The guilty (silicon) mind

Mens rea in human-machine teaming

Brendan Walker-Munro and Zena Assaad

A human-machine team (HMT) represents a novel technology where a computer or artificial intelligence (AI) system is linked to a human being, to assist or aid the human – in the military context – in completing the objectives of missions or military operations. However, the linking of a human being to an artificial system raises questions about attributing liability: do the actions give rise to civil liability (where the remedy is usually compensation or some remedial order of the court) or criminal liability (where the remedy is usually imprisonment for natural persons, both as a form of punishment and to protect innocent members of society)? In this policy brief, we will discuss some of these liability issues which may arise on the development and potential use of HMT.

Defining HMT

The ADF defines a HMT to be the 'incorporation of autonomous or robotic systems within military teams to achieve tactical outputs that neither machines nor people could deliver independently'. Although this definition has direct application to Australian HMT operations, it does not cover the full scope of possible HMT operations.

Instead, it should be recognised that what makes HMT unique is the "bi-directionality of communication". Existing military capabilities already involve collaboration between humans and machines: a pilot observing a radar screen in a fighter, a soldier using the viewfinder of an anti-tank missile. However, in an HMT, the human can communicate with the artificial element – whether it contains AI or not – and receive feedback on the achieving of shared goals. For example, the machine component of a HMT may communicate with its human capability by displaying sensor information or the projected results of a weapon detonation, whilst a human may instruct the machine component to select, track or engage targets presented.

It is the bi-directional communications between human and machine which pose the biggest issue in determining "who" or "what" is to blame following an accident, misfire or negligent injury. Consider the following examples based on existing military technologies:

- The pilot of an attack aircraft, assisted by uncrewed sensor drones, attacks a convoy based on the drones' assessment of those vehicles as being legitimate military targets. Following an investigation, it is revealed that the convoy contained fleeing refugees and the sensor data was incorrectly interpreted by the drones;
- The captain of a Naval destroyer is linked to the automated defences of their ship. The radar detects an aircraft approaching and assesses its behaviour as benign; the captain, however, believes the aircraft is adopting an attack profile and opens fire. The aircraft was in fact an allied fighter in an adjacent battlegroup; and
- A platoon of soldiers is conducting a patrol in a foreign country, assisted by an armed robotic companion that is teamed with one of the platoon soldiers. Unbeknownst to the platoon, the software under-pinning the robot has been hacked by enemy forces and suddenly presents false threat warnings. The teamed soldier opens fire, killing one of his platoon members.

Issues of Liability In HMT

Under most Western legal systems, individuals carry legal responsibility for their actions – also called "blameworthiness". Where those actions harm another person or damage some property, the individual becomes liable either for restitution or compensation (civil) or to face the punishment of society and discourage other offenders in the future (criminal). It is for this reason that the criminal law imposes a higher standard and burden of proof than in civil law. It can be said then that the criminal law "bans" a risky activity by threat of imprisonment, whilst the civil law "prices" a risky activity by imposing costs for performing it.

This is also why in legal terms, many offences involve the concept of *mens rea* or a "guilty mind". Good motives cannot rescue or defend wrongful conduct. So, who has the "guilty mind" when a HMT makes a bad or wrong decision? Can the human being really be held accountable for errors in programming or code in a HMT? Or is the human completely immune to the reach of the law because any decisions they make cannot be separated from the operation of the machine?

In a legal system where the focus is on the punishment of unlawful conduct or the remediation of breaches of rights, any circumstance influencing the blameworthiness of an agent will have serious ramifications for attribution of liability, because:

> [A]n agent can only be held responsible if they know the particular facts that surround their action, they are able to freely form a decision to act, and are able to select one of the suitable available alternative actions based on the facts of the given situation.

In summary then, the human component of a HMT will face liability for their actions if the following three conditions are met: they have a *knowledge* of the facts of the incident, there were *suitable* *other alternatives* which were not taken up, and the HMT had the *freedom of action* to decide on one of them in the circumstances.

