Israel Law Review (2025), 1–24
doi:10.1017/S0021223725100071

CAMBRIDGE UNIVERSITY PRESS

ARTICLES

Applying Precautions in Target Verification with AI Decision Support Systems

Renato Wolf

University of Queensland, Brisbane, Australia and University of New South Wales, Canberra, ACT, Australia

Email: r.wolf@uq.edu.au

Abstract

This article examines how the obligation to take precautionary measures to verify targets under international humanitarian law (IHL) can be applied with artificial intelligence decision support systems (AI-DSS). It uses the reported deployment of systems like 'Lavender' and 'Where's Daddy?' by Israel in the Gaza War as an illustrative example, breaks down the use of AI-DSS into stages – legal qualification, classification, and identification/location – and evaluates how precautions to verify can reduce the risk of false positives in each of these stages. It argues that precautions to verify must be applied at all stages, and discusses factors that affect their feasibility. The article concludes that while human oversight remains essential, precautions specific to AI-DSS outside the realm of the human operator are possible, and at times, necessary to ensure compliance with IHL.

Keywords: weapons; distinction and precautions under IHL; artificial intelligence decision support systems; Israel–Gaza War

1. Introduction

Military artificial intelligence decision support systems (AI-DSS) are 'computerized tools … designed to assist humans at different levels in the chain of command to complete decision-making tasks', making 'use of AI to benefit from powerful computing tools to better collect, integrate, manage and analyse large and complex data sets'.[1] Such systems may, for instance, propose the targets that could be attacked, calculate the collateral damage expected from an attack, or suggest a certain course

[1] International Committee of the Red Cross (ICRC) and Geneva Academy of International Humanitarian Law and Human Rights, 'Artificial Intelligence and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts: Current Developments and Potential Implications', 13 May 2024, 8, https://www.icrc.org/en/publication/expert-consultation-report-artificial-intelligence-and-related-technologies-military; see also Klaudia Klonowska, 'Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare' (2020) 23 *Yearbook of International Humanitarian Law* 123, 124.

of action.[2] In contrast to autonomous weapon systems, AI-DSS by definition have a human in the loop – that is, they may propose a certain course of action or a person as a target, but the ultimate decision remains with a human operator.[3]

AI-DSS are going to be part of future warfare.[4] Understanding how, in practical terms, the risks that the use of AI-DSS may pose to civilian populations can be reduced is therefore highly relevant. Since many AI-DSS appear to be used to support the selection of targets, and thereby may contribute to civilians or civilian objects being falsely made the object of an attack, measures to avoid or reduce that risk are particularly crucial. In legal terms, these measures are mandated as precautionary measures to verify that the targets to be attacked are military objectives (precautions to verify), most prominently enshrined in Article 57 of Additional Protocol I to the Geneva Conventions (AP I).[5]

The need to take precautions to verify has been underlined in many academic publications that discuss AI-DSS from the perspective of international humanitarian law (IHL).[6] However, the discussion of precautions to verify in the use of AI-DSS to date has focused on human operators, in particular regarding the time in which they have to decide, or their knowledge and biases.[7] Some actors, including Israel

[2] Anna Nadibaidze, Ingvild Bode and Qiaochu Zhang, 'AI in Military Decision Support Systems: A Review of Developments and Debates', Center for War Studies, University of Southern Denmark, November 2024, 6–7; Sarah Grand-Clément, 'Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain' (2023), United Nations Institute for Disarmament Research (UNIDIR), 15–16, https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI_Beyond_Weapons_Application_Impact_AI_in_the_Military_Domain.pdf.

[3] Alexander Blanchard and Laura Bruun, 'Bias in Military Artificial Intelligence', Stockholm International Peace Research Institute, December 2024, 9–10.

[4] For a description of states' motivation to use AI-DSS see Blanchard and Bruun, ibid 2; Grand-Clément (n 2); Nadibaidze, Bode and Zhang (n 2) Ch 3.

[5] Protocol Additional to the Geneva Conventions of 12 August 1949 (entered into force 7 December 1978) 1125 UNTS 3 (AP I); for claimed customary IHL equivalent see Jean-Marie Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law, Vol 1: Rules* (Cambridge University Press 2005) Ch 5 (ICRC Study).

