Uncertainty and Meaningful Human Control

May 24, 2021

Penn Symposium on Social Implications of Autonomous Weapons Systems

Lisa Miracchi, Ph.D.

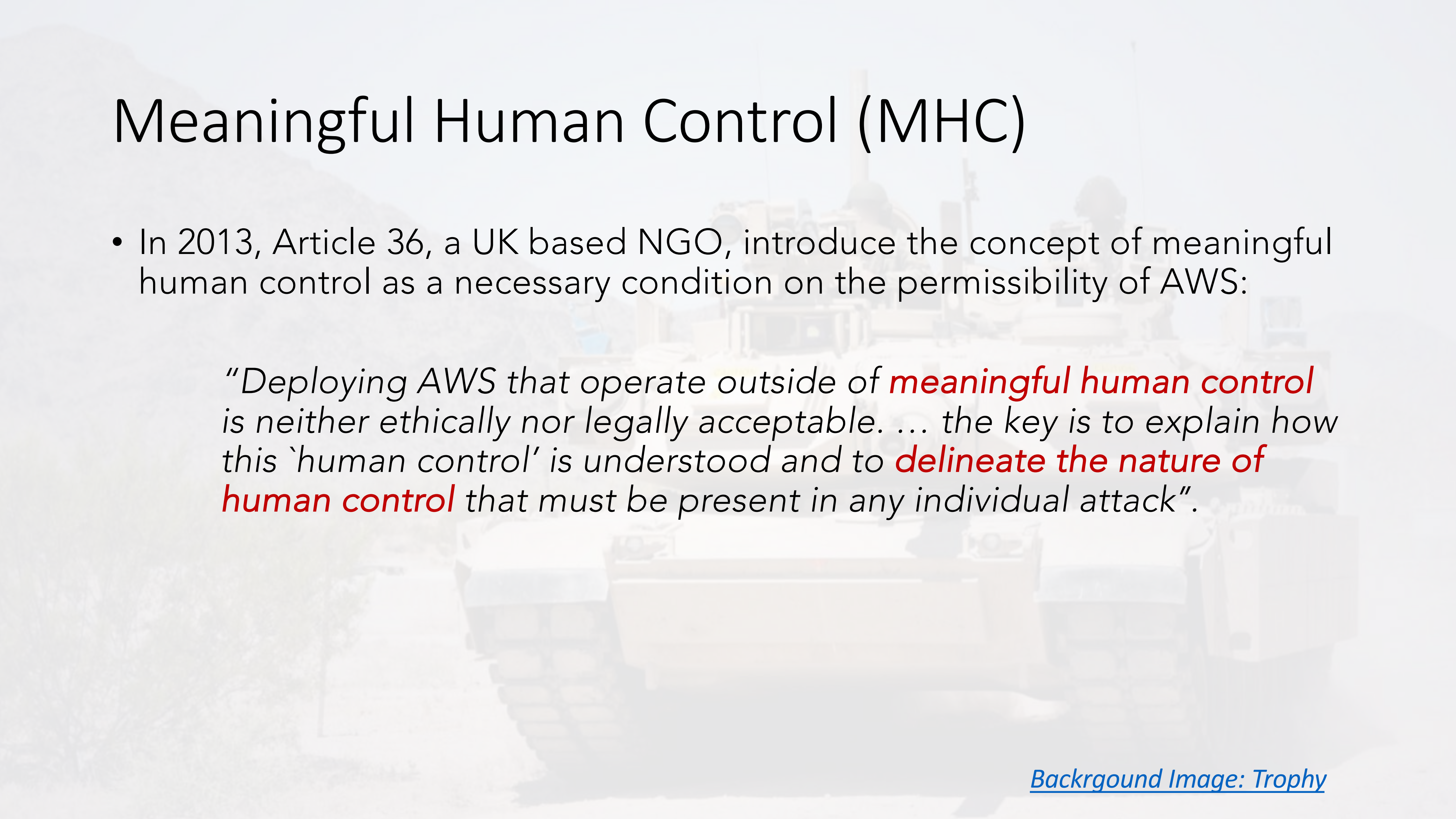
Assistant Professor of Philosophy

Affiliated Faculty, GRASP

Affiliated Faculty, MindCORE

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

mindCORE
Center for Outreach, Research, and Education



Meaningful Human Control (MHC)

- In 2013, Article 36, a UK based NGO, introduce the concept of meaningful human control as a necessary condition on the permissibility of AWS:

> *“Deploying AWS that operate outside of meaningful human control is neither ethically nor legally acceptable. … the key is to explain how this `human control’ is understood and to delineate the nature of human control that must be present in any individual attack”.*

Backrgound Image: Trophy

What’s the challenge for MHC in AWS?