Problems Applying Liability to HMT

Where a machine can be attributed with blameworthiness, there comes the question of how to achieve a penalty or restitution in a manner that is relevant to the machine. Alternately, there is a question of how to apply a remedy to a human who may have had no conscious control of or over the actions they are now alleged to have engaged in. This in turn raises serious questions about liability.

Which actor within an HMT, whether the machine or human actor, is the one 'making' a decision? If the human has taken what they consider was the only 'reasonable' option, are they really making a decision? The decision has already been made by the machine—perhaps inadvertently—by presenting the information in a way that only one option was possible.

HMTs as a technology will also face challenges in a court setting. Much of the technology, automation, or software underpinning HMTs is likely to be protected by trade secrets or military secrecy. Further, the opacity of AI/automation programs in HMT means that even where such the code of such programs can be exposed, the apparent nature of decision-making by that code is not readily discernible in a manner understandable by jurors or judges.Thirdly, each jurisdiction in Australia (and indeed other countries) will have various types of defences for the attribution of blame, including impairment, automatism, or insanity defences. How then will a HMT be assessed as meeting any of those defences where the machine and human element were closely linked? The varying degrees, scope, and application of these defences will lead to entirely varied treatments of HMTs in circumstances where judges are called to assess the 'voluntariness' of actions to assign blameworthiness.

A Proposed Framework for Liability

The resolution of these various concerns with liability is not easily completed. Instead, we suggest that military HMTs might look to "chains of responsibility" to deal with attribution of blameworthiness issues.

HMT to ensure the safety of their individual activities so far as is reasonably practicable. At each level, from design, through manufacture and testing, to 'handover' to military authorities and eventual deployment in military operations, an HMT must be rigorously tested in all intended operational environments. Legal and ethical advice should be sought and incorporated into the design, manufacture, and testing stages. Such testing must be performed by both the manufacturers and military authorities, and testing performed at any specific stage should not be regarded as being conclusive. A 'cut-off' or similar system should always be included in any HMT that permits a human operator (or other person acting remotely) to deactivate the machine component in the event of a failure or incident.

Any safety defects, issues, and injuries must be rigorously investigated and either remediated or repaired, or a mandatory warning provided in relation to conduct likely to cause that issue again. In both training and operational use, military commanders bear an additional non-delegable duty to ensure their staff are trained on HMTs and deemed competent in their use. In the absence of clear legal guidance to the contrary, principles of both domestic and international law should be deemed to always apply to the use of HMTs in operational military environments. In the event of an accident or incident, an investigation is conducted that examines the entire logistic chain to determine where the duty was breached, and by which agent. Breaches of that duty of care may result in the commencement either of civil action (involving pecuniary penalties) or criminal offences (involving potential for penal sentences in severe cases).

Conclusion

What is required for the use of HMTs in Defence is a nuanced and purposeful regulatory regime which considers the reasoning for attribution of responsibility, whilst also providing appropriate mechanisms for restitution and punishment. This is much for the benefit of our armed forces as for the protection of the rules-based global order: military officers and personnel need to know the legal limits of their conduct, what can be done in war and peacetime, and what consequences might attach when they step outside those boundaries.

The exact parameters of technologies designed to constitute HMT and how they are defined in law will need a more comprehensive examination. The definitions will need to be expansive enough to capture those technologies at the forefront of military and civilian research, but also those yet to be contemplated. Alternately, new legal definitions for those technologies will need to be included in their own regulatory regime to eliminate grey areas and ambiguity. Just like our treatment of AI, we need to ensure that the definition is clear, unambiguous, and is not leading to inaccurate or oversimplified definitions of the technology.

Brendan Walker-Munro and Zena Assaad, 'The Guilty (Silicon) Mind: Mens Rea in Human-Machine Teaming' (2023) 8(1) Cambridge Law Review 1–24

This research received funding from the Australian Government's Next Generation Technologies Fund through Trusted Autonomous Systems, a Defence Cooperative Research Centre. The views and opinions expressed are those of the author, and do not necessarily reflect the views of the Australian Government or any other institution. They also do not constitute legal advice.

Law and the Future of War
School of Law
The University of Queensland
law.uq.edu.au/future-war

TRUSTED AUTONOMOUS SYSTEMS
DEFENCE CRC

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA