THE AI INITIATIVE

THE FUTURE SOCIETY at Harvard Kennedy School

Artificial Intelligence and the Law of Armed Conflict: *Parameters for Discussion*

The Future Society at the Harvard Kennedy School of Government

Although there is near universal agreement on the customary norms governing armed conflict there has been no international discussion on applying these standards to the incorporation of Artificial Intelligence (AI) agents used in support of military operations.

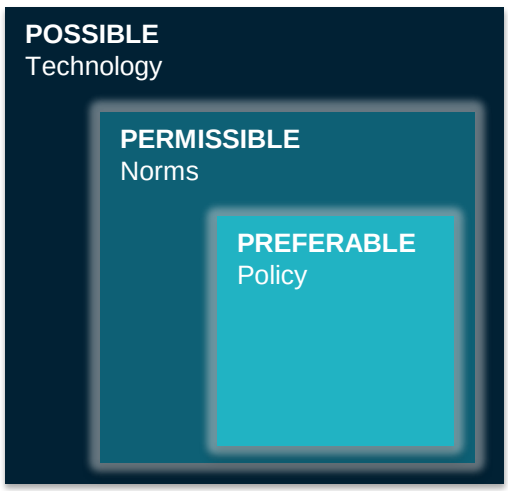
Technology, Law, and Policy: The Three P's

Proposed model to allow for the building of a common theoretical frameworks to help key stakeholders synchronize academic research, operational planning, and high-level policymaking.

Possible (Physical constraints): Consists of everything that technologists have or can deliver without violating the laws of physics.

Permissible (Normative constraints): Nestled within the possible is the permissible which reflects the range of lawful options for policymakers to consider.

Preferable (Policy constraints): Within the permissible is the preferable which encompasses the policy options which are most compatible and aligned with the political landscape, cultural sentiment, and national interests.



The chief benefit of this construct is that it provides a common framework that allows the three indispensable disciplines within cyber to avoid talking past each other as they discuss limitations on policy options, capabilities development, or operational courses of action.

Legal Safe Harbor for Command Responsibility

Have the seven sets of norms which form the basis of the Law of Armed Conflict (LOAC) been satisfied in mission planning and execution? For fully autonomous operations, has the AI been programmed (or taught) how to independently apply the most fundamental of these norms at cyber-speed, without immediate human supervision? For hybrid operations, do humans have the appropriate resources immediately available for real-time decision making in concert with the AI?

Here are the first 35 questions which should be asked by the human-AI team before embarking on a potentially destructive mission in cyberspace:

1. Participation: Does the Law of Armed Conflict (LOAC) apply to this cyber operation? Are there any geographical limitations we are bound to respect in planning the operation? Is the conflict international or non-international? What is the criminal responsibility of leaders who are aware or should have been aware of violations of LOAC? What is the criminal responsibility for operators in carrying out a patently unlawful order?

2. Attacks Generally: Are the attacks being carried out by members of the armed forces? Is the attack part of a cyber *levee en mass*? Do the operators carrying out the cyber operation meet the international definition of "mercenaries'? Are the operators civilians?

3. Attacks against Persons: Are the operators aware of the general prohibition of attacks against civilians? How much doubt is there as to the status of the person being attacked? Is the person a lawful object of attack? Are civilians directly participating in a cyber attack? Can the cyber attack be characterized as a terror attack?

4. Attacks against Objects: Is the operator aware of the general prohibition of attacks against civilian objects? In targeting, has distinction been made between civilian objects and military objectives? Have targeters distinguished between objects used for civilian and military purposes? How much doubt is there as to the status of the object to be attacked?

