Proposed Principles for the Combat Employment of Weapons Systems with Autonomous Functionalities

Robert O. Work

Ethics for Intelligent Autonomous Systems
Technical Exchange Meeting
3 March 2021

State of play

- International debate over lethal autonomous weapon systems (LAWS) has been underway for nearly a decade
- Striking the match: publication of DODD 3000.09, *Autonomy in Weapon Systems*
- The fire erupts: Campaign to Stop Killer Robots
- Debate has been hampered by the lack of an agreed upon definition for LAWS
- 2019: state parties to the UN CCW agreed “human responsibility” for the decisions over the use of LAWS and the use of force “must be retained”
- Discussions now tend to focus on the type and degree of human involvement in the employment of LAWS to ensure compliance with IHL

Despite the disagreement over terms (or perhaps because of it) DoD needs to take this debate far more seriously, and be far more active in trying to shape it

- The danger can be traced to DODD 3000.09's lack of acknowledgement of the long history of autonomy in weapon systems: as a result, some of the more extreme definitions and arguments against lethal autonomous weapon systems include virtually every guided munition now in service or under development, and would impose exceptionally onerous operational guidelines on their employment:

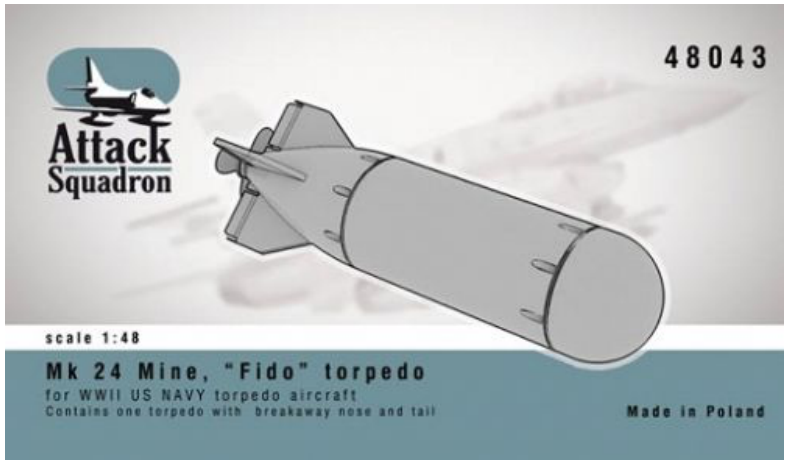
  - E.g., International Committee on Robot Arms Control: " If a 'semi-autonomous weapon system' may have capabilities to autonomously acquire, track, identify, group and prioritize targets, and to control their engagement once a "go" signal is given, conversion to full lethal autonomy could be as simple as throwing a (software) switch…Since verification of the non-existence of an autonomous option in software is virtually impossible, and would be deemed far too intrusive, **a tamper-proof system will be needed that can verify, after the fact, that an attack in question was under direct control of a human being ("in the loop," not "on the loop"). This could be achieved by keeping the records of each engagement and making the records of specific engagements available to a Treaty Implementing Organization, on request, when sufficient evidence exists to support suspicions of illegal autonomous operation.**"

Possible next step

- Several scholars argue the discussions over LAWS should focus on “developing objective, commonly held, and ***functions-based understandings of autonomy in the military context***
- Premise of this presentation is the best way to achieve such an understanding is to develop, debate and agree upon some commonly held principles for the ***employment of weapon systems with autonomous functionalities in armed conflict***
- DOD should lead the way and publish such principles