[6] ICRC, 'Submission to the United Nations Secretary-General on Artificial Intelligence in the Military Domain', April 2025, 6, https://www.icrc.org/sites/default/files/2025-04/ICRC_Report_Submission_to_UNSG_on_AI_in_military_domain.pdf; ICRC and Geneva Academy of International Humanitarian Law and Human Rights (n 1) 16–17, 25; Michael N Schmitt, 'Israel – Hamas 2024 Symposium – The Gospel, Lavender, and the Law of Armed Conflict', 28 June 2024, https://lieber.westpoint.edu/gospel-lavender-law-armed-conflict; Marta Bo and Jessica Dorsey, 'Symposium on Military AI and the Law of Armed Conflict: The "Need" for Speed: The Cost of Unregulated AI Decision-Support Systems to Civilians', 4 April 2024, *OpinioJuris*, https://opiniojuris.org/2024/04/04/symposium-on-military-ai-and-the-law-of-armed-conflict-the-need-for-speed-the-cost-of-unregulated-ai-decision-support-systems-to-civilians; Gal Dahan and Tal Mimran, 'Artificial Intelligence in the Battlefield: A Perspective from Israel', 20 April 2024, *OpinioJuris*, https://opiniojuris.org/2024/04/20/artificial-intelligence-in-the-battlefield-a-perspective-from-israel.

[7] Bo and Dorsey (n 6); Jessica Dorsey, 'Israel's AI-Enabled Targeting of Hamas Members Jeopardizes Moral and Legal Standards of Warfare', Utrecht University', 18 July 2024, https://www.uu.nl/en/achtergrond/israels-ai-enabled-targeting-of-hamas-members-jeopardizes-moral-and-legal-standards-of-warfare; Christopher Elliot, 'Expedient or Reckless? Reconciling Opposing Accounts of the IDF's Use of AI in Gaza', *OpinioJuris*, 26 April 2024, https://opiniojuris.org/2024/04/26/expedient-or-reckless-reconciling-opposing-accounts-of-the-idfs-use-of-ai-in-gaza; see also ICRC, 'Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach' (2020) 102 *International Review of the Red Cross* 463, 471–72; Emelie Andersin, 'The Use of the "Lavender" in Gaza and the Law of

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

in a reaction to media reports, have also argued that one should take into account the broader targeting process in which AI-DSS are used, and which targets are verified, in the same way as they would be if no AI-DSS were used.[8] Undoubtedly, human operators play a crucial role in the application of precautions to verify, and many established precautions to verify can and must be applied when using AI-DSS. However, whether there are precautions to verify that are both AI-DSS-specific and outside the domain of the human operator is so far an underexplored question.[9]

This article therefore aims to identify possible precautions to verify that are specific to AI-DSS, with a focus on precautions that are not centred on human operators. However, as the risk to the civilian population that emerges from the use of AI-DSS is the result of a cascade of decisions about the application of possible precautions to verify, relevant aspects of the role of the human operator will be considered where necessary.

1.1. Examining the alleged use of 'Lavender' and 'Where's Daddy?' to illustrate the challenges and possibilities of taking precautions to verify

A general difficulty of discussing legal aspects of AI-DSS is limited access to reliable and detailed descriptions of their actual or intended use, which are crucial for a real-world legal analysis. An article in *+972 Magazine*, titled '"Lavender": The AI Machine Directing Israel's Bombing Spree in Gaza', published in April 2024, provided a rare detailed account of how such a system may have been used.[10] The factual basis of this report is debated.[11] Despite the uncertain accuracy of the report, the description of the use of AI-DSS appears to reflect how many envision these systems are being used, or will be used in the nearer future.[12] This article will therefore examine two of the AI-DSS described in the article by *+972 Magazine* – dubbed 'Lavender' and 'Where's Daddy?' – to illustrate possible precautions to verify that are specific to

Targeting: AI-Decision Support Systems and Facial Recognition Technology' (2025) *Journal of International Humanitarian Legal Studies* 1, 14.

[8] Israel Defense Forces (IDF), 'The IDF's Use of Data Technologies in Intelligence Processing', https://www.idf.il/210062; Dahan and Mimran (n 6).

[9] The most detailed discussion of precautions specific to AI-DSS to date was that provided by Andersin, but with a focus on AI-DSS that use facial recognition technology: Andersin (n 7) 28–33.

[10] Yuval Abraham, '"Lavender": The AI Machine Directing Israel's Bombing Spree in Gaza', *+972 Magazine*, 3 April 2024, https://www.972mag.com/lavender-ai-israeli-army-gaza.

[11] Israel Defense Forces (n 8); Dahan and Mimran (n 6); for a discussion of the differences between the *+972 Magazine* article and Israel's claims see Elliot (n 7); Elizabeth Dwoskin, 'Israel Built an "AI Factory" for War. It Unleashed It in Gaza', *The Washington Post*, 29 December 2024, https://www.washingtonpost.com/technology/2024/12/29/ai-israel-war-gaza-idf.