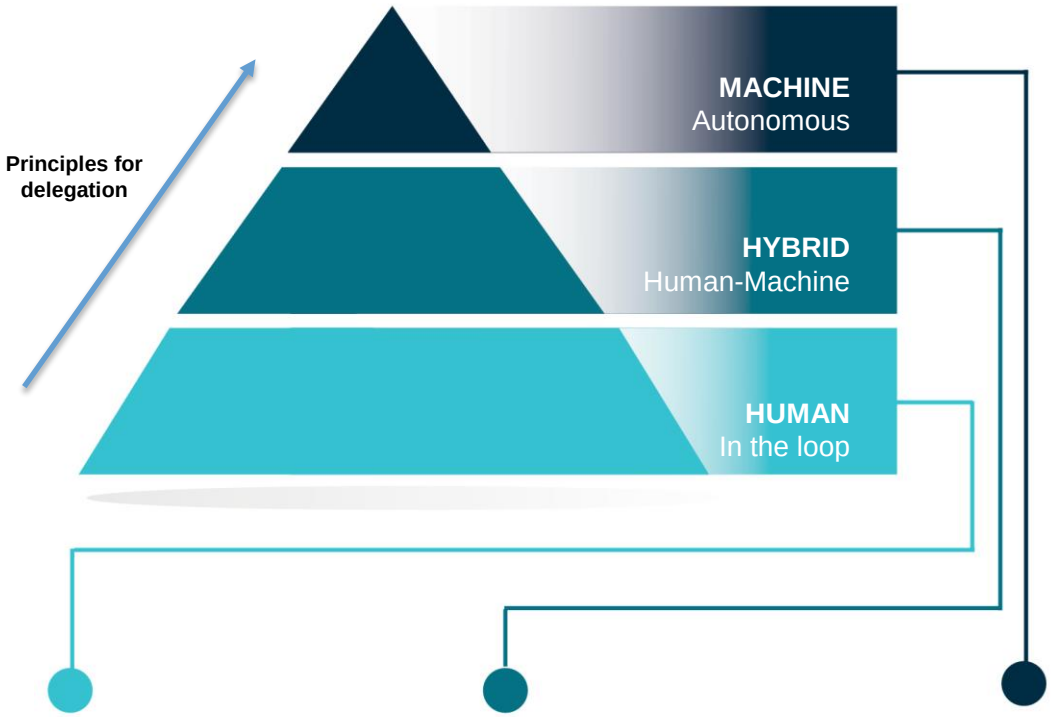
5. Means & Methods of Warfare: Is the operator aware of the definition of 'means and methods of war'? Will the cyber attack cause superfluous injury or unnecessary suffering? Are operators employing an indiscriminate means or method? Doe the operation involve laying a cyber booby trap? Is the cyber operation an indispensable component of a larger operation that will result in the starvation of civilians? Can the operation be characterized as a belligerent reprisal? Does the alternative definition of belligerent reprisal of Additional Protocol I apply to the country conducting the cyber operation? Has the cyber weapon been the subject of a legal review to ensure that it is not per se indiscriminate?

6. Conduct of Attacks: Is this an indiscriminate attack? Is the operation being conducted against clearly separated and distinct military objectives? Has the principal of proportionality (attacks may not cause unreasonable collateral damage to civilians or their property with respect to the concrete and direct military advantage expected to be gained?)

7. Precautions in Attack: Has constant care been exercised in the planning and execution of the cyber attack? Have the targets been verified? Have the appropriate limitations been placed upon the choice of means and methods to attack? Have precautions as to proportionality (the expected collateral harm to civilians and their property weighed against the concrete and direct military advantage to be gained) been taken? Have the proper restrictions on choice of targets been enforced? Do mission rules provide for the suspension of cancellation of attack when certain legal criteria have been met? If required, have appropriate warnings been issued with sufficient lead time?

The AI-Human Pyramid of Interaction

Designed to vertically integrate legal considerations across the range of autonomous operations, the AI-Human Pyramid of Interaction provides operators, mission planners and decision-makers greater operational awareness by stratifying the requirements for autonomous, semi-autonomous and directed operations.



Level where the human operator has the most control and fully runs the machine.

Level where the machine reports back to the human operator for input. Involves comprehensive decision trees based around the law of armed conflict and organizational ethics.

Level where machine operates autonomously and independently makes distinction on what can and can't be destroyed.

We recommend that the best way forward to start tackling this issue is by convening an global group of experts and stakeholders from the public, private and NGO sector to address the following:

- In search of a common ground, how can the global community reach a consensus around a working definition of "AI"?.

- How can the interdependent dynamics between technology, law, policy and society best be framed?

- What are the most effective/appropriate technological, legal and policy safeguards that would need to be implemented to ensure that the military, intelligence, and law enforcement operations are kept within the bounds of accepted international norms? What forum would be most appropriate to develop these?

Authors

Thomas Wingfield, J.D., LL.M / Professor of Cyber Law / National Defense University
Thomas.Wingfield [at] ndu.edu

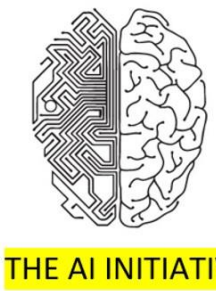
Lydia Kostopoulos, PhD
L [at] Lkcyber.com

Cyrus Hodes, Director, The AI Initiative, The Future Society @ HKS
cyrus [at] ai-initiative.org

Nicolas Miailhe, Co-founder and Director for Projects, The Future Society @ HKS
Nico [at] thefuturesociety.com

The Future Society (TFS) at the Harvard Kennedy School of Government (HKS) works to raise awareness regarding the consequences of the NBIC revolution (Nanotechnology, Biotechnology, Information Technology and Cognitive Sciences) by placing key policy-making questions at the center of its work.

The AI Initiative engages a wide range of stakeholders globally to foster discussion and facilitate policy-making.



Contact

The AI Initiative | The Future Society @ Harvard Kennedy School
www.thefuturesociety.org **|** cyrus.hodes [at] ai-initiative.org.org **|** nico [at] thefuturesociety.org

@hksfuture

www.linkedin.com/groups/The-Future-Society-8431532