Image

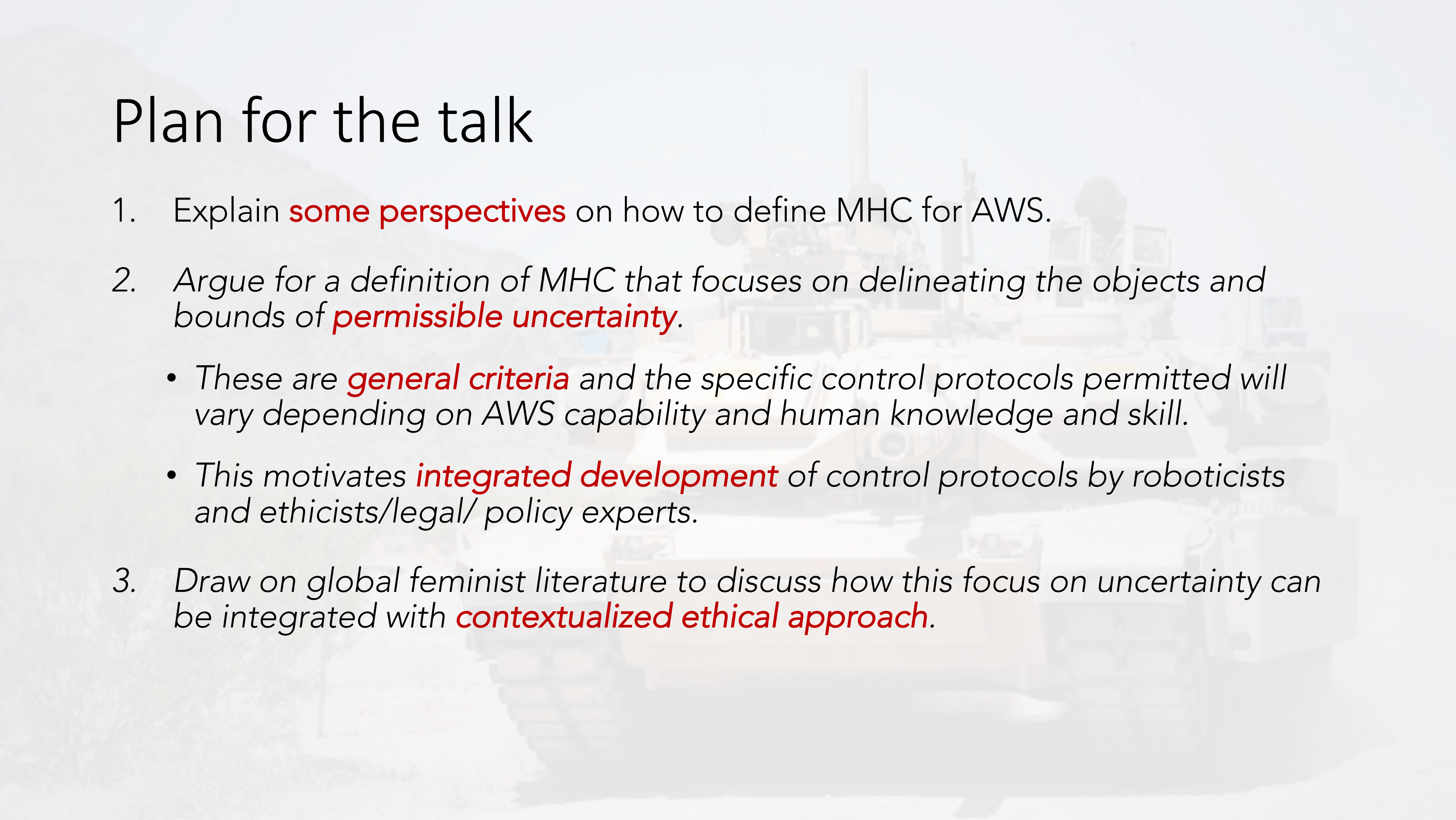
“When that engagement occurs at beyond-visual-range, that pilot has meaningful human control even though the pilot makes the decision to fire entirely based on information received from sensors and computer processors – machines – and computers then guide the missile onto the target.”

- Horowitz and Scharre (2015)

MHC ***cannot*** require:

1. Total information.
2. Infallibility.
3. Ability to abort throughout.

Image



Plan for the talk

1. Explain some perspectives on how to define MHC for AWS.

2. *Argue for a definition of MHC that focuses on delineating the objects and bounds of permissible uncertainty.*
   - *These are general criteria and the specific control protocols permitted will vary depending on AWS capability and human knowledge and skill.*
   - *This motivates integrated development of control protocols by roboticists and ethicists/legal/ policy experts.*

3. *Draw on global feminist literature to discuss how this focus on uncertainty can be integrated with contextualized ethical approach.*

Which human control protocols allow for MHC?

- Strategies for answering this question:

  1. Specify control features of permissible AWS systems or human-AWS systems (from Amoroso & Tamburrini 2020):
     - Uniform human control
       - E.g. boxed autonomy, denied autonomy, supervised autonomy.
     - Differentiated human control
       - E.g. AWS selects possible targets, human chooses.

  2. Specify general criteria a human control protocol would need to meet, independently of particulars of the AWS or human-AWS system.

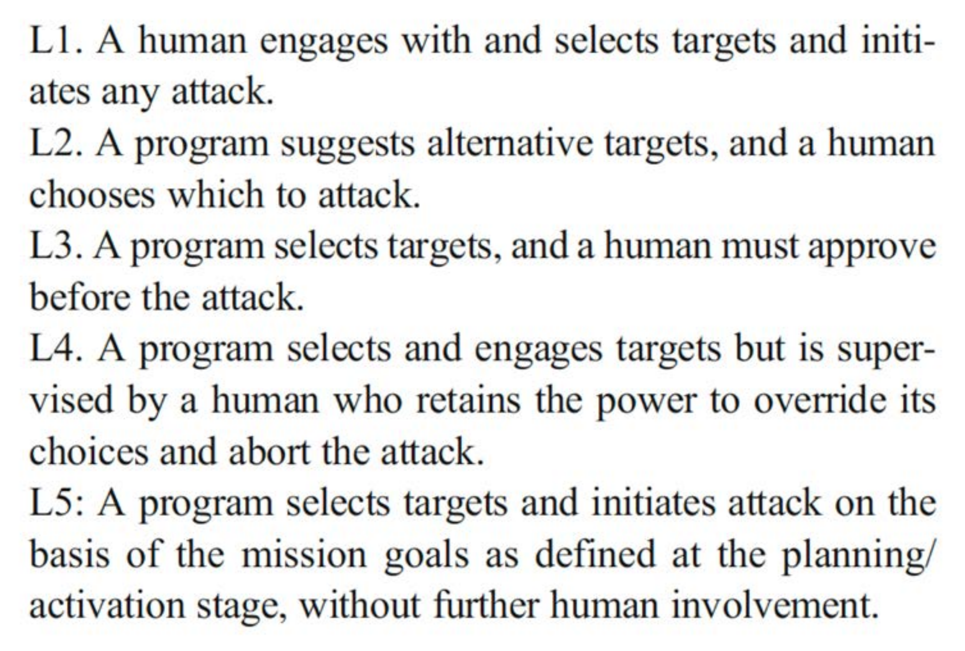
Daniele Amoroso

Guglielmo Tamburrini

Control feature approach

*"The prudential character of this policy is embodied into the following **default rule:** low levels of autonomy L1–L2 should be exerted on all weapons systems and uses thereof, unless the latter are included in a **list of exceptions** agreed on by the international community of States."*

~ Amoroso & Tamburrini 2020.



L1. A human engages with and selects targets and initiates any attack.

L2. A program suggests alternative targets, and a human chooses which to attack.

