilian use of artificial intelligence and machine learning applications for human development; and

    f. Ensuring human responsibility for AI-enabled military systems and the consequences of their actions.

5. The focus of this Code of Conduct is on AI-enabled military systems and components as they affect security relationships. Types of military AI systems and components addressed by and relevant to this Code include, but are not limited to, weapons systems; command and control systems; early warning systems and other systems to support human decision-making which may lead to the use of military force; intelligence, surveillance, and reconnaissance (ISR) systems; and other AI-enabled systems that may affect international peace and security. This Code is also relevant to systems designed originally for commercial purposes but subsequently adopted for military use.

General Principles

6. **Legitimate Uses of AI**. States have legitimate rights to pursue technologies, including AI, for civilian use, economic development and national security. States should develop and use AI-enabled military systems in a manner that is consistent with international law and does not undermine international security and stability. States should consider developing and using AI-enabled systems in a manner aimed at reducing civilian casualties and human suffering and avoiding armed conflicts.
7. **Lawful, Ethical, and Justifiable**. States should subject the development and use of AI-enabled military systems to the highest legal and ethical standards. There should be justifiable reasons behind the design, deployment, and launch of AI-enabled military systems and development of related AI equipment and technology. Military AI research and development should support international peace and security and accord with international law, including legally binding UN instruments, and applicable arms control and disarmament instruments.
8. **Robustness and Reliability**. States should ensure that AI-enabled military systems should be reliable and robust and perform in a predictable fashion in accordance with their design specifications. Before deploying AI-enabled military systems, States should ensure they have a high degree of confidence in the system's reliability and in the state's ability to conduct the operation in accordance with international law and the commander's intent.
9. **Human Control and Responsibility**. Any use of AI-enabled military systems should be subject to a clear line of military command and control and human responsibility. Because of the life and death consequences that could result, States should design AI-enabled military systems to ensure human responsibility for the development, deployment, and use of AI systems and to avoid automation bias or ceding all human judgment to AI systems. Human judgment is also necessary for accountability, as humans, not computer systems, are subject to rights and responsibilities under law.
10. **Technically Constrainable**. AI-enabled military systems should have certain reasonable constraints in time and space domains, in order to avoid misuse and unexpected civilian casualties and to allow for termination of an AI-enabled weapon that has passed its intended target.
11. **AI Weapons-Free Zones**. States may wish to consider limiting the deployment of some forms of AI-enabled or autonomous systems to some geographic regions (e.g., "AI weapons-free zones"). Examples could include an agreement not to deploy AI-enabled military systems near, or so as to cause damage to, key civilian infrastructure, including dams, airports or high-use commercial airways and waterways, and not to deploy lethal or untested AI-enabled military systems in areas of enhanced political and military tension.
12. **Prohibiting Certain Types of AI-Enabled Weapons Systems**. As AI is a new technology, States should consider making certain types or fields of AI-enabled military applications off limits. Examples of types of systems that might be declared off-limits for AI include, but are not limited to:
    - small arms and light weapons, due to their significant proliferation risks and less prominent role in state-on-state conflict;
    - nuclear command and control, due to their destructive force and the requirement of human control over nuclear launch decisions; and
    - lethal weapons systems that (a) defy human control, or (b) are autonomous and capable of evolution in unforeseen or dangerous ways after deployment.
13. **Competence**. Policymakers and military leadership must have a grounded understanding of AI systems, including their capabilities, limitations and risks. States should commit to ensuring their relevant political and military leadership comprehend the underlying technologies an AI system uses and, in particular, their limitations. Military, technical, and legal competence are all relevant to those commanding AI-enabled weapons systems.

Design and Development

Design Principles

14. States should design AI-enabled military systems to be reliable, robust, traceable, governable, and free from unintended bias, demonstrably fit for purpose and in a way that increases the transparency and explainability of system functions. To the extent possible consistent with their national security, States should disclose the assumptions, limitations, and a comprehensive evaluation of performance of AI-enabled military systems.

15. States should comprehensively document, for their own internal purposes, all aspects of the design, development, test, and evaluation associated with the development of AI systems. This should include correctly timestamped and version controlled copies of all code and data as well as access to the experimental platform/architecture used during development. States should consider sharing the documentation of the robustness of their systems to the extent consistent with their national security.

16. States should acknowledge the tension between speed and quality of decision-making as an element of the design of AI-enabled military systems. States should develop systems to enhance decision-making efficiency without sacrificing the quality of decisions or the role of human judgment. Design parameters aimed at improving decision-making efficiency of AI-enabled military systems must not obstruct human control or other legal or technical requirements.

