WRITTEN BY ANNA TUREK WITH RICHARD MOYES

AUTONOMY IN WEAPONS: 'EXPLICABILITY' AS A WAY TO SECURE ACCOUNTABILITY

Article36

WWW.ARTICLE36.ORG
INFO@ARTICLE36.ORG
@ARTICLE36

Article 36 is a specialist non-profit organisation, focused on reducing harm from weapons.

KEY MESSAGES

- Where technologies work in ways that are 'opaque' – such that their functioning cannot be effectively understood or explained – it raises challenges for predicting specific outcomes and ensuring adequate accountability. Such challenges are particularly acute in the context of autonomy in weapons because the outcomes involved include severe harms.
- In the civilian space, policy and legal responses to new technologies have recognised these challenges and have imposed obligations for 'explicability' both as a system requirement and as part of any response to people who experience harm from automated data processing.
- In the context of autonomy in weapons systems, establishing a legal requirement for 'explicability' (as once component of a legal response) would prohibit certain forms of system functioning. It would also provide a basis for scrutiny of technologies under development (such as in national weapon reviews) and would facilitate legal judgements and accountability around the use of systems that are not prohibited.

INTRODUCTION

As the technologies that are used within new weapons become increasingly complex, a number of questions arise with regards to their moral and legal implications. Advances in sensing, data processing, robotics and machine learning, amongst other areas, are producing concerns that find expression in international discussions regarding 'autonomy in weapons systems'. In this paper we will focus on two inter-related aspects of the subject matter, namely:

i. the problem of 'opacity' in the context of 'autonomy' – with opacity presenting a barrier to a user's understanding of a system and therefore, *inter alia*, to predictability of outcomes, and accountability for outcomes; and
ii. the notion of 'explicability' as one form of response to that problem.

In line with recent Article 36 papers, the weapons systems we are considering in this paper fall within a broad category of systems where, after a point of human activation, force will be applied on the basis of data collected by sensors, without human evaluation of that data, and without a human setting the time and place of that application of force.[1]

Within that scope, it is important to note that arguments presented in this paper do not relate to autonomous systems using sensors intended to identify *people* as objects to which force is applied. Article 36 argues that such a process for targeting people should be subject to outright prohibition regardless of the sorts of explicability issues considered in this paper.[2] So the paper is broadly considering systems that use sensors to determine when and where to apply force, but which are not used to target people *per se*.

The purpose of this paper is to outline a way of thinking about how some of the challenges arising from greater complexity in weapons systems could be approached by using the concept of explicability – understood as a basic ethical principle. It is a principle, now finding expression in civil law and policy, that relates to intelligibility of the inner workings of technologies and that can help to enable accountability for their use.[3] The paper reflects on issues of 'opacity' in the context of autonomy in weapons, including in relation to legal obligations and requirements for accountability. It then considers ethics and the notion of 'explicability' as they have been approached in policy and legal responses to machine decision making in everyday life. Finally, it suggests implications from those emerging ethical orientations to issues arising in the military context.

‘OPACITY’ IN THE CONTEXT OF WEAPONS SYSTEMS

The term ‘opacity’ is used when discussing difficulties in understanding *how* certain technologies work and *why* they produce specific outcomes in response to particular inputs. It is recognised that in certain systems (for example, those using neural networks for machine learning) the full processes within a machine through which outputs are arrived at cannot be seen and so cannot be fully understood. In the example of neural networks, designers may understand the initial ‘how it learns’ but they cannot see the detail of ‘what’ it has learned or the subsequent ‘how’ by which certain inputs produce certain outputs. Thus, the value or values produced by such algorithms, in response to external stimuli, are not pre-defined or encoded in the design phase. On the contrary, post-design experience significantly influences the output that will be produced. As a result, we find ourselves in a situation where no one is able to explain why a certain output was produced, beyond the level of ‘after the learning process this is the output produced in response to that specific set of inputs.’ Specific cases may match the pattern of inputs to intended output from the ‘learning’ process, but *how* the specific case is being matched into the pattern is unknown and that *it will* fit the pattern remains a matter of probability.

