Reconciling trust and control in the military use of artificial intelligence

Tim McFarland

In efforts to develop AI technologies, the desired human-AI relationship is often framed in terms of trust. In this paradigm, the aim is to make AI systems sufficiently trustworthy and to foster appropriate levels of human trust in AI systems the workings of which may often be inscrutable and the behaviour of which not readily predictable. The notion of trust acknowledges that some uncertainty about the actions an AI system might 'decide' to perform in operation is part of the human-AI relationship.

At the same time, in legal discussions the challenge is most often framed as ensuring that humans will be able to exert appropriate control over AI. The focus is on seeing that responsible human operators can reliably guide the behaviour of an AI system. Much of the opposition in international forums to the continued development of autonomous weapons has related to fears that people might lose control over AI-powered weapons, and such opposition has been met with reassurances also cast in terms of humanity's ability to retain control over its creations.

Obviously, developers of AI and autonomous military systems also speak in terms of 'control' in the sense of control theory and control systems engineering, but this paper deals with 'control' and 'trust' as paradigms which seek to shape the human-machine relationship in ways which may or may not be compatible with each other.

Trust in the human-AI context

The first challenge faced when investigating trust is simply defining it. Jacovi et al offer a formal, contractual model of human-AI trust. This model is based on the notion of a 'contract' between human and AI whereby the human requires the AI to behave in a specific way or perform a specific action in specific circumstances: to accurately navigate a platform or munition from an origin to a destination, to classify a potential target as military or civilian with a clear explanation of the reasoning leading to the assessment, and so on.

In this model, the task of ensuring a successful outcome from human-AI teams which operate based on trust is a matter of calibrating the user's trust to the AI model's capability to maintain the contract in question. In other words, it is a matter of ensuring that trust in the AI is warranted, or that the user's trust corresponds to the AI's trustworthiness.

(Regulatory) control in the human-AI context

The notion of control has featured heavily in the work of the Group of Governmental Experts on Lethal Autonomous Weapon Systems, the primary international forum for discussions about regulating autonomous weapons. The Chairperson's Summary published in April 2021 provides some insight into the international community's views on 'control' over AI:

> Measures based on a concept of human control could require considerations based on the specific characteristics of a weapon, on the operational environment, on the time-frame of autonomous operation, scope of movement over an area and on human-machine interaction. Such measures could also specify: the degree of predictability required in a weapon system's behaviour; the required degree of training and understanding of a weapon system; and the ability of a human to deactivate or override the operation of a weapon system … Effective human control …may not necessarily equate to direct, manual control but rather contextual factors including boundaries placed on the weapon and environment of use, and requirements for human-machine interaction. Further work is needed within the Group to understand various aspects of human control …

Some States have linked the idea of control more closely with direct human involvement in weapon system operation while others have drawn attention to the distributed nature of State control over military force. The common element has been the focus on the capacity of humans to ensure specific (or, at least, sufficiently tightly constrained) behaviours of AI-driven systems. The idea of relying on 'trust' in the system to behave as desired is almost absent from regulatory discussions.

Reconciling trust and control

Existing law is virtually silent on the nature of the relationship that must exist between human beings and the weapon systems they operate. In the absence of settled law, the views being expressed by States and others, the behaviour of systems being developed for use by armed forces and the practices being adopted in the use of those systems are together shaping the norms which will govern the military use of AI. That is why it is important that the paradigms adopted by participants in technical and regulatory efforts are consistent with each other.

Unfortunately, very little work has been done on reconciling trust and control in this context. Some work has been done in the field of organisational governance, albeit in an interpersonal context, and parallels may perhaps be cautiously drawn. Two popular views can be identified in the literature: that trust and control are substitutive, and that they are complementary.

The substitution perspective treats trust and control as being inversely related: low levels of trust require high levels of formal control, and higher levels of trust allow for lower levels of formal control. The complementary perspective sees trust and control as mutually reinforcing, or able to contribute simultaneously to managing a relationship.

One possible resolution begins with the proposition that, as the law already seems to allow room for both trust and control between human and AI, further developments should not impinge on that relationship any more than necessary. The control required by legal frameworks should aim to set limits without intruding unduly into how those limits are met. AI designers should accordingly ensure that the need for operators to trust those systems does not extend beyond the limits set by legally-required control. That in turn would require that legal limits be specified in terms that are compatible with the nature of human-machine trust and would consequently impact understandings of a range of legal obligations, such as the conduct of weapons reviews and accountability regimes. ●

Tim McFarland, 'Reconciling Trust and Control in the Regulation of Autonomous Military Systems' (2022) 30(4) *International Journal of Law and Information Technology* 472–483

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. Cover image by Hanjin / Adobe Stock.

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