17. States should establish and adhere to performance indicators for AI-enabled weapons systems, including accuracy, reliability, robustness, security and other performance indicators. Under normal circumstances, the security and safety requirements for offensive AI-enabled weapon systems should be greater than those for defensive AI-enabled weapon systems.

Failsafe and Signaling

18. States should design their AI-enabled weapon systems to include embedded failsafe functions so that if failures occur, they minimize the probability and consequences of accidents that could lead to civilian harm or unintended military harm and potential unintended escalation. These functions should transfer control of AI-enabled weapons systems to a human, where possible, in the event of a failure. This is especially important for lethal AI-enabled weapon systems, which States should consider designing with embedded self-destruct or other deactivation mechanisms.

19. In designing AI-enabled military systems, States should consider unique issues related to how autonomous and AI-enabled systems may inadvertently signal or indicate hostile intent. States should consider design mechanisms that allow systems to signal in clear, human-identifiable ways non-hostile intent. Systems should be rigorously tested to ensure that actions that would be understood as hostile intent are only displayed if intended. When such systems are fielded, States should consider when additional signaling and communications mechanisms specific to AI systems will be mutually beneficial to reduce inadvertent escalation or risk of accidents.

Test and Evaluation

20. States should ensure comprehensive test, evaluation, validation and verification of AI-enabled military systems in realistic operational conditions prior to deployment.

21. States should establish a suitable institutional framework for test and evaluation of AI-enabled systems, which includes the early consideration of legal and technical issues through consultation with relevant experts.

22. In order to comply with international treaty and customary law, States should subject new weapons to a legal review in order to establish whether they would in some or all circumstances be contrary to their obligations under international law. This obligation applies equally to AI-enabled weapons systems. Consequently, States should ensure that legal reviews become an integral part of test and evaluation of AI-enabled systems from the earliest possible time.

23. States should design test environments and test parameters in good faith and in a manner that provides a realistic impression of the system's performance and its ability to be used in compliance with applicable law and ethical standards. In the absence of international rules or standards on reliability and predictability of AI-enabled weapons systems, States should consider establishing their own minimum standards against which the performance of systems under testing and review can be measured.

24. States should ensure that test and evaluation, including legal reviews, is an ongoing and iterative process which should be initiated at the earliest possible time in the development and/or acquisition and adoption phase and repeated throughout the life cycle of the system, to include accounting for continuous deployment, such as when fielded systems receive regular AI algorithm or model updates.

25. States should establish procedures to monitor changes in system functions in order to be able to decide whether and when additional or supplementary reviews are required to ensure compliance with IHL and other relevant rules of international law.

Deployment and Use

Oversight

26. States should ensure that those responsible for the use of AI-enabled weapons systems should take all necessary precautions to limit attacks to military objectives and combatants and avoid or minimize incidental loss of civilian life and damage to civilian property. Depending on the circumstances of the individual attack, necessary precautions may include but are not limited to:

    a. Restrictions on the geographical scope of the operation;

    b. Restrictions on the temporal scope/duration of the operation;

    c. Restrictions on categories of targets;

    d. Restrictions on target recognition criteria; and

    e. Requirement of target approval or mission-supervision by a human operator.

27. States should ensure that humans are responsible for exercising judgment over the use of force, especially decisions over the use of force that may result in the loss of human life. States should ensure that authority for life and death decisions is never delegated to machines.

28. Deployment of AI-enabled military systems in cyberspace may present unique challenges, especially related to attribution, accountability, and human control. Nonetheless, deployment of AI-enabled military systems in cyber space should be subject to the same principles and limitations as in other domains.

29. Because of the destructive potential of nuclear weapons, States must ensure that humans remain in control of nuclear launch decisions. States should ensure that nuclear command-and-control systems are designed such that affirmative human action is needed to initiate a nuclear launch and that a technical accident cannot result in an inadvertent launch. For example, early warning systems should be physically separate from nuclear launch systems, with humans providing a critical oversight role at every step of a potential nuclear weapon launch sequence. Similarly, uninhabited vehicles (e.g. drones) should never be nuclear-armed or used as nuclear delivery platforms because of the risk of accidents and loss of control over nuclear weapons.

Competence and Responsibility

30. States should ensure that design requirements are understood and approved by those with military expertise.

31. States should ensure that decisions about use of AI-enabled military systems remain the responsibility of military commanders and are taken solely by those with the appropriate authority. States should ensure that commanders have the technical competence to interpret, explain, and use AI-enabled military systems under their command, including an appreciation of the impact of uncertainty on an AI-enabled military system's performance, the likely sources and scale of errors in deployment, and the costs of such errors. States should make human experts available who can support commander-led decision-making.

Post-Use Assessment and Accountability

32. States should ensure that an investigation is conducted in the event of any accidental, unlawful, unethical, or unexpected activity resulting from deployment of an AI-enabled military system. States may wish to consider sharing the facts and outcome of all AI-related incidents to the extent consistent with their national security.

33. States responsible for the deployment of an AI-enabled military system should ensure that any necessary remedial, corrective, and retraining actions after an accidental, unlawful, unethical, or unexpected use are taken as soon as possible and in all instances before the system in question is placed back in operation.

34. States should share information about corrective and retraining steps with other interested parties, including those involved in incidents of accidental, unexpected, unethical, or unlawful use, to the maximum extent possible consistent with their national security.

Information Sharing, Trust Building and International Cooperation

Transparency and Identification

35. States should consider transparency measures about the functionality of autonomous systems to avoid the risk of miscalculation and accidents and to allow verification of safety measures in the event of a post-incident investigation.

36. To reduce the risk of uncertainty as to origin and command, States should recognize and adhere to existing international legal obligations to affix distinctive national emblems to AI-enabled military systems and their related components. States should consider how to ensure such emblems are recognizable, including to human adversaries and to other AI-enabled systems.

37. States may wish to consider means of identifying and communicating to other parties various characteristics of deployed AI-enabled military systems, including but not limited to: the extent to which a vehicle/vessel/aircraft is crewed ("manned") or uninhabited ("unmanned"); the level and type of autonomous functionality of the system and the extent of human oversight; and whether or not the system is armed/unarmed. This could be accomplished through a variety of identification and marking measures, including but not limited to physical markings such as paint, flags, lights, or other emblems and/or electromagnetic means such as a radiofrequency beacon or message.

38. States may wish to consider development of a universal signal for an AI-enabled military system that has been disarmed or reverted to a failsafe mode.

Information Sharing

39. States should, to the extent consistent with their national security, share procedures for and the outcomes of (a) their legal reviews and (b) their testing and evaluation of AI-enabled military systems.

40. States should consider adopting a common set of performance metrics and evaluation criteria for evaluating the effectiveness, robustness, and safety of AI-enabled systems. States should ensure and certify that their AI-enabled military systems achieve agreed minimum performance.

41. States should consider whether revisions to internal rules of engagement are necessary to accommodate engagement of military AI systems, and may wish to consider sharing aspects of such amendments that might reduce accidents or miscalculation. States may wish to consider whether there is mutual benefit in documenting and sharing command structures as they relate to AI-enabled military systems.

42. States should exchange research on the basic science of machine-learning interpretability and robustness and should, to the extent consistent with their national security, consider exchange of information about related applications in order to, among other things, advance the goal of ensuring human oversight. The longstanding tradition of open publication of academic advances in AI/ML systems can be an important source of transparency, providing a level of detail that can be shared in circumstances where national security prevents sharing of applied system details.

International Cooperation

43. States should establish regular dialogue on military doctrine, capabilities and uses of AI to reduce misunderstanding and address questions about specific AI systems and how their attributes may undermine the safety and security of all nations. In the case of States operating in close proximity deploying AI-enabled military systems, States should, consistent with their national security, develop no-fail methods for crisis communication.

44. States should cooperate on counter-proliferation efforts to increase transparency about the proliferation of potentially dangerous AI systems to state or non-state actors. Specific measures could include marking systems as to national ownership, sharing information about international arms transfers, technical safeguards, and/or agreeing not to proliferate certain AI systems or applications to certain actors.

45. States may wish to consider creating, supporting and using a neutral, technically competent third-party organization that could:

    a. Recommend international standards and best practices for reliability, robustness, assurance, and verification of AI-enabled military systems;

    b. Propose a set of internationally agreed confidence building measures (CBMs) for AI-enabled military systems.

    c. Propose a common set of metrics and standards for test and evaluation for AI-enabled military systems;

    d. Contribute to test, evaluation, verification, and validation in a way that supports the protection of sensitive information and capabilities;

    e. Establish training programs that would equip commanders and their teams with technical and legal knowledge relevant to having responsible command and control of AI-enabled military systems; and

    f. Clarify ill-informed representations of AI capabilities or doctrine (so called “myths”).

Centre for
Humanitarian
Dialogue

Mediation for peace

www.hdcentre.org