UNIDIR

DATA ISSUES AND MILITARY AUTONOMOUS SYSTEMS

EXECUTIVE SUMMARY

by Arthur Holland Michel

All autonomous systems run on data. Indeed, one way to think of "autonomy" is as a process by which machines respond to data inputs with a corresponding output or action, without any human direction. If there are issues with these data inputs, autonomous systems can exhibit suboptimal performance or fail.

Data are therefore central to the ongoing discussion among policymakers about the harms that could arise from the use of autonomous weapon systems and other forms of military artificial intelligence. To be sure, it has been speculated that autonomous systems could potentially exhibit better performance in certain tasks than traditional means or methods of warfare, leading to a reduction in some kinds of unintended harm. But because autonomous system failures could also *increase* harms, any potential future policy actions related to human control and the application of law and responsibility to autonomous military systems, or to the "operationalization" of the Group of Government Expert on LAWS's guiding principles, will likely have to account for the vagaries of data and its effects on these technologies.

THE SCIENCE OF DATA ISSUES

In order to perform as desired, autonomous systems must collect data that are complete, relevant, accurate and of high quality. Most importantly, these data must not differ too much from the data for which the system was developed and tested. But in the real world, the data that any autonomous system will rely upon to execute its mission are unlikely to ever be perfect. Conflict environments are harsh, dynamic, and adversarial. More generally and crucially, there will always be more variability in the real-world data of the battlefield than the limited sample of data on which autonomous systems are built and verified.

Because they are complex systems, autonomous weapons encountering such issues could likewise fail in a complex and unpredictable manner. Specific data issues that cause dangerous failures come to be known through the development of the weapon, the testing and legal review of that weapon, and that system's previous track record. However, such issues are just a subset of all the actual issues a system might encounter in the real world. It is impossible to know every single data vulnerability that any given autonomous system might have, just as it is impossible to predict every single data issue that such systems will encounter. As such, all autonomous systems will be prone to inevitable accidents which cannot be foreseen. We know that such problems exist either now or will emerge in the future, but we cannot characterize or specifically anticipate them. One might call such data issues "known unknowns".

To be sure, all complex weapons systems can have failure modes that cannot be foreseen. But it may be harder to anticipate, quantify and characterize the risks associated with those issues in autonomous weapons due to the inadequacy of present-day testing and verification processes for AI, the difficulty of characterizing AI failure points, the low relative reliability of AI, and the unpredictable conditions and effects of autonomous system deployments. As a result, it will be comparatively more challenging for militaries facing a complex conflict environment to determine whether and how data issues are likely to affect deployed autonomous system and, by extension, where on the scale of reliability and risk that system will perform.

THE IMPLICATIONS OF DATA ISSUES

This has potentially complex implications. International law requires those employing autonomous weapons to anticipate and respond to data issues that could cause unintended harm. While there are a variety of factors that determine the degree to which parties could anticipate and respond to issues – including the mode of human-machine interaction, environmental conditions, and the type of operation – their ability *and thus responsibility* to do so ultimately hinges on whether these issues are known *or* unknown.

The fact that some autonomous system vulnerabilities are "known unknowns" could therefore create ambiguity as to state or individual responsibility for unintended harm resulting from such issues: Because such issues are unforeseeable with current testing and risk assessment measures, those employing these technologies may not be required to anticipate or respond to those issues – but because these issues are also by their very nature inevitable, these actors would need to address or account for them, and could be held responsible for any harm stemming from a failure to do so.

This unusual facet of complex autonomous systems, which mirrors and potentially further confounds questions of how to assign human responsibility to unpredictable machine "decisions", could lead to widely differing interpretations of how the law applies to these technologies. Those actors attempting to make a good faith assessment of the likelihood of unintended harm in the use of an autonomous system will find little guidance on how to navigate the matter of known unknowns. On the other hand, those with a looser threshold of required certainty in operations will have few compunctions about employing a potentially fallible system if that system's vulnerabilities are not specifically known. The unknowability of system errors may even provide convenient legal cover against responsibility for harm that may not, in fact, have been totally accidental.

SOLUTIONS FOR DATA ISSUES

A range of technical approaches are often cited as potential solutions to prevent – or at the very least make known – the failures that data issues cause. These include hybrid intelligent systems, anomaly detection, expanded "training" data sets, and multi sensor collection systems. However, while these approaches show promise, they are all still emerging research areas. And though they may reduce incidences of failure they may also increase the complexity of autonomous systems, thus creating new unknown vulnerabilities. It is therefore safe to say that at least in the near- and mid-term future, technological solutions alone will not resolve the paradox of known unknown data issues.

This suggests that policy options may be necessary to navigate the challenges discussed in this report. A range of such options have been proposed in recent years, including limits or moratoria on use, overhauled liability and due diligence frameworks, enhanced legal review processes, recursive testing and review regimes, and international knowledge-sharing. However, many of these, while potentially helpful, require significant additional research. Furthermore, none of these options is likely to fully address the problem of data issues if implemented in isolation.

The following five avenues for action could bolster efforts to minimize the risks of unintended or unaccountable harms arising from the use of military autonomous systems. Like all international initiatives relating to autonomous military systems, they will require close cooperation between stakeholders from all domains, including governments, militaries, civil society, academia and the technology sector.

1. **Perform advanced, collaborative research on the legal review process.** Legal reviews are likely to be key to addressing data issues. Developing legal review procedures that resolve the many ambiguities described in this report will require significant new research, collaborative dialogue and knowledge-sharing.

2. **Develop classification criteria for data issues and resulting failures; specifically, develop criteria to distinguish *known unknown* issues from *unknown unknown* issues, and frameworks to assign appropriate responsibility in cases of harm arising from such issues.** A finer-grain scheme for differentiating between different types of failure – and a clearer framework designating the actors for whom those failures should be knowable – could aid efforts to quantify risk in operations and assign due responsibility for unintended harm arising from data issues.

3. **Share specific knowledge on technical and normative approaches to data and risk in relation to autonomous military systems.** Given the formidable challenge of characterizing data issues, to say nothing of addressing them through technical approaches, all stakeholders should be encouraged to share knowledge across political and disciplinary divides. This especially applies to sharing of best practices, given that even good faith efforts to minimize the risks of data issues in autonomous systems could be frustrated by the complexity and ambiguity of data issues. A number of militaries already possess significant shareable relevant knowledge (for example, sophisticated risk assessment tools and procedures) that could serve as a foundation for assessing autonomous systems risks; the distribution of these resources would be beneficial for all actors seeking to mitigate the risks of autonomous systems.

4. **Study adversarial measures and their effects on autonomous weapons.** No autonomous system is "unattackable", and many of the most dangerous and unpredictable data issues for autonomous systems could arise from adversarial actions. By foregrounding the science of adversarial measures, the international community will better place itself to model their effects and, as necessary, take adversariality into account in the development of norms or policies for the development and use of autonomous systems.

5. **Adopt a system-of-systems approach to studying data issues.** Failures in autonomous systems arise from the interaction of a range of subsystems: not just sensors and algorithms but also actuators, power sources, communications devices and other systems in the battlespace. Taking all these interacting systems into account will help guide parties to more grounded solutions than discussions that solely focus on the algorithmic element of autonomous technologies.