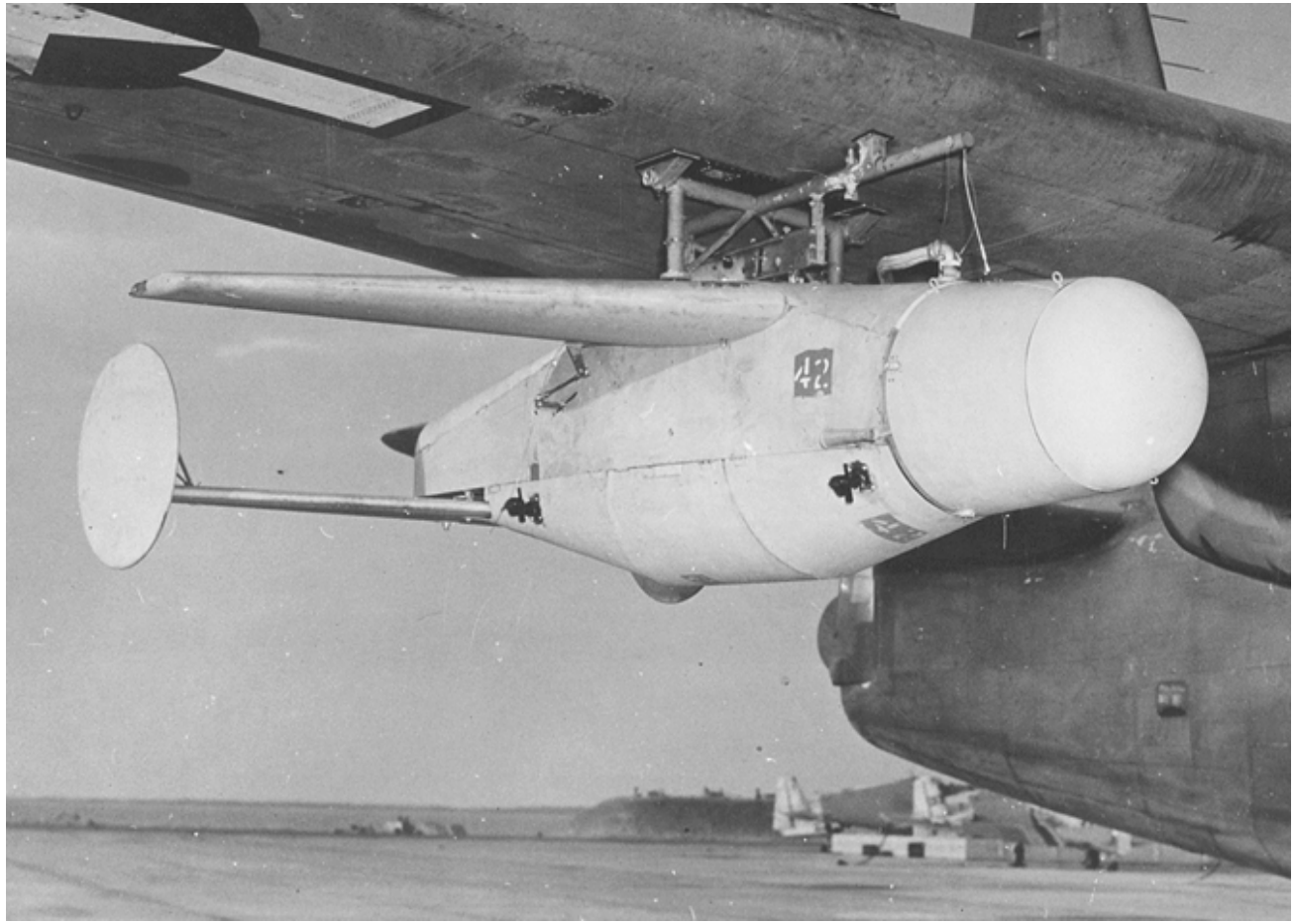
A short history of weapon systems with autonomous functionalities