Systems working in this way can exceed the capabilities of humans and of more conventionally coded software in certain tasks, and for many applications the attendant opacity is of little consequence. For example, however opaque the inner workings of DeepMind’s Alpha Zero, the ‘problems’ that such opacity can produce are entirely limited if it is only playing computer chess. However, the situation changes dramatically if we consider opacity in the context of weapons systems that apply force and as a result cause harms, including loss of human life. *Thus, characteristics of system functioning can come to present challenges or problems when they are considered in a certain context.*

The ‘problem’ of opacity comes to the fore when harms resulting from a system’s operation demand an explanation and the assigning of responsibility. If a system operator’s explanation can add nothing more than, ‘this is just what happened in this situation,’ then it raises questions about the appropriateness of using such a system in such a context.

INCREASING AUTONOMY IN WEAPONS

New technologies are enabling machines to perform more and more complicated tasks with less direct human control or supervision – machines are increasingly autonomous in their functioning. Weapons’ technological capabilities are no exception here. Various already existing weapons systems are autonomous in the sense that, once activated by a human operator, they will then apply force, based on sensor data and algorithmic calculations.[4] Examples include, *inter alia,* anti-material defensive weapons which are deployed to protect specific areas or facilities from incoming attacks with missiles or other type of projectiles.[5]

It is important to recognise that there are limits on how this type of weapons system functions in practice – such limits include, in particular, constraints on the type of targets against which they are likely to apply force (based on a specific and understood target profile), limits to the geographical area where they can operate, and mechanisms to curtail the timeliness and duration of their operation (e.g. a person can switch the system on and off). When we talk of ‘increasing autonomy’ we tend to mean that these constraints are becoming less narrow as weapons systems are enabled to operate without human supervision in wider space, for longer time and/or with less specificity or clarity as to the circumstances that may trigger an application of force.

In this model of ‘increasing autonomy,’ opacity is primarily a problem that derives from decreasing specificity or clarity regarding the circumstances that can trigger applications of force. However, where systems can move autonomously and over longer periods of time, the problem of opacity is likely to be exacerbated:

- Firstly, greater space and duration of operation increases the likelihood of a system encountering circumstances that produce an unexpected outcome, or one that later demands explanation. This is to say, circumstances that bring out problems resulting from opacity in the process triggering force are more like to occur.
- Secondly, any opacity in the wider process of system functioning becomes implicated in the outcome, suggesting an expansion from uncertainty about why force was applied in a particular situation, to uncertainty as to how the system came to be in that situation in the first place.

In the paragraphs that follow we will briefly present the main challenges that opacity and autonomy in weapons create for compliance with international humanitarian law (IHL).

FUNDAMENTAL LEGAL OBLIGATIONS ARISING FROM IHL

IHL that governs the conduct of hostilities imposes a number of legal obligations on belligerents. Obligations particularly relevant to the use of weapons regulate the conditions under which force may be applied and how that application of force is undertaken. Put in a broadly chronological order they include obligations in the following areas:

- before use of force: assessment of proportionality in attack, precautionary measures to be undertaken before the attack;
- during the attack: precautionary measures to be undertaken during the attack;
- after use of force: obligation to hold accountable persons alleged to have committed or to have ordered to be committed grave breaches of IHL.

PROPORTIONALITY AND PRECAUTIONS IN ATTACKS

Both principles of proportionality and precautions in attacks constitute norms of customary IHL as well as part of treaty law (Additional Protocol I[6]). In accordance with the principle of proportionality: “Launching an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated, is prohibited.”[7] In

accordance with the principle of precautionary measures in attacks: "In the conduct of military operations, constant care must be taken to spare the civilian population, civilians and civilian objects. All feasible precautions must be taken to avoid, and in any event to minimise, incidental loss of civilian life, injury to civilians and damage to civilian objects."[8] In order to comply with these obligations those responsible for planning and carrying out attacks are required to make informed, context-specific, value-based legal judgements on each individual attack.[9] These assessments are dynamic and depend on the circumstances present on the ground that may change rapidly in the situation of armed conflict.

Using functionally opaque weapons systems would impede military commanders from making the required legal judgments. At a practical level this could result in the risk of excessive or otherwise inappropriate death, injury or damages that cannot be effectively justified under IHL rules. Similar risk arises when we think of increasing autonomy in weapons in terms of expanding systems' independent operation in time and space. For example, we should note that the legal judgement made by a commander to use a system may not continue to be valid if circumstances on the ground change in a way that would require them to cancel or suspend the attack. Thus, if once activated a weapon system will search for a target for long time and across a wide geographic area, and a military commander is not able to adjust their decision if circumstances change, there is a significant risk of force being applied inappropriately.