L3. A program selects targets, and a human must approve before the attack.

L4. A program selects and engages targets but is supervised by a human who retains the power to override its choices and abort the attack.

L5: A program selects targets and initiates attack on the basis of the mission goals as defined at the planning/activation stage, without further human involvement.

For the General Criteria Approach

1. The General Criteria Approach allows us to treat the ethics of AWS as continuous with that of other weapons.

   - The issue is not "letting machines make decisions" but rather making sure that the decisions of humans are appropriately constrained.

   - AWS as tools, not agents. Reframe the debate from "killer robots:

     > *"Treating a human as an object is what happens when LAWS are allowed to kill. The victim, be she combatant or civilian, is reduced to a data point in an automated killing machinery that has no conception of what it means to take a human life." (Rosert and Sauer 2019).*

   - *Aside: Work to change incentive structure to change language.*

2. We want ethical guidelines that continue to apply even as technology changes and develops.

Anti LAWS:

Elvira Rosert

Frank Sauer

H&S’s General Criteria Approach

- Horowitz and Scharre (2015) clarify the concept of MHC as having three essential components:

  1. Human operators are making informed, conscious decisions about the use of weapons.
  2. Human operators have sufficient information to ensure the lawfulness of the action they are taking, given what they know about the target, the weapon, and the context for action.
  3. The weapon is designed and tested, and human operators are properly trained, to ensure effective control over the use of the weapon. (14-15)

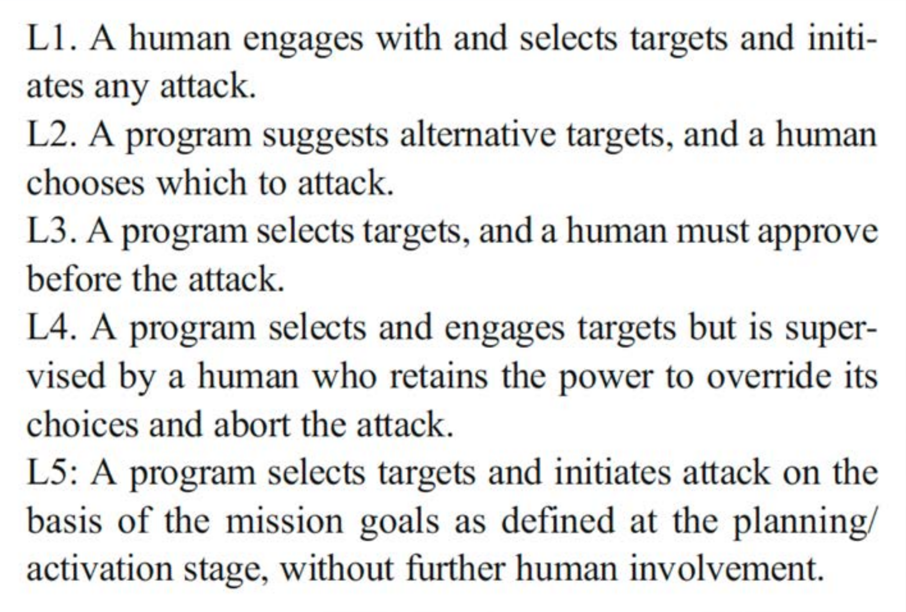
- All of these conditions are crucially epistemic. I suggest that we focus on these epistemic conditions to provide general criteria for permissible uncertainty with contextualized applications.

Michael Horowitz

Paul Scharre

Specifying permissible uncertainty:

- We should be asking epistemic questions at every level of control.
- L1/ L2: How might the AWS system bias for the selection of certain errors or against the selection of legitimate targets?
    - Analogy to racism in face recognition software.