- Combat debut: 1943
  - Mk-24 FIDO acoustic homing torpedo



Led to numerous “fire and forget” weapons

- 1945
  - SWOD-9 BAT



Two-stage guided munitions designed to dispense guided submunitions

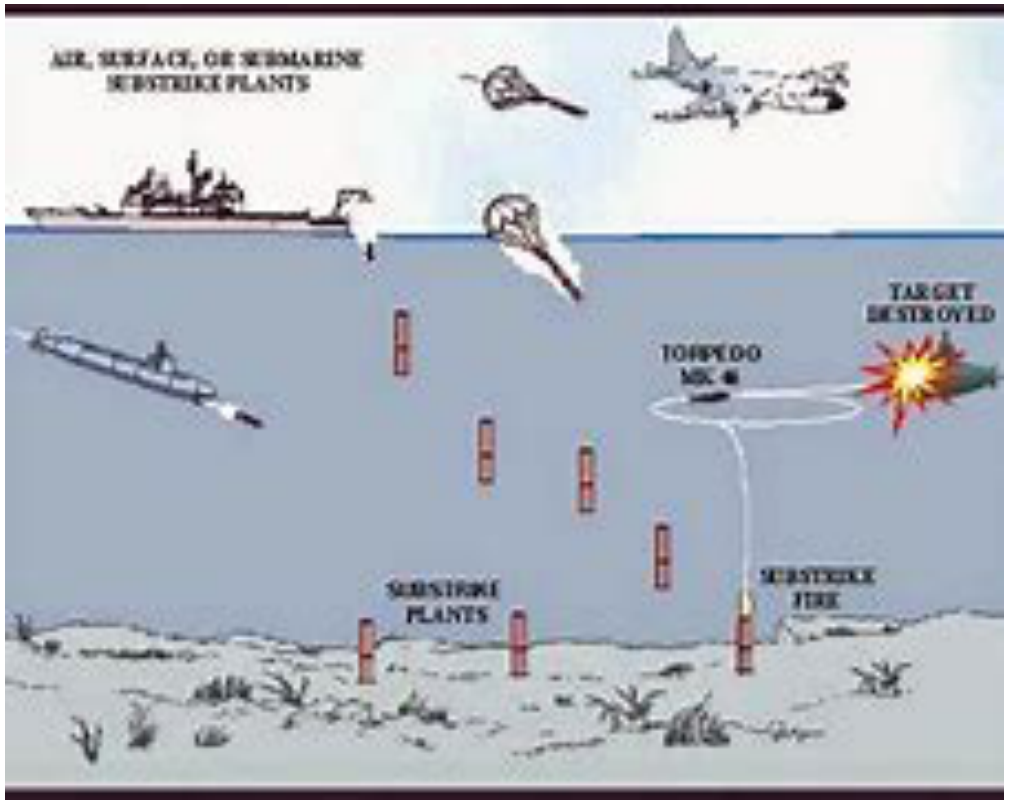
ATACMS + BAT

CBU 105 + SFW

- Two-stage guided munitions delegated authority to the submunition to discriminate, select and engage the final target amongst a group/type of distant targets designated for destruction by a human operator

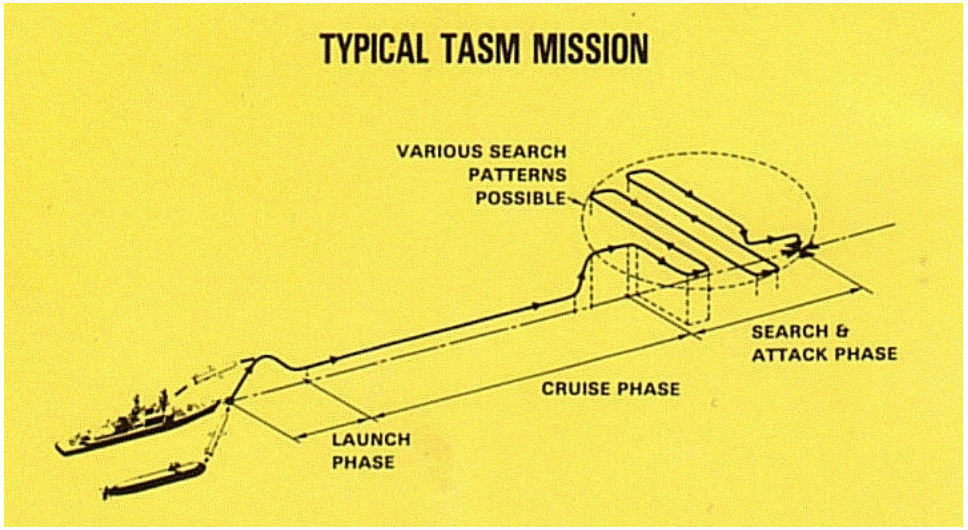
Autonomous weapons: Static search weapons

- CAPTOR

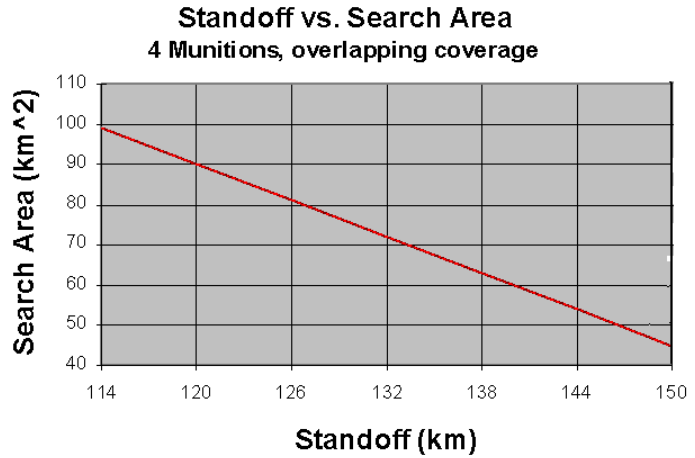


Autonomous weapons: Bounded search weapons

- TASM



- LOCAAS



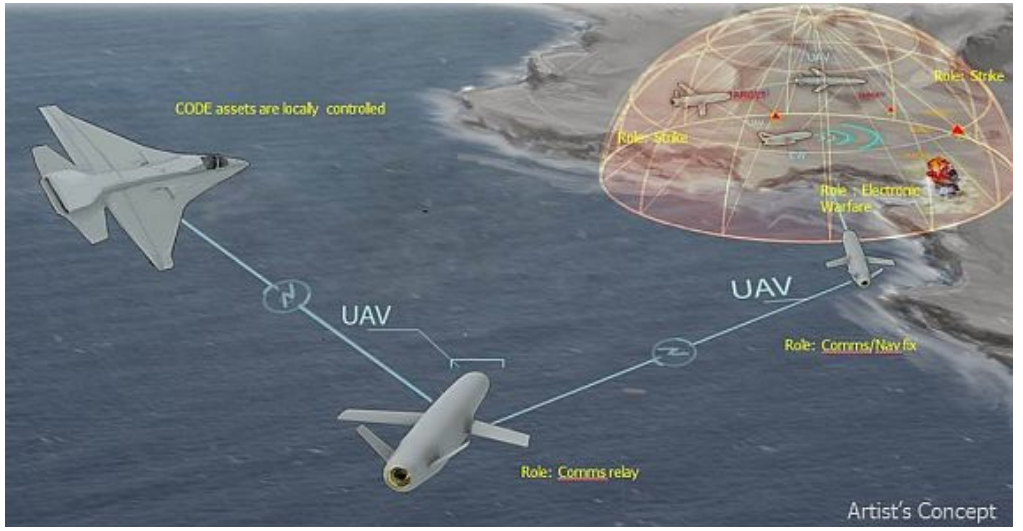
Autonomous weapons: Human-supervised

- Patriot

- Aegis

AI-enabled autonomous weapons: collaborative attack

- DARPA CODE project



- Combat swarms

AI-enabled autonomous weapons: improved target discrimination

- Fewer blue on blue engagements
- Fewer blue on green engagements
- Fewer unintended engagements (civilian casualties)
- Less collateral damage of civilian infrastructure

Key themes for “lethal autonomous weapon” or “fully autonomous weapon”

- “Machines with the power and discretion to take lives without human involvement”
- Weapons capable of deciding a course of action…without depending on human oversight and control”
- “ LAWS should be understood as implying a total absence of human supervision, meaning there is absolutely no link (communication or control) with the military chain of command…”
- Proposed definition: **a LAWS, or FAWS, is an independent, unsupervised, self-targeting weapon system**