At one level, opacity and autonomy create risks of excessive or inappropriate harm in relation to existing legal rules. Beyond that, it can be seen that at a certain level these characteristics make claims of rule application implausible. For example, it is not possible to claim to have 'strictly implemented the rule of proportionality' whilst also acknowledging that one doesn't practically understand the circumstances that will result in a system applying force, or that one doesn't actually know the context within which that force will be applied. There is therefore a threshold of understanding regarding both system functioning and the context of use that must necessarily be met if claims to have duly applied legal rules are to be plausible.

ACCOUNTABILITY

In accordance with the provisions of the 1949 Geneva Conventions[10] states have undertaken to search for and prosecute those who have committed or ordered to commit any of the grave violations of IHL. The potential and capacity to assign individual criminal responsibility for unlawful actions is an important mechanism for building and promoting observance of legal rules. On the one hand it is aimed at deterring future crimes by showing that their commitment is inevitably followed by trial and punishment and on the other hand it is indispensable for bringing justice for victims and their families. Yet assigning individual criminal responsibility for harms resulting from the operation of systems that have been approved for use but which function in an opaque manner would face serious obstacles. Holding a person criminally liable requires proving fault in his or her behaviour. This would be difficult if those against whom charges would be brought, had not been in a position sufficiently to understand how a weapon system identifies targets and applies force, or if they were not in a position to exercise sufficient control over where and when applications of force might occur.

If the individual that gives final authorisation for a system's use is operating within a framework that approves such use, and that approves of their limited personal understanding of the system and of the specific context, then it is difficult to hold them personally accountable for harms that result from such limitations of understanding. A tension arises then between individual responsibility for choices in attacks and the wider bureaucratic structures that develop, produce and approve systems for use. If the limitations of understanding arising from opacity and autonomy do not bear on the individual user but are diffused into that wider bureaucratic structure then both the meaning of legal observance and the pressure to ensure it are diffused also.

Just as greater 'autonomy' could serve to push towards wider and longer notions of 'an attack' as an important unit of human legal application in the law, so opacity and autonomy can serve to diffuse the specificity of responsibility and reasonable accountability. In both cases there is a weakening of the fabric of the law resulting from a prioritisation of generalisations and prior assumptions over contextually more specific and active human deliberation. This in turn would significantly hinder judicial oversight and scrutiny of actions taken by the militaries on the battlefield. Increasing vagueness in combat situations, and in the meaning of the law, would lead to lack of certainty in the courtroom, making norms of IHL unenforceable and lowering the standards of protection IHL should provide.

ETHICS AS AN UNDERLYING FOUNDATION FOR RESPONDING TO OPACITY AND AUTONOMY

Various stakeholders, including policy and law makers, industry and academia, in the field of emerging technologies look to ethical principles as a starting point for thinking on how to shape a regulatory framework for the development and use of new technologies so that they can positively serve humankind. There is a wide consensus among experts and governments acting within the frameworks of international organizations, such as the European Union (EU) or the Organisation for Economic Co-operation and Development (OECD)[11], as well as G20[12], that values-based, ethical approaches constitute a bedrock for developing and implementing new technologies. Only when fundamental principles, in particular: respect for human rights, respect for human autonomy, prevention of harm and accountability are abided by, can societies fully benefit from the technologies and innovations that they bring. As examples we note in particular the Code of ethical conduct for robotics engineers adopted by European Parliament,[13] the *Ethics Guidelines for Trustworthy Artificial Intelligence* proposed by expert group set up by the European Commission,[14] and the Institute of Electrical and Electronics Engineers (the IEEE) *Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.*[15] These approaches assume that for certain innovative technological solutions to be adopted in society it is essential that they gain public acceptance and are thought to be 'trustworthy'.[16] This in turn is considered only to be possible if the development and use of the technologies in question occurs while observing ethical principles,

principles which reflect commonly shared values, ideals and understandings, and which constitute a shared foundation for trust. Thus, ethics provide an underlying foundation upon which rules governing design, development and implementation of technologies should be based.

Away from weapons systems, issues of autonomy and related opacity are present as broad thematics, internationally recognised in the civilian space as raising particular concerns and as demanding ethically founded responses. Examples of these can be found not only in the form of policies or codes of conduct mentioned above, but also in binding laws that have already been brought into force. Prominent examples come from data protection legislation enacted in 2016 by the European Union.