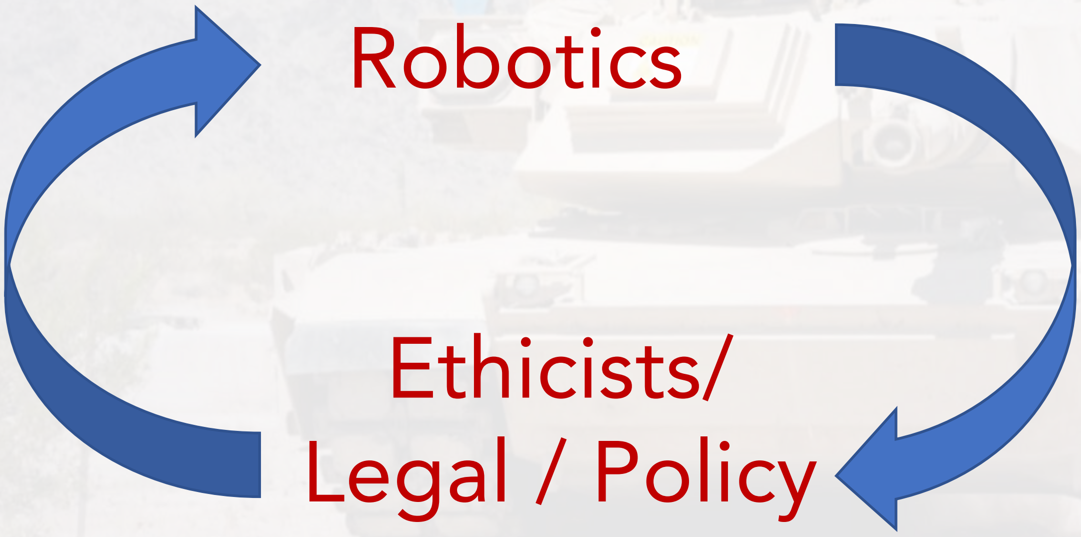


L1. A human engages with and selects targets and initiates any attack.
L2. A program suggests alternative targets, and a human chooses which to attack.
L3. A program selects targets, and a human must approve before the attack.
L4. A program selects and engages targets but is supervised by a human who retains the power to override its choices and abort the attack.
L5: A program selects targets and initiates attack on the basis of the mission goals as defined at the planning/ activation stage, without further human involvement.

- L4: What options for unforeseen events between engagement and completion of attack are likely enough to warrant the need for the ability for human intervention?
    - Compare SARMOs (Mantis, Phalanx, C-Ram) which have low risk of human harm.

Key roles for robotics community in clarifying permissible uncertainty.

- A focus on explainability of AWS in context requires integration of AI and robotics expertise with ethical and legal experts.

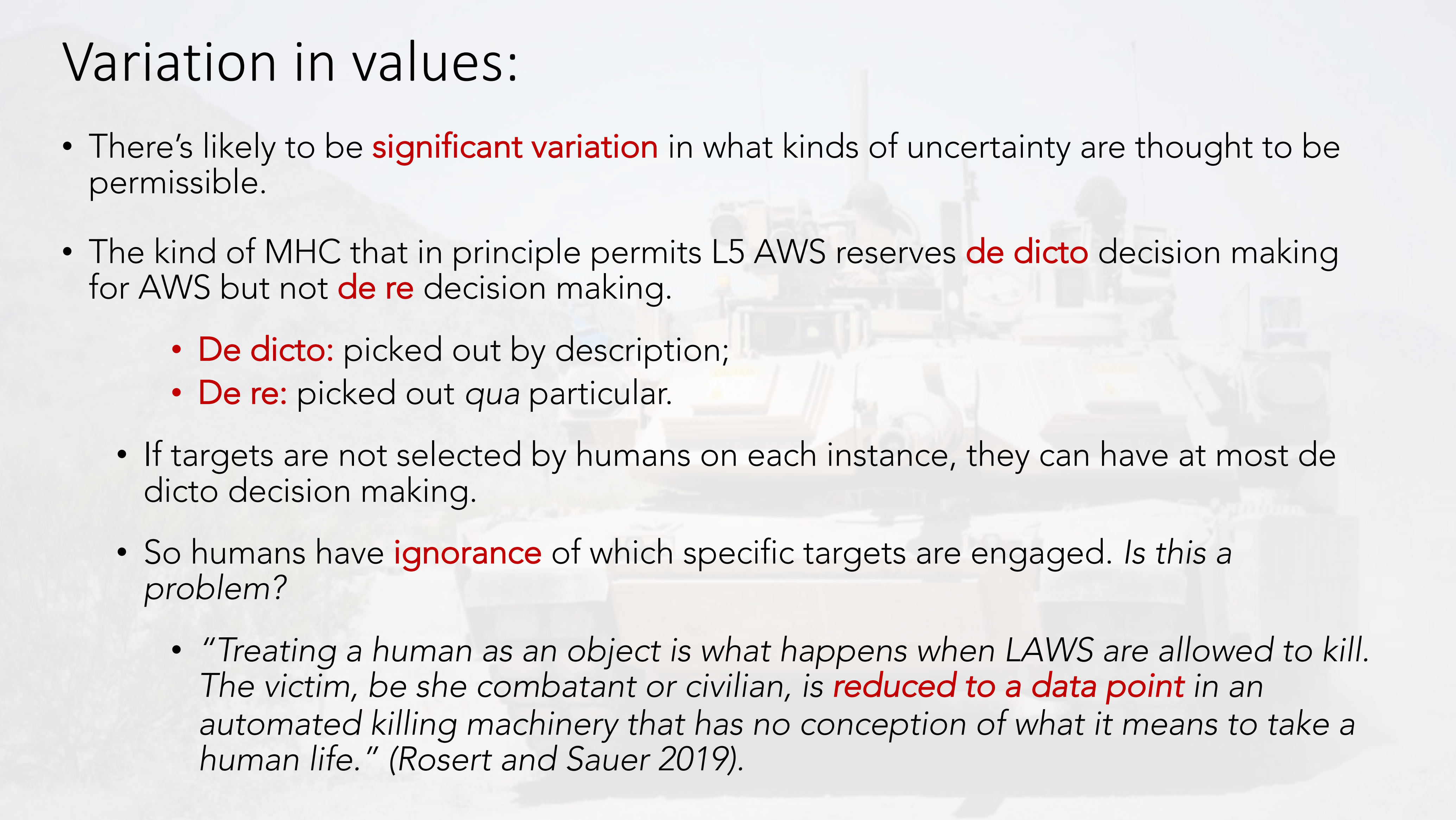


Robotics

Raise and refine uncertainty questions

Provide concrete grounds for justifying/ denying MHC claims, comparing between weapons systems.

Ethicists/ Legal / Policy



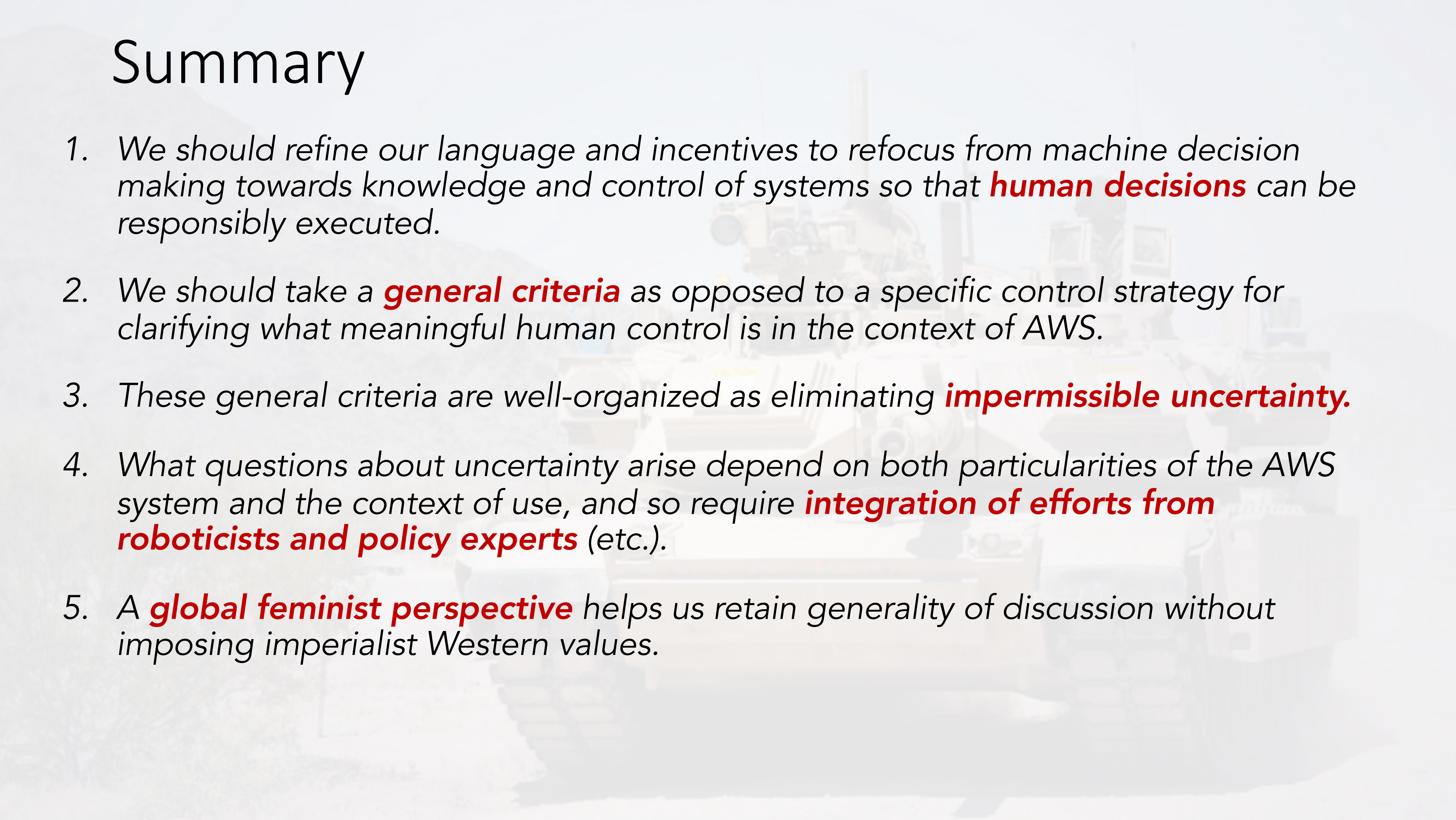
Variation in values:

- There's likely to be significant variation in what kinds of uncertainty are thought to be permissible.
- The kind of MHC that in principle permits L5 AWS reserves de dicto decision making for AWS but not de re decision making.
    - De dicto: picked out by description;
    - De re: picked out *qua* particular.
  - If targets are not selected by humans on each instance, they can have at most de dicto decision making.
  - So humans have ignorance of which specific targets are engaged. *Is this a problem?*
    - *"Treating a human as an object is what happens when LAWS are allowed to kill. The victim, be she combatant or civilian, is reduced to a data point in an automated killing machinery that has no conception of what it means to take a human life." (Rosert and Sauer 2019).*

Variation – a global feminist perspective

- *From Serene Khader,* **Decolonizing Universalism** *(2019):*
  - *"… feminism requires **universalist opposition** to sexist oppression, but feminism does **not** require universal **adoption of Western** … values and strategies" (3).*
    - *E.g. secularism, not wearing head coverings, etc.*
- *Analogously, we might say:*
  - *Ethical use of AWS requires MHC, which **requires the elimination of impermissible uncertainty**. It does **not** require universal adoption of Western values about which kinds of uncertainty are (im)permissible or strategies for eliminating uncertainty.*

*So, **international discussions** about MHC can be reframed in mutually respectful ways centering on the common goal of eliminating impermissible uncertainty and working towards **mutually acceptable standards.***



Summary

1. *We should refine our language and incentives to refocus from machine decision making towards knowledge and control of systems so that* ***human decisions*** *can be responsibly executed.*

2. *We should take a* ***general criteria*** *as opposed to a specific control strategy for clarifying what meaningful human control is in the context of AWS.*

3. *These general criteria are well-organized as eliminating* ***impermissible uncertainty.***

4. *What questions about uncertainty arise depend on both particularities of the AWS system and the context of use, and so require* ***integration of efforts from roboticists and policy experts*** *(etc.).*

5. *A* ***global feminist perspective*** *helps us retain generality of discussion without imposing imperialist Western values.*

THANKS!

Lisa Miracchi, Ph.D.
miracchi@sas.upenn.edu
LisaMiracchi.com

MIRA GROUP

LEGES SINE MORIBUS VANAE

mindCORE
Center for Outreach, Research, and Education