Proposed Principles for the Combat Employment of Weapon Systems with Autonomous Functionalities

Prologue

- The law of war does not specifically prohibit or restrict the use of autonomy to aid in the operation of weapons; neither does it expressly approve of its use
- U.S. Department of Defense policy is that *any and all* weapons, including weapon systems with autonomous functionalities, must be developed and used in compliance with the law of war, policy, applicable treaties, weapon system safety rules, ethical guidance, and rules of engagement
- Weapon systems with autonomous engagement functionalities have met this standard for nearly eight decades
- (TEVV) of any munition or weapon system with autonomous functionalities must demonstrate that it can reliably and repeatedly meet mission objectives in realistic operating environments while conforming to the law of war, policy, applicable treaties, weapon system safety rules, ethical guidance and rules of engagement
- In addition to TEVV, a separate legal review of the weapon and its intended use is also required to ensure compliance with the law of war and DOD policy, as is the case for all weapons developed by DoD

Principle #1

- ***Any use of weapon systems with autonomous functionalities must be guided and overseen by a responsible chain of human command and control***

Principle #2

- ***Decisions to initiate a sequence of actions, including autonomous actions, that may result in the loss of human life through the use of force (i.e., a kill chain) are the sole province of human intent and judgment***

Principle #3

- ***Human responsibility for decisions over the use of force cannot be transferred to machines under any circumstances***

Principle #4

- ***To make a valid determination about the lawfulness of an attack on a specific target, persons who authorize the use of, direct the use of, or operate weapon systems with autonomous functionalities must have sufficient information about the system's expected performance and capabilities, doctrine for use, the intended target, the environment, and the context for use (e.g., the presence of non-combatants in the engagement area)***

Principle #5

- ***Once a human being initiates a sequence of actions that is intended to end with the application of lethal force, weapon systems with autonomous functionalities may complete the sequence on their own without further human oversight***

Principle #6

- ***As long as a weapon system's selection and engagement of a target occurs as part of a sequence of actions tied directly to a deliberate human decision to carry out a lawful attack, the standard of appropriate human judgment over the use of lethal force is met***

Principle #7

- ***Commanders must take appropriate action if they obtain evidence that weapon systems with autonomous functionalities may be operating in a manner contrary to expected performance, the law of war, policy, applicable treaties, ethical guidance, and rules of engagement***

BACKUP

DODD 3000.09, Autonomy in Weapon Systems

- Not meant to be a comprehensive policy on the military requirement for autonomy, autonomous operations or autonomous systems and weapons
  - Written by Acquisition, Technology and Logistics (AT&L)
  - Need to read the entire document carefully to discern points

- Applies to "the design, development, acquisition, testing, fielding and employment of ***autonomous*** and ***semi-autonomous*** weapon systems, **including guided munitions that can independently select and discriminate targets"**
  - Introduced new terms for weapons, rather than focusing on autonomous functionalities in weapons
  - Swept guided munitions into the definitions of autonomous and semi-autonomous weapons; in hindsight, this was a mistake

- Intent of the document was clear: to outline the guidelines and processes to minimize the probability and consequences of failures in autonomous and semi-autonomous weapons systems that could lead to unintended engagements

DODD 3000.09 introduced three new types/definitions of weapon systems

- **Autonomous weapon system (AWS**): a weapon system that, one activated, can select and engage targets without further intervention by a human operator
  - Descriptions elsewhere make clear the autonomous functions of interest are target discrimination, selection and engagement

- **Human-supervised AWS (HSAWS)**: a weapon system that, one activated, can select and engage targets without further intervention by a human operator, but is designed to allow human operators to override its operation
  - Generally thought of as human-on-the-loop control