[12] Grand-Clément (n 2) 15–18; ICRC (n 6) 5; UN General Assembly (UNGA), 'Current Developments in Science and Technology and Their Potential Impact on International Security and Disarmament Efforts: Report of the Secretary-General' (23 July 2024), UN Doc A/79/224, 3, https://documents.un.org/doc/undoc/gen/n24/218/85/pdf/n2421885.pdf; ibid 14–15; ICRC and Geneva Academy of IHL and Human Rights (n 1) 13–14; Wen Zhou and Anna Rosalie Greipl, 'Artificial Intelligence in Military Decision-Making: Supporting Humans, Not Replacing Them', *Humanitarian Law & Policy Blog*, 29 August 2024, https://blogs.icrc.org/law-and-policy/2024/08/29/artificial-intelligence-in-military-decision-making-supporting-humans-not-replacing-them; Arthur Holland Michel, 'Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making', April 2024, 17, https://shop.icrc.org/decisions-decisions-decisions-computation-and-artificial-intelligence-in-military-decision-making-pdf-en.html.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

AI-DSS.[13] However, the article does not imply that these systems were actually used as reported.

1.1.1. Lavender

Lavender has been described as a system that analyses large data sets on the residents of Gaza with the aim of identifying Hamas or Palestinian Islamic Jihad (PIJ) operatives.[14] For that purpose, the system reportedly assigns most residents with a rating between 1 and 100, and any resident with a rating higher than a defined threshold would be proposed as a Hamas or PIJ operative (that is, a target).[15] Once a target had been proposed by Lavender, a human operator would 'conduct a single check: ensuring that the AI-selected target is male' and, in that case, put it on a target list.[16] If the proposed target were female, it would be assumed that the system had made a mistake as there were no known female operatives.[17]

1.1.2. Where's Daddy?

Once a target was put on a target list, the *+972 Magazine* report claims, a second software, named 'Where's Daddy?', would then track the target and alert the operators when it reached the target's home, after which an attack would be launched.[18] There are no further descriptions of how that system works, the sensors it uses, or the data it takes into account.

1.2. Structure

The article will identify possible precautions to verify that are specific to AI-DSS by first analysing the obligation to do everything feasible to verify that the target is a military objective from the perspective of the AI-DSS. To that end, it will break down the process of targeting, using AI-DSS, into a number of different stages at which precautions to verify can potentially be taken. This will include a discussion of the term 'everything feasible' in Article 57(2)(a)(iii) AP I. The remainder of the article will then identify the possible precautions at each stage of use and discuss the factors that could affect their feasibility.

2. The obligation to take precautionary measures to verify that the target is a military objective

Precautionary measures to verify that the target is a military objective are most prominently mandated in Article 57(2)(a)(i) AP I. The article reads as follows:

> 2. With respect to attacks, the following precautions shall be taken:
> (a) those who plan or decide upon an attack shall:
> (i) do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protection

[13] Abraham (n 10); Dwoskin (n 11).
[14] Abraham (n 10).
[15] ibid; see also Dwoskin (n 11).
[16] Abraham (n 10); Dwoskin (n 11).
[17] Abraham (n 10).
[18] ibid.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

> but are military objectives within the meaning of paragraph 2 of Article 52 and that it is not prohibited by the provisions of this Protocol to attack them;

This rule is likely also to be part of customary IHL, applicable to both international and non-international armed conflicts.[19]

2.1. The stages where precautions to verify can be taken

Considering that AP I in general, and its Article 57 in particular, aim to protect the civilian population from the effects of war, it would appear that the aim of precautionary measures to verify is to reduce the probability that a civilian or a civilian object is mistakenly made the object of attack and hence killed or destroyed.[20] In turn, this probability can be reduced at various stages of the process of using AI-DSS (see Figure 1).[21]

The first stage at which precautions to verify can be taken – by which the risk of civilians or civilian objects mistakenly being made the object of attack can be reduced – is the stage of deciding which persons or objects, or which classes of persons or objects, legally qualify as military objectives that can be lawfully attacked (legal qualification).

In principle, potential targets can be legally qualified individually, such as 'this particular person is a lawful target' or 'this bridge here is a civilian object', or collectively, such as 'we consider all persons who are active members of armed group X (other than those who are *hors de combat*) as lawful targets'.[22] An example of collective legal qualification is the practice of qualifying certain members of armed groups as 'declared hostile', after which they 'may be engaged with deadly force' '[a]s soon as they are positively identified'.[23] For the purpose of this article, persons or objects that were not individually but collectively qualified as military objectives are referred to as a class of target, such as the class of targets of Hamas and PIJ operatives. It is possible, if not likely, that AI-DSS are also used in the process of targeting persons or objects that were individually legally qualified. However, because collective legal qualification may pose additional legal challenges, this article will focus on processes where targets are collectively legally qualified.