The General Data Protection Regulation (EU) 2016/679 (GDPR), applicable in civilian space, provides specific regulations regarding decisions made by algorithms that produce legal effects concerning an individual or similarly significantly affects him or her. The mere fact that such decisions are planned is sufficient to categorise an operation as 'likely to result in a high risk to the rights and freedoms of natural persons' and thus to trigger the obligation for 'controllers' (those responsible for personal data processing) to conduct a thorough assessment process aimed at identifying, evaluating and mitigating those risks. Additionally, controllers are obliged to inform 'data subjects' (individuals whose data shall be processed), upon collection of personal data, about the existence of automated decision-making and to provide 'meaningful information' about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. Subsequently, when in an individual case a decision has been made, data subjects must be guaranteed with at least the right to obtain human intervention from the controller; the right to express his or her point of view; and the right to contest the decision originally produced.[17]

Similar guarantees can be found in Data Protection Law Enforcement Directive (EU) 2016/680, which regulates personal data processing for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties. The Directive requires that the EU Member States shall presume a decision based solely on automated processing, including profiling, which produces an adverse legal effect concerning the data subject (or significantly affects him or her), to be prohibited unless authorised by the EU or Member State law to which the controller is subject. Overcoming the presumption of prohibition requires ensuring appropriate safeguards for the rights and freedoms of the data subject and at least the right to obtain human intervention on the part of the controller. In the preamble of the Directive it is indicated that 'suitable safeguards' in such cases include provision of specific information to the data subject and the right to obtain human intervention, in particular to express his or her point of view, and to obtain an explanation of the decision reached after such assessment or to challenge the decision. Additionally, the Directive also imposes an obligation on Member States to ensure that the controllers are obliged to carry out data protection impact assessments if a given processing operation 'is likely to result in a high risk to the rights and freedoms of natural persons' (however, unlike the GDPR, it does not say that automated decisions per se should trigger the obligation to carry out the required impact assessment).[18]

These examples clearly highlight that automated decisions affecting the wellbeing of people are already recognised as raising particular concerns and requiring particular legal responses. The types of response demanded by the law relate to ensuring awareness and understanding on the part of those who may be subject to such algorithmic decision-making that they may be subject to such a process, how it works, and what the consequences may be, and to providing recourse to human intervention and review in situations where an initial outcome is contested. In both the prior and the responsive obligations, the notion of 'explicability' provides a common thread.

EXPLICABILITY

Explicability in general terms requires that those who design, deploy or are affected by technologies understand how these technologies work.[19] Both what can actually occur from the use of technologies as well as what is intended to be achieved should be understood by technology users.

In the context of weapons two broad domains of explicability should be considered:

- Firstly, 'how autonomous weapons systems work' must be sufficiently clear for their designers and militaries that intend to use them (inner workings of weapons must be practically intelligible before these weapons can be deployed).
- Secondly, it is necessary to ensure that if the use of a system causes harm then an adequate explanation can be provided as to how the event in question occurred (explicability necessary to facilitate accountability).

Thus, explicability can be considered vital for various stakeholders - for those that might be affected by unintended results of the use of autonomous weapons systems, but also for designers, manufacturers, militaries and governments deciding on their development and use. It should also be noted that the requirements of explicability, as sketched out here, are not explicitly required by the letter of the law (in terms of IHL) – rather they are ethically demanded, if the law is to be practically implemented towards its ethically derived purpose.

In the general civilian discourse, the importance of the principle of explicability in the context of new technologies has been widely recognized by various stakeholders – from academia and law makers. Notably, it has been incorporated into *Ethics Guidelines for Trustworthy Artificial Intelligence* prepared by High-Level Expert Group on Artificial Intelligence set by the European Commission in 2018[20], as one of four basic ethical principles that characterize 'trustworthy Artificial Intelligence'. A very clear statement on the subject matter has also been made by the European Parliament in a resolution on Civil Law Rules on Robotics, noting:

> *'(...) it should always be possible to supply the rationale behind any decision taken with the aid of AI that can have a substantive impact on one or more persons' lives;*
> *(...) it must always be possible to reduce the AI system's computations to a form comprehensible by humans;*
> *(...) advanced robots should be equipped with a 'black box' which records data on every transaction carried out by the machine, including the logic that contributed to its decisions.'* [21]