- **Semi-autonomous weapon system (SAWS)**: a weapon system that, once activated, is intended only to engage individual targets or specific target groups that have been selected by a human operator
  - Includes weapons with autonomous functionalities such as acquiring, tracking, and identifying potential targets, cueing potential targets to human operators; prioritizing selected targets; timing of when to fire; or providing terminal guidance to home in on selected targets
  - Includes “fire and forget” and “lock-on-after-launch” homing munitions
  - This definition covered most, but not all, guided munitions fielded before 2012

One key aim of DODD 3000.09 was to avoid making any guided munitions then in the Joint force inventory non-compliant with policy

- Policy therefore aimed to reflect only the tactical capabilities of guided weapons in the Joint inventory at the time
  - **Semi-autonomous weapons** can be used to apply lethal or non-lethal, kinetic or non-kinetic force
    - In cases of degraded or loss communications, system cannot select and engage individual target or specific target groups/types that have not been previously selected by an authorized human operator
  - **Human supervised autonomous weapon systems** can be used to select and engage targets, **with the exception of humans**, for local defense to intercept attempted time-critical or saturation attacks for static defense of manned installations or onboard defense of manned platforms
    - Included Phalanx CIWS; Aegis and Patriot missile systems
    - Established the targeting of humans as a special case
  - **Autonomous weapon systems** can be used only to apply non-lethal, non-kinetic force, such as some forms of electronic attack, against material targets
    - MALD, MALD-J missiles were the exemplar used; both applied non-lethal, non-kinetic force

3000.09 sought to establish a baseline for weapons in the inventory and outline procedures for seeking approval to develop and field new weapons with more advanced autonomous functionalities

- Did not cover all fielded weapons: no mention of the Quickstrike naval mines, autonomous weapons capable of applying lethal force; no mention of two-stage guided munitions
- Did not cover weapons that had been fielded but subsequently retired: no mention of the Cold War CAPTOR mine, another autonomous weapon system capable of applying lethal force
  - If the Department wanted to build a new CAPTOR, it would have to seek new approval to do so
- Did not cover weapons designed and tested but not fielded: LOCAAS
  - Directed hunter-killer guided weapons
    - While the weapon was proven to be operationally effective, it was shelved after Air Force leaders insisted a data link be added to provide human monitoring and intervention after release
    - This could have been an important talking point for subsequent debates
- No mention of potential future swarming weapons or weapons capable of collaborative attack
- Because the directive did not acknowledge or consider autonomous functionalities long approved in weapon systems, some of its definitions were flawed
  - For example, the definition of SAWS is a weapon system that, once activated, is intended only to engage individual targets or specific target groups that have been selected by a human operator
  - But the targets engaged by two-stage guided munitions are selected by the munition, not a human. The human has designated targets for destruction, but delegates authority to the munition to make the final target selection

The debate is hampered by the disagreement over what constitutes a "lethal autonomous weapon"

- U.S. and Campaign to Stop Killer Robots: "a weapon system that, once activated, can select and engage targets without further intervention by a human operator."
  - Campaign is talking about the development of future weapons; DODD 3000.09 suggests the same

- As is Japan (2016): "a weapon capable of pursuing, without human intervention, autonomous deployment and recovery, identification of a target, judgment/decision of the attack and application of lethal force to the target, specifically a human target, **and that such LAWS do not exist at present."**

- As does the United Kingdom (2017): "An autonomous system is capable of understanding higher-level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be."
  - Describes the capabilities suggested by Terminator
  - Not possible with current so-called (second wave) narrow AI"

- France (2016): " LAWS should be understood as implying a total absence of human supervision, meaning there is absolutely no link (communication or control) with the military chain of command…The delivery platform of a LAWS would be capable of moving, adapting to its land, marine or aerial environments and targeting and firing a lethal effector (bullet, missile, bomb, etc.) without any kind of human intervention or validation."
  - Been there, done that: CAPTOR

Autonomy as part of a 3OS

- As the 3OS defined it, autonomy resulted *from freedom to develop and select a course of action required to achieve a higher authority task or objective*
  - Both human operators and intelligent machine systems could enjoy autonomy in battle network operations
  - Importantly, the delegation of authority to intelligent machine systems to develop and select among COAs would be confined to *specific tasks* assigned by a human commander or operator

- The 3OS posited the Joint force might potentially extend its military technical advantage in the near to mid-term by developing and placing into each battle network grid a variety of human supervised *narrow task systems* with improved and expanded AI-enabled autonomous functionalities
  - These human supervised narrow task autonomous systems would be designed to restrict their decisions and performance to a particular problem or domain—meaning the system would be programmed or trained to operate within the bounds of a defined problem and knowledge base
  - In practice, human supervised narrow task autonomous systems would "have a degree of self-government and self-directed behavior with [a] human's proxy for decision"

- The framers of the 3OS deliberately emphasized human supervised narrow task autonomous systems, as opposed to unsupervised *mission* task autonomous systems, which would require AI that more closely mimics human intelligence and reasoning
  - Consensus was such systems would require more advanced third wave "general AI" which was thought to be decades away, even if it was technically achievable

The 3OS emphasizes human supervised, narrow task autonomous systems

- It was hard to imagine any commander trained in the West delegating authority to a machine to clear a village of insurgents, since the commander—not the machine itself—would be held accountable for the machine's actions

- In contrast, it was easy to imagine assigned missions being accomplished by human commanders and operators operating in conjunction with narrow task autonomous systems under their direct supervision
  - These narrow task autonomous systems might manifest themselves through *autonomy at rest*—that is, virtually, in software, in things like intelligence support, planning and expert advisory, and predictive maintenance systems
    - Because the 3OS assumed all these type systems would be designed to help *humans* make better decisions, it described autonomy at rest in terms of ***human-machine collaboration***
  - Narrow task autonomous systems might also manifest themselves through *autonomy in motion*—reflected in the physical world in the form of robotics and autonomous unmanned vehicles, systems and weapons
    - Since the 3OS conceived of these systems as working hand-in-hand with humans to solve complex battlefield problems, it referred to autonomy in motion in terms of ***human-machine combat-teaming***

- Although the 3OS envisioned supervised narrow task autonomous systems as improving human performance, it fully expected that in some cases they might also perform independent battlefield tasks—especially if the tasks were dull, especially dangerous, or involved operations in contaminated environments
  - In these cases, the 3OS envisioned the expanded use of ***machine-machine combat teaming***, often in the form of combat swarms, again under human supervision

The postulated end state of the 3OS was a new type of human-machine collaborative battle network

- The 3OS hypothesized an aggressive insertion of supervised narrow task autonomous systems in all battle network grids would ultimately lead to more powerful ***human-machine collaborative battle networks*** that could make:
  - More rapid sense-making of high heterogeneity and volume of data;
  - More rapid understanding of the operational environment;
  - More rapid development of a common Joint multi-domain operational picture, shared more quickly throughout the force;
  - More rapid development of relevant courses of action and plans;
  - More rapid force-wide understanding and appreciation of commander's intent; and
  - More rapid, more relevant decisions, promulgated faster to manned, unmanned, human-machine and machine-machine combat teams, able to apply faster, more discriminate effects across every operating domain

- The net effect would be a dramatic increase in operational tempo in all domains, which was thought to provide a decisive combat advantage. In addition, widespread autonomous systems and operations would likely allow the Joint force to:
  - Operate more effectively during periods of denied or intermittent communications;
  - Conduct operations requiring high degrees of complexity and coordinated action, especially during multi-domain operations; and
  - Reduce the danger to human operators

Human-machine collaborative battle networks

- The 3OS hypothesized more powerful human-machine collaborative battle networks would lead to:
  - More rapid termination of battles;
  - Fewer blue-on-blue engagements;
  - Fewer blue-on-green engagements;
  - Fewer non-combatant casualties;
  - Less collateral damage of physical infrastructure
- In other words, although never explicitly stated, the framers anticipated human-machine collaborative battle networks would build upon the ethical and moral foundation of guided munition-battle networks