[19] ICRC Study (n 5) 55.

[20] See also Claude Pilloud and Jean Pictet, 'Article 57 – Precautions in Attack' in Yves Sandoz, Christophe Swinarski and Bruno Zimmermann (eds), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (ICRC and Martinus Nijhoff 1987) 680.

[21] See also Nadibaidze, Bode and Zhang (n 2) 6.

[22] AP I (n 5) art 41.

[23] Gary D Solis, *The Law of Armed Conflict: International Humanitarian Law in War* (Cambridge University Press 2010) 507–508; US Department of Defense, *Law of War Manual*, June 2015, updated July 2023, para 5.7.1.1, https://media.defense.gov/2023/Jul/31/2003271432/-1/-1/0/DOD-LAW-OF-WAR-MANUAL-JUNE-2015-UPDATED-JULY%202023.PDF; United States Army, The Judge Advocate General's Legal Center and School, *Operational Law Handbook* (2024) 108; Camilla Guldahl Cooper, *NATO Rules of Engagement: On ROE, Self- Defence and the Use of Force during Armed Conflict* (Brill Nijhoff 2020) 123–24; for the question of whether IHL requires a case-by-case analysis see also Ian Henderson, *The Contemporary Law of Targeting: Military Objectives, Proportionality and Precautions in Attack under Additional Protocol I* (Martinus Nijhoff 2009) 83–87; Alan Cole and others, *Sanremo Rules of Engagement Handbook* (International Institute of Humanitarian Law 2009) 37–38.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

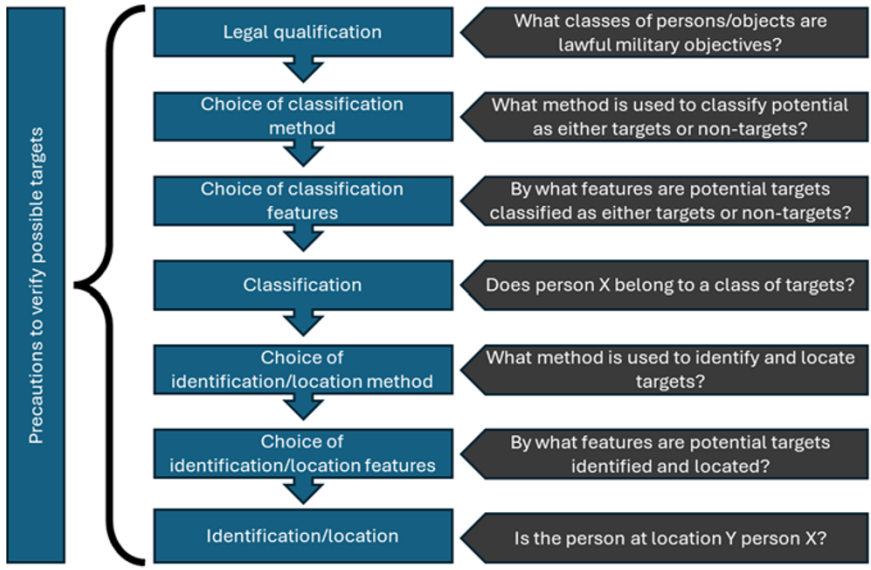

Figure 1. Stages of target verification with AI-DSS.

Once a class of persons or objects has been legally qualified as lawful targets, it is then necessary to classify all persons or objects that are considered as potential targets into either the class of targets or the class of non-targets.[24] For instance, assume, in line with the claims of the *+972 Magazine* article, that Hamas and PIJ operatives have been determined to be a class of targets. Each resident of Gaza, analysed by Lavender, would then need to be classified as either belonging to the class of targets of Hamas and PIJ operative, or to the class of non-targets, which consists particularly of civilians. While precautions to verify can be taken during the stage of the classification itself, two additional stages in which precautions can be taken precede this classification: first, the stage of selecting the method of classification; second, the stage of selecting the classification features.

Selecting the method of classification is the stage at which the method by which individual persons or objects will be classified as belonging to a class of either targets or non-targets is determined. Examples of such methods could be the classification by soldiers upon physically encountering a person or object, manual classification by soldiers based on data available about a person or object, or partially automated classification with the help of an AI-DSS. Each such method will have unique advantages and downsides.

