

Autonomous Weapon Systems: Dismantling False Assumptions

Sarah Scheffler and Jacob Ostling

Autonomous Weapon Systems and the Law of Armed Conflict

An **autonomous** weapon system makes targeting decisions and fires without intervention from a human. We respond to two false assumptions in the literature debating banning versus regulating AWS.

Under the Law of Armed Conflict, a weapon must adhere to the following five principles:

Suffering: The weapon must now cause suffering to combatants with no military purpose. Not an Al issue.

Military Necessity: The weapon must provide an advantage for legitimate military objectives. Al provides many.

Distinction: The weapon must always distinguish between civilians and combatants/military objectives.

Proportionality: The weapon cannot cause incidental harm to civilians excessive compared to military advantage.

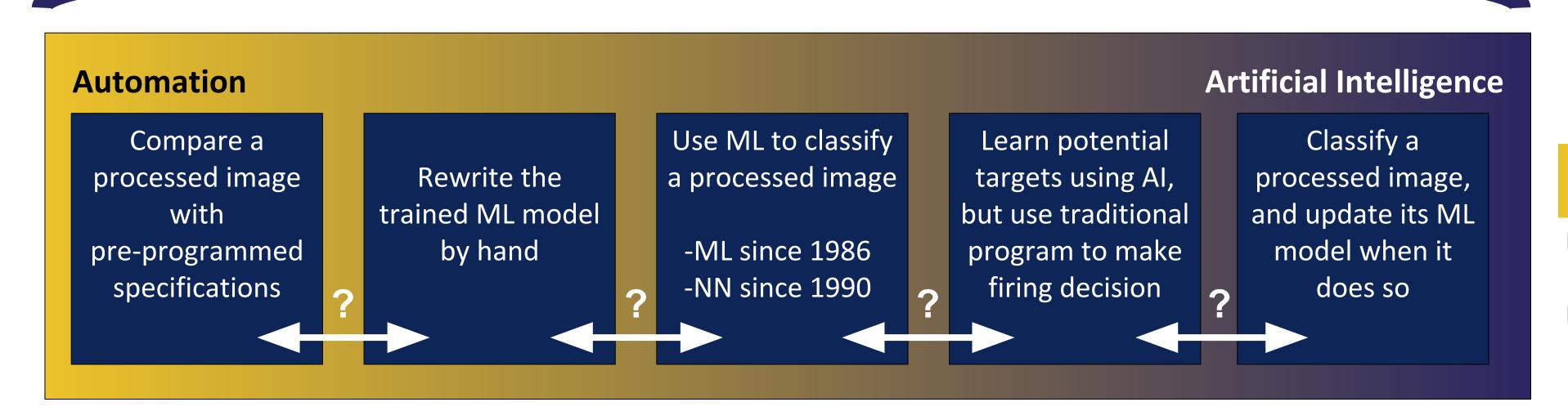
Command accountability: Superiors deploying the weapon must be held liable for war crimes committed with it.

No clean separation between AI and automation: Include both as AWS

False assumption: "Automated" systems and "artificial intelligence" are distinct (thus AI can have a different legal standard)

Our response: There is no clean line between automated systems and AI, so be over-inclusive and hold all systems that choose and fire on targets without human intervention to the same legal standard

Hypothetical Targeting Systems



Tomorrow's AI may fix current shortcomings: Plan accordingly

False assumption: The shortcomings of today's artificial intelligence are inherent, even for future Al

Need for reduced communications

Our response: There is evidence that AI will improve in performance and predictability (especially for distinction and command accountability), and we should plan accordingly

Autonomous vehicle improvements aid AWS

Roadmap for predicted future AWS improvements

Potential technical improvements Potential non-technical changes Better detection of unexpected (low-accuracy) input, "Military advantage" comes to be failure modes are more predictable understood more objectively Predictability improvements lead to better command Hardware forces deactivation after accountability practices short time out of contact Testing methods for AI improve Human supervisors prevent existing failures from worsening Security/anti-tamper methods improve Commanders still held accountable for all AI actions Improvements in image processing for removing Pace of combat increasing clutter, decoys, identifying dynamic targets

LOAC Considerations

Distinction:

- Does it matter if the AI powers the *targeting* decision, versus the *firing* decision?

Proportionality:

- AWS can programatically determine acceptable distributions of collateral damage using existing frameworks, but "military advantage" is a subjective, case-by-case evaluation

Command accountability:

- Commanders must be able to judge what AWS will do to be held accountable for their actions. AI makes this more difficult, but does not require a higher legal standard.
- Al must be predictable *enough* for commanders to be held accountable for AWS actions

All be held to the same (high) standard, despite technical differences