Engineering communities also consider explicability as one of the basic principles that should be taken into consideration while developing autonomous and intelligent systems. A prominent example is provided by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems – a broad initiative that drew on more than a thousand experts from six continents, representing not only engineers active in the industry, but a wide range of stakeholders including civil society and academia.[22] These are just a few amongst many initiatives towards ethical standards for developing technologies all of which clearly state the need for securing explicability in terms of how technologies work, why they produce specific outcomes or why they act in a certain way.[23]

In the documents noted above, the importance of explicability is generally considered to be dependent, in part, on the type of unwanted consequences that can be produced by a given technology. In general terms, the more severe the consequences of the technology's effects (whether it is functioning properly or malfunctioning), the more vital explicability becomes.[24] In the GDPR, for example, as a general rule, any new products or services that involve personal data processing must be subject to prior assessment (in the design phase) in order to ensure that they comply with obligations resulting from GDPR. However, if it is likely that a given processing might result in a high risk then there are obligations to carry out additional measures.[25] In general terms, the 'risk-based approach' adopted in GDPR requires implementing measures adequate to the risks that a given processing involves - the higher the risks are, the higher standards for protection of rights of individuals that are required. Of course, harms inflicted through weapons can be very severe, including the loss of life and long-term impairments. Furthermore, given that the capacity to cause such harms is fundamental and intrinsic to weapons systems, it is arguably all the more important that attention is paid to the processes that underpin that capacity.

TOWARDS EXPLICABILITY IN WEAPONS SYSTEMS

Against this background, it is appropriate to try to develop thinking on how explicability as an ethical principle could be embedded into new legal rules, policies or practices aimed at addressing some of the challenges posed by opacity and autonomy in weapons systems that we sketched out earlier. An important initial point to pin down might be the recognition that IHL, as a legal framework regulating armed conflict, recognises socially derived ethics as a source for its obligations. 'The requirements of public conscience' are recognised as the fundamental point of reference for determining the rules that should be identified and applied to protect civilians and belligerents in cases not covered specifically by existing laws.[26] Whilst this, 'the Martens Clause', may not point to specific regulatory responses, it does clearly indicate the relevance of an external ethical reference point as basis for further legal development.

In the sections below we suggest ways in which 'explicability' could be used in developing a regulatory structure in relation to autonomy in weapons systems.

PROHIBITING THE USE OF 'INEXPLICABLE' WEAPONS

First of all, explicability can serve as a criterion for determining whether a given type of weapons system should be subject to outright prohibition at the international level. Based on the material presented here, weapons which would be complex and opaque to an extent that excludes the possibility of understanding the logic behind their functioning and thus to foresee how they will operate, should be prohibited. Such a rule could be articulated at a broad level, with specific technical formulations presented alongside such a general rule. For example, Article 36 has argued elsewhere that systems where 'target profiles' might change after a system's activation and without human authorisation would fall foul of a requirement for explicability. Likewise, systems where 'target profiles' are built on the basis of machine learning such that a human operator does not know the actual pattern of physical characteristics that will trigger an application of force may not be sufficiently explicable.[27] Rules in this area could be formulated as prohibitions, but also as positive obligations to ensure explicability in any systems that are developed.

EXPLICABILITY *ANTE BELLUM* - NATIONAL WEAPONS REVIEW MECHANISMS

Article 36 of Additional Protocol I to the Geneva Conventions imposes obligation for states parties to conduct legal assessments of new weapons and means or methods of warfare.[28] Implementation of this obligation requires states parties to Additional Protocol I to adopt national weapons review mechanisms. Introducing an international legal prohibition on 'inexplicable' systems, or a positive obligation toward explicability, would create a legal test against which national review mechanisms would need to operate. Implementing such a general requirement into national weapons review processes could provide an important mechanism for promoting explicability as a consideration through the conception, design and development of new systems. Promoting engagement with these requirements early in the development process would facilitate explicability at later stages, such as in situations of use and in any subsequent assessments of resultant harms.

By comparison with many other areas of technology, assessments of specific new weapon systems at a national level are often complicated by (inter-related) issues such as:

- narrow conceptions of the purpose of reviewing technology (e.g. to assess explicit legality/illegality as opposed to underpinning ethical matters);
- divergent interpretations of existing IHL requirements and divergent review methodologies;
- that fact that the technology is often intended to cause some form of harm (and therefore the boundaries between appropriate or inappropriate harms are often uncertain or contingent);
- that populations likely to experience intended or unintended harms are not represented as stakeholders in assessments;
- national and corporate preferences for secrecy vs transparency and accountability;
- bureaucratic acceptance of processes within the military domain that don't rise to the standards expected in other areas of society.[29]

Coupled with the relatively low number of states that have indicated consistent implementation of such reviews (despite it being a legal obligation), these factors posit against a reliance on national review mechanisms *alone* to respond to challenges posed by opacity and autonomy.

Without clear legal obligations to prohibit use of inexplicable systems and to ensure explicability in the development of systems, such weapon reviews would lack a clear point of reference against which to interrogate these matters. However, given such a reference point, national review process would be an important tool for the management of future technologies.

EXPLICABILITY *IN BELLO* - FACILITATING LEGAL JUDGEMENTS ON EACH ATTACK

As we noted earlier in this paper, explicability is also essential for meaningful legal judgments on proportionality and precautionary measures that military commanders are obliged to make when planning and deciding on launching each attack. It is indispensable that military commanders understand how weapons that they intend to use work, what are their limitations and what risks their use involves. Explicability here relates to how weapons systems function – ensuring it is possible to be meaningfully informed on such matters as:

- how a target object is identified by a system;
- what might be identified as a target erroneously;
- in what form is force applied and on what scale;
- what mechanisms are available to change or cease a system's operation.

Explicability in the use of systems would also relate to the context in which a system would function and within which force might be applied. Understanding on these points, and others, would seem to be necessary in order for a military commander to make informed and justified legal judgements.

For example, if in the area where a system will operate, military objectives can be found together with civilian objects, a military commander will only be able to make required legal judgements if he or she is aware of the mechanisms applied within weapons systems to recognize specific objects as targets, and assess the risks of 'false positives' given the technical limitations of the system in question. A commander would also need to understand the form and scale of force that would be applied in order to assess any risks of wider harm even if a military objective is being identified as a target.

Prohibiting the use of systems that cannot be explained and requiring explicability of systems during development would both enable explicability in use. However, positive legal obligations to ensure meaningful human control in the use of systems would be necessary to ensure that the ethical effect of these requirements are transitioned through into practice. For example, positive obligations to sufficiently constrain the location and duration of a system's functioning to enable meaningful legal judgements, could also be framed in terms of making those human legal judgements explicable.

Embedding obligations for explicability into the law would strengthen and reinforce the application of existing *in bello* requirements and could in turn be facilitated through good practice in information provision, technical manuals and training. The purpose here would be to ensure that those who use weapons systems understand how these weapons function, what are their capabilities as well as limitations, and how these relate to possible effects or patterns of effect. It is from such understandings, coupled with understandings of these context of use, that meaningful legal judgements can be made.

EXPLICABILITY *POST BELLUM* - FACILITATING ACCOUNTABILITY

Explicability is also very important after force has been used, especially if unlawful harm may have been inflicted as a result of that use of force (or if a dispute arises as to whether a given use of force was lawful or not). Explicability in this context can enable analysis by courts or other assessment bodies in relation to the inner workings of systems, and so enable determinations as to where in a chain of events critical decisions led to the effects experienced. It would be an important tool for untangling issues of human intent, human judgment, human error, technical malfunction or unforeseen technical circumstances which could all point responsibility in different directions. Understanding how systems work and why they produce specific outcomes is essential for determining who should bear responsibility for harms or violations.

Again, a broad positive obligation to ensure sufficient explicability in systems being developed would be the fundamental starting point. It is, however, argued that computerised technologies enable detailed documentation and logging of events and processes that occur during their functioning.[30] Such capabilities enable a movement from explicability as to how a system works in general to the documentation of specific cases. Such a capability will not be feasible in all circumstances – notably those in which any sensing and algorithmic decision making occur within a unit that is destroyed in the process of applying force (such a certain 'sensor fuzed' artillery shells currently deployed). However, where such a capability is feasible it would seem to offer the prospect of greater clarity and accountability. Such mechanisms could be required where feasible, or promoted as a presumption, and would help to establish, in individual cases, how a specific event occurred – *inter alia* when the weapons system was activated, what contextual data and sensor data matched the target profile, where and when, and what system events occurred subsequently. Clearly, however, mechanisms to facilitate post-use accountability must not be taken as a basis for allowing systems, or system uses, that continue to present prior concerns with respect to ensuring meaningful human control.

The ability to evaluate the inner workings of systems after the fact does not alleviate the need for explicability and practical intelligibility for system users ahead of time. However, understanding how systems work and the logging of actual system events are only likely to work in favour of greater accountability.

CONCLUSIONS

As presented here, explicability can provide a general ethical principle that might thread from the development of a system, through use, and into subsequent determinations of responsibility and accountability. It is a concept that is accepted as fundamentally important in other social domains that are responding to challenges presented by opacity and autonomous decision making. Other domains have also already recognised that algorithmic decision-making requires a specific legal response where it has implications for people's wellbeing.

Embedding explicability obligations that apply *ante bellum*, *in bello* and *post bellum*, as outlined above, would contribute significantly to ensuring morally and legally adequate controls on autonomous weapons systems (as defined and caveated earlier in this paper). Along with a prohibition on targeting people, and positive obligations to ensure meaningful human control in the use of systems, prohibitions and obligations regarding explicability are vital components for an international legal instrument addressing autonomy in weapons systems.

Practically understanding the workings of weapons, their capabilities and risks that their use involves - combined with understanding the context for their use - constitute the essential underpinnings for the application of IHL and for other policy constraints on the choice of weapons in the use of force. Building recognition of the importance of such understandings may seem very modest in its implications, yet it is a central requirement for enabling control over emerging technologies as well as to exerting better controls over weapons already widely in use. The discursive space of technical understandings and explicability regarding weapons systems is a space that critical assessments working to strengthen civilian protection need to occupy.

A THOUGHT ON THE LEGAL TRAJECTORY

Part of the argument in this paper seeks to leverage the growing legal acceptance, in the civilian space, that people who experience harm as a result of machine decision making have rights to certain forms of redress, including human explanation of that decision. The legal structures provide mechanisms of recourse for those subject to such processes (victims, perhaps, in the case of harm being experienced), as well as obligations of prevention and response for those who establish and use such systems. These legal structures, such as the EU's 2016 GDPR, have been developed subsequent to 1977 Additional Protocol I of the Geneva Conventions and they do not have a straightforward corollary in existing IHL. Technologies, and technologically enabled bureaucratic and commercial processes, have been developed that were recognised as producing the need for this legislation. Further developing and adopting analogous technologies into the military space, without adopting rules that respond to the concerns that have been identified in civilian life, would be a reckless abandonment of identified moral responsibilities. It would also amount to an active expansion of the extent to which IHL derogates from the protections that would be demanded in other circumstances.

ENDNOTES

1 'Autonomy in weapons systems: mapping a structure for regulation through specific policy questions', Article 36, 2019, http://www.article36.org/wp-content/uploads/2019/11/regulation-structure.pdf.

2 See the proposed structure for regulating autonomous weapons in: 'Autonomy in weapons systems: mapping a structure for regulation through specific policy questions', Article 36, 2019, http://www.article36.org/wp-content/uploads/2019/11/regulation-structure.pdf.

3 L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, Ch. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4 People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, p. 700, https://www.researchgate.net/publication/329192820_AI4People-An_Ethical_Framework_for_a_Good_AI_Society_Opportunities_Risks_Principles_and_Recommendations.

4 Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) 11-15 April 2016, Geneva, Views of the International Committee of the Red Cross (ICRC) on autonomous weapon system, 11 April 2016, p. 2; https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system.

5 Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) 11-15 April 2016, Geneva, Views of the International Committee of the Red Cross (ICRC) on autonomous weapon system, 11 April 2016, p. 2; https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system.

6 Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977 (Protocol Additional I).

7 J.M. Henckaerts, L. Doswald-Beck, Customary International Humanitarian Law, Volume I: Rules, Cambridge 2005, p. 46, Rule 14; Article 51.5. b) and Article 57.2. a) iii) and b) of the Protocol Additional I.

8 J.M. Henckaerts, L. Doswald-Beck, Customary International Humanitarian Law, Volume I: Rules, Cambridge 2005, p. 51, Rule 15; Article 57 and 58 of the Protocol Additional I.

9 See: 'Meaningful Human Control, Artificial Intelligence and Autonomous Weapons', Article 36, 2016, p. 5, http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf; see also: C. Pilloud, J. Pictet, Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949, edit. Y. Sandoz, Ch. Swinarski, B. Zimmermann, Geneva 1987, p. 684.

10 Article 49 of Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field. Geneva, 12 August 1949; Article 50 of Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea. Geneva, 12 August 1949; Article 129 of Convention (III) relative to the Treatment of Prisoners of War. Geneva, 12 August 1949; Article 146 of Convention (IV) relative to the Protection of Civilian Persons in Time of War. Geneva, 12 August 1949.

11 OECD Council Recommendation on Artificial Intelligence, 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

12 G20 AI Principles, 2019, https://www.meti.go.jp/press/2019/06/20190610010/20190610010-1.pdf.

13 European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html.

14 Ethics Guidelines for Trustworthy AI, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, 8 April 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

15 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https://standards.ieee.org/content/

ieee-standards/en/industry-connections/ec/autonomous-systems.html.

16 See e.g. Ethics Guidelines for Trustworthy AI, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, 8 April 2019, p. 4: "In a context of rapid technological change, we believe it is essential that trust remains the bedrock of societies, communities, economies and sustainable development.", https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

17 See: Article 13.2 (f), Article 22.3 and Article 35 of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

18 Article 11, Article 27 of Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA; see also: Recital (38) of the preamble to this Directive.

19 L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, Ch. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4 People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, p. 700, https://www.researchgate.net/publication/329192820_AI4People-An_Ethical_Framework_for_a_Good_AI_Society_Opportunities_Risks_Principles_and_Recommendations.

20 Ethics Guidelines for Trustworthy AI, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, 8 April 2019, p.12, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

21 European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html.

22 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

23 As examples we note also: (i) the Montreal Declaration for a responsible development of artificial intelligence (this document refers to 'intelligibility' and 'transparency' requirements: <<The decisions made by AIS affecting a person's life, quality of life, or reputation should always be justifiable in a language that is understood by the people who use them or who are subjected to the consequences of their use. Justification consists in making transparent the most important factors and parameters shaping the decision, and should take the same form as the justification we would demand of a human making the same kind of decision>>), https://www.montrealdeclaration-responsibleai.com/the-declaration; (ii) EU guidelines on ethics in artificial intelligence: Context and implementation, European Parliamentary Research Service, September 2019 (this document refers to 'transparency' and 'explainability': <<The wide-ranging concept of explainability is about making explanations on an algorithmic decision-making system available. The requirement for explainable AI addresses the fact that complex machines and algorithms often cannot provide insights into their behaviour and processes.>>), https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf.

24 Ethics Guidelines for Trustworthy AI, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, 8 April 2019, p. 13, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

25 Article 25, Article 35 GDPR.

26 This principle, known as "Martens clause" has been first articulated in the Convention with respect to the laws of war on land (Hague II) of 1899: "Until a more complete code of the laws of war is issued, the High Contracting Parties think it right to declare that in cases not included in the Regulations adopted by them, populations and belligerents remain under the protection and empire of the principles of international law, as they result from the usages established between civilized nations, from the laws of humanity, and the requirements of the public conscience". Martens clause has been since then adopted in various international treaties, including, inter alia, Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land of 1907 or the Protocol Additional I.

27 See: 'Target Profiles', Article 36, 2019, http://www.article36.org/wp-content/uploads/2019/08/Target-profiles.pdf

28 In accordance with Article 36 of Protocol Additional I: "In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party."

29 For some background discussion on the challenges of weapon reviews see Brian Rappert, Richard Moyes, Anna Crowe, and Thomas Nash, 'The roles of civil society in the development of standards around new weapons and other technologies of warfare', International Review of the Red Cross, Volume 94 Number 886 Summer 2012, available at: https://www.icrc.org/eng/assets/files/review/2012/irrc-886-rappert-moyes-crowe-nash.pdf

30 Liability for Artificial Intelligence and other emerging technologies, Report from the Expert Group on Liability and New Technologies – New Technologies Formation set up by the European Commission, p. 47, https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608.

Acknowledgements:

Lead author on this paper was

Anna Turek
Attorney-at-law, member of the Human Rights Section at Warsaw Bar Association
annaturek.3@gmail.com