Selection of the classification features is the stage when it is determined by which features a potential target is classified as belonging to a class of either targets or of

[24] See also JP 3-60 ('The detect phase is designed to acquire the targets selected in the decide phase'): United States Joint Chiefs of Staff, 'Joint Publication 3-60 – Joint Targeting', 31 January 2013, C–1, II–21, https://www.justsecurity.org/wp-content/uploads/2015/06/Joint_Chiefs-Joint_Targeting_20130131.pdf.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

non-targets.[25] Take the example of enemy combatants, where it is determined that they can be identified by the country-specific uniform they are expected to wear (the 'fixed distinctive sign recognizable at a distance').[26]

The classification itself is the stage at which a potential target is categorised as belonging to a class either of targets or non-targets by the chosen classification method and features, involving a human operator in some way. For instance, an AI-DSS proposes the classification of an encountered person into the class of targets, identified by the country-specific uniform, and the human operator then confirms or rejects this classification.

Once a person or object has been classified as a target and their whereabouts are unknown at the time of the classification – as would appear to be the case with AI-DSS, as in the reported use of Lavender – the target must be identified and located for the purpose of being attacked. For instance, person X has been classified as a target and needs to be located. In order to be able to say that a given person located at a given location is person X, that person needs to be identified as person X.

In principle, where AI-DSS are used to identify and locate a target, the same stages occur in the identification as described in the process of classifying potential targets: namely, the choice of the method of identification, the choice of identification features and the identification itself, including human operator involvement.

2.1.1. The stages in current targeting procedures

The above description of the stages at which precautions to verify can be taken is an analytical framework and is not intended to describe how states conduct targeting, or how they should conduct targeting with AI-DSS. However, the stages described can, in principle, be found in established targeting procedures such as the NATO Allied Joint Doctrine for Joint Targeting AJP-3.9 or the United States Joint Publication 3-60.[27] For instance, the stage at which targets or classes of targets are legally qualified corresponds with what is typically referred to as the target validation or, more broadly, the target development phase.[28] Similarly, provisions in the rules of engagements (ROE) may provide guidance as to the choice of the classification and/or identification method, and target selection standards and target characteristics may determine the choice of the classification and/or identification features.[29] Furthermore, the actual classification and/or identification would occur during the 'find' and 'fix' phases of the F2T2EA process (Find, Fix, Track, Target, Engage, Assess).[30] However, not currently reflected in the targeting procedures

[25] See also JP 3-60, ibid I–2 ('Every target has distinct intrinsic or acquired characteristics that form the basis for target detection, location, identification, and classification for ongoing and future surveillance, analysis, engagement, and assessment'); see also Cole and others (n 23) 38–39.

[26] Geneva Convention relative to the Protection of Civilian Persons in Time of War (entered into force 21 October 1950) 75 UNTS 287 (GC III), art 4.

[27] NATO Standard AJP-3.9, 'Allied Joint Doctrine for Joint Targeting', Edition B, v.1, November 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1033306/AJP-3.9_EDB_V1_E.pdf; US Joint Chiefs of Staff (n 24).

[28] US Joint Chiefs of Staff (n 24) Ch II–11, II–5; NATO Standard AJP-3.9 (n 27) Ch 1.3.32.

[29] Cole and others (n 23) 38–39; NATO Standard AJP-3.9 (n 27) LEX-18; US Joint Chiefs of Staff (n 24) Ch I–2; Commonwealth of Australia, Australian Defence Force, 'Operations Series ADDP 3.14 Targeting', 2 February 2009, Chs 1–4; see also NATO Standard AJP-3.9 (n 27) Ch 1.3.18.

[30] US Joint Chiefs of Staff (n 24) Ch II-21–26; NATO Standard AJP-3.9 (n 27) Chs 5–3.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

described is the distinction between the stages of classification and identification. This is probably because current targeting procedures typically apply to situations where the decision that a person or object is one that has been legally qualified as a target is taken upon physically encountering them. For these situations, it is only ever necessary to conduct either a classification or an identification, but not both. Where, for instance, it has been decided that all members of armed group X are lawful targets, it is sufficient to classify an encountered person as either a member or non-member of group X, whereas an identification (is this person X or Y?) is not necessary. Where, however, an individual person – say person X – is legally qualified as a lawful target, it is only necessary to identify an encountered person as either person X or someone else, and no classification is necessary. In other words, while established targeting procedures would appear, in principle, to cover both the classification and identification stages, only one stage would typically be relevant for a given target, whereas for AI-DSS both stages would be relevant.

2.1.2. Each stage as a source of risks to civilians or civilian objects

Each of the stages described here, from legal qualification to identification and location of the target, can be a source of risks for civilians to be killed or injured by an attack. During the legal qualification stage, a class of civilians or civilian objects may be qualified incorrectly as lawful targets and consequently attacked, as a result of a failure in the application of the relevant rules of IHL. In the choice of classification method, one may be selected that, because of its inherent limitations, classifies a civilian person or a civilian object as belonging to a class of targets. In the choice of identification features, some may be chosen that may be shown by persons or objects that do not belong to a class of targets. In the classification stage itself, civilians who do not show the classification features could incorrectly be classified as targets as a result of a malfunction of the target classifier, whether human or AI-DSS. Furthermore, in the stages at which targets are identified, persons or objects who have not been classified as targets could be identified as such, as a result of a failure in each of the respective stages of identification.

2.2. Does the obligation to do everything feasible to verify extend to all stages of the use of AI-DSS?

Because civilians and civilian objects can incorrectly be made the object of attack as a consequence of a failure in each of the stages outlined above, precautionary measures in each of these steps could prevent or minimise the number of civilians killed or civilian objects destroyed. However, does the legal obligation in Article 57(2)(a)(i) AP I[31] extend to all stages where precautions to verify are possible, and are state parties therefore required to do everything feasible to verify in each of these stages?[32]

The examples provided in the literature for the obligation to do everything feasible to verify are typically precautions for the qualification of objects or persons as

[31] AP I (n 5).

[32] See also Renato Wolf, 'The Legal Review of Autonomous Weapons Systems' (University of Queensland 2024) 117–19.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

military objectives. For instance, in its commentary on the Additional Protocols of 8 June 1977, the International Committee of the Red Cross (ICRC) – while using the word 'identification' – discusses issues that are clearly related to the legal qualification of objects as either military or civilian, such as the knowledge of a commander about a target and its nature.[33] Similarly, Henderson discusses the gathering of information to determine whether 'the bridge is a military objective', elaborating on the question of whether the bridge qualifies as a military objective, and not on whether it is indeed the bridge that has been selected as a target.[34]

Some authors, however, also list examples of misidentification, such as the attack on the Chinese Embassy in Belgrade during the Kosovo war. The Embassy was struck not because it was mistakenly believed that it qualified as a military objective but because it was incorrectly identified as a different building.[35] That both qualification and identification are covered by the obligation is also indicated by the International Criminal Tribunal for the former Yugoslavia (ICTY), which has summarised the obligation as follows:[36]

> A military commander must set up an effective intelligence gathering system to collect and evaluate information concerning potential targets. The commander must also direct his forces to use available technical means to properly identify targets during operations.

It is indeed difficult to see how IHL could protect the civilian population, and states could fulfil their obligation 'to distinguish at all times', if the obligation to take all feasible precautions to verify did not cover both the qualification and the identification of targets.[37] If the obligation covered only the qualification of targets and no precautions were mandated for their proper identification, then the careful qualification of targets could lawfully be nullified by sloppy identification, resulting in unnecessary attacks on civilians or civilian objects. Take, for instance, a scenario where for the legal qualification of a factory as either a military or a civilian object, additional evidence was gathered where necessary, and all the available evidence was carefully considered before the object was qualified as a military objective and therefore as a lawful target. However, in order to reduce risks for the pilots in executing the attack on the factory, the attack was ordered to be conducted from very high altitudes, making it difficult to correctly identify the factory and to distinguish it from similar-looking factories that were considered to be civilian. As a consequence, despite the efforts taken to legally qualify the factory correctly, civilian objects were attacked because they were falsely identified as the object in question.[38]

[33] Pilloud and Pictet (n 20) 680–81.

[34] Henderson (n 23) 234.

[35] Anthony PV Rogers, *Law on the Battlefield* (2nd edn, Juris 2004) 107; referring to the same example Henderson (n 23) 165.

[36] International Criminal Tribunal for the former Yugoslavia (ICTY), 'Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia', para 29, https://www.icty.org/x/file/Press/nato061300.pdf.

[37] AP I (n 5) art 48.

[38] For a practical example of this scenario see Rogers (n 35) 107; ICTY (n 36) para 64.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

Along the same line of argument, the apparent aim of Article 57(2)(a)(i) AP I to protect the civilian population could be achieved only if its obligations extend to all stages where precautions can reduce the probability that civilians or civilian objects are mistakenly made the object of an attack. Therefore, all feasible precautions to verify must be taken during all stages where the probability can be increased that the target to be attacked is a lawful military objective.

2.3. What precautions are feasible?

Many of the possible precautions to verify at the various stages described above come with military costs, so far as they may slow down operations, delay their launch, use more resources, increase the risk for own troops, decrease the probability that a target is successfully attacked, and so on.[39] For instance, where a human operator is given more time to review a proposed target, fewer targets can be processed by the same number of operators, or where standards of accuracy are increased to avoid misclassification of civilians as targets, more actual targets may slip through the net and remain undetected.[40]

Can states take these military costs into account when deciding whether to adopt a certain precaution to verify? Can they, in other words, argue that they are not taking a certain precaution because its military costs are too high, even though it may save the lives of civilians? That implementing a rule causes some sort of cost (and not just financial costs) is no particularity of Article 57(2)(a)(i) AP I. However, states cannot typically claim that the military costs of measures to implement a given rule are too high and that therefore they will not apply them. Take, for instance, Article 52(1) AP I, which prohibits intentional attacks against civilian objects. States could not argue that the (military) costs of this prohibition were too high and that, therefore, they could lawfully intentionally attack civilian objects. However, Article 57(2)(a)(i) AP I does not require that *everything possible*, regardless of military cost, is done to

[39] Michael N Schmitt, 'Precision Attack and International Humanitarian Law' (2005) 87 *International Review of the Red Cross* 445, 453–62; Program on Humanitarian Policy and Conflict Research at Harvard University, *Commentary to the HPCR Manual on International Law Applicable to Air and Missile Warfare* (Cambridge University Press 2013) 27–28. One could argue that the application of a legal rule may have costs is not a particularity of Art 57(2)(a)(i) AP I, but is a normal occurrence in the application of any legal rule. While this may be true, if indeed the feasibility proviso, as argued below, allows states to take into account the costs of applying a certain measure, then Art 57(2)(a)(i) AP I would be different from most rules of IHL or international law in general. Provided that the interpretation of the feasibility proviso presented below is correct, states can lawfully argue that they cannot take a given possible precaution because its military costs would render it unfeasible (always provided that an attack without the precautionary measure would suffice the minimum standards of distinction under Art 48 AP I and the rule of proportionality of Art 57(2)(a)(iii) AP I). For most rules of IHL, such an argument would not be valid. Take, for instance, Art 52(1) AP I, which prohibits intentional attacks against civilians or civilian objects. States cannot argue that the (military) costs of this prohibition were too high and that, therefore, they could lawfully intentionally attack civilians or civilian objects.

[40] See also Schmitt (n 6) ('[i]ndeed, an analyst could perform the same task manually, albeit in most cases without comparable speed and comprehensiveness'. See also Section 5.3 for a discussion of the relationship between the rates of false positives and false negatives.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

verify that targets to be attacked are military objectives.[41] Rather, states must *do everything feasible*.[42] The meaning of the term 'feasible' was discussed extensively during the drafting phase of AP I and is typically described as meaning 'practicable or practically possible' and taking into account humanitarian and military considerations.[43] Considering that IHL aims to protect the victims of armed conflict – and, in the context discussed here, civilians, in particular – then it would appear that the key humanitarian consideration is the number of civilians not killed or injured when taking precautionary measures, compared with not taking them (that is, the humanitarian benefits of a precautionary measure).[44] In turn, military considerations would appear to be the military costs of taking a given measure, such as a reduced likelihood that a given target can be struck, or that the resources spent cannot be used to achieve a military advantage, and so on (that is, the military costs of precautionary measures).[45] A measure is therefore feasible when the humanitarian benefits of taking the measure outweigh the military costs of not taking it.[46]

Given that states can take into account the military costs of possible precautions to verify, and as military costs therefore affect the precautions to verify that are legally required, the question of precautions to verify that states must take when using AI-DSS cannot be answered without taking into account the military costs that such precautions may incur.

3. Precautions to verify in the legal qualification of potential targets

The first stage at which everything feasible to verify must be done is that of legal qualification: deciding the classes of persons or objects that can be lawfully attacked. While there are some considerations about introducing AI tools for supporting legal advice in military operations, there are few concrete suggestions so far that states have intentions of using AI-DSS to support the legal qualification of targets in the

[41] Marco Sassòli and Yvette Issar, 'Challenges to International Humanitarian Law' in Andreas von Arnauld, Nele Matz-Lück and Kerstin Odendahl (eds), *100 Years of Peace Through Law: Past and Future* (Duncker & Humblot 2015) para 8.332.

[42] ICRC Study (n 5) 55; see also Pilloud and Pictet (n 20) para 2198; Henderson (n 23) 161–62; Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict* (4th edn, Cambridge University Press 2022) 189–90; see also Wolf (n 32) 119–22; US Department of Defense (n 23) para 5.2.3.2.

[43] ICRC, *Draft Rules for the Limitation of the Dangers Incurred by the Civilian Population in Time of War* (ICRC 1956) 10–11; ICRC, 'Official Records of the Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts', Vol I Pt III, 17; ICRC, 'Official Records of the Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts', Vol XV, 285; see also Pilloud and Pictet (n 20) para 2198; see also Program on Humanitarian Policy and Conflict Research at Harvard University (n 39) 26–27; Michael Bothe, Karl Josef Partsch and Waldemar A Solf, *New Rules for Victims of Armed Conflicts: Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949* (2nd edn, Martinus Nijhoff 2013) 404–405; Sassòli and Issar (n 41) para 8.332.

[44] See also Pilloud and Pictet (n 20) 680.

[45] See also Bothe, Partsch and Solf (n 43) 405.

[46] This is not to indicate that the weighing up between the military costs and the humanitarian benefits is governed by the same 'proportionality' yardstick that is imposed on the relationship between the expected collateral damage and the expected military advantage. The relationship between military costs and humanitarian benefits is not necessarily required to be 'balanced' or proportional. However, it is intrinsically an act that must take into account two quantities (military costs and humanitarian benefits) and decide which of the two prevails over the other.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

same way as they are using it for classification and identification.[47] In the reporting of the alleged use of Lavender in the Gaza War, for instance, there are no indications that Lavender or other AI-DSS have played any substantial role in legally qualifying potential targets (as opposed to classifying or identifying them).[48] In other words, it is likely that Lavender would not, colloquially speaking, have told its operators, 'I think Hamas or PIJ operatives are lawful targets according to the rules of IHL'. Instead, the decision to qualify Hamas or PIJ operatives as military objectives was likely to have been made without the involvement of Lavender or other AI-DSS, and in the same way that it would have been made if no AI-DSS were used to classify or identify targets.

For the legal qualification of a group of persons or objects as a class of targets, the legally necessary precautions to verify, therefore, in principle are not affected by the use of AI-DSS. All feasible precautions to verify must nevertheless be taken, as for any attack, whether involving AI-DSS or not.

As states do not appear to use or envisage the use of AI-DSS to support the stage of the legal qualification of targets, it will not be analysed in detail here. However, if states were to use AI-DSS to support the legal qualification of potential targets, they would be likely to be faced with similar questions and similar possible precautions to verify to those discussed in the context of classification below. By what method are the potential targets qualified? By what features? What is the acceptable rate of false positives (that is, the rate at which classes of persons or objects are qualified as targets even though they are civilians)?[49]

4. Precautions in selecting the method of classification

The second stage at which states have an obligation do everything feasible to verify is that of the choice of the method by which a potential target is classified as belonging to a class either of targets or non-targets. Each method of classification – for instance, using humans for the classification, or an AI-DSS in conjunction with humans – will typically have different inherent limitations as to the rate of false positives (that is, the rate at which persons or objects belonging to the class of non-targets are falsely classified as belonging to a class of targets). In turn, choosing one method of classification over another has the potential to increase or decrease the rate of false positives and therefore the number of civilians killed or injured, or civilian objects damaged or destroyed.

For instance, assume a state has two methods of classification available: a traditional, slower method using only humans, and a faster one using AI-DSS to support classifications. If one further assumes that the traditional classification method results in fewer false positives than the method that uses AI-DSS, then a possible precaution to decrease the rate of false positives would be the choice of the traditional method of classification over the AI-DSS method. This reveals the prototypical dilemma of precautions between humanitarian benefits and military costs. Choosing

[47] See Grand-Clément (n 2) 15–20; see also Trent Kubasiak, 'AI Proves Powerful Legal Ally', Association of the United States Army (AUSA), 2 January 2024, https://www.ausa.org/articles/ai-proves-powerful-legal-ally.

[48] Abraham (n 10).

[49] See also Andersin (n 7) 7.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

the traditional method over the AI-DSS method would have military costs, such as a lower number of potential targets classified by the same number of operators, or the same number of potential targets classified by more operators, balanced against fewer civilian casualties. The question of whether, in this example, one should choose the traditional over the AI-DSS method for classification is therefore a question of feasibility, as discussed above (Section 2.3). It should be noted, however, that the assumption that the AI-DSS method results in more false positives than the traditional method merely serves to illustrate the example. There is no intrinsic reason why AI-DSS must have a higher rate of false positives, and that, therefore, there would always be a humanitarian benefit in using a non-AI-DSS method.

5. Precautions in selecting classification features

The next stage that requires precautionary measures to verify is that of determining the features by which potential targets are classified into a class of either targets or non-targets (classification features).

In principle, potential targets can be classified by at least two distinct types of feature: their legally determinative features or some proxy feature(s). The first type of classification parameters is identical to the features that determine that the persons or objects of a class of targets legally qualify as military objectives (legally determinative features). For instance, assume it was determined that persons of a certain class of targets are military objectives because they are transporting munitions for an armed group. If an AI-DSS were able to detect when a person is engaged in such transportation for the armed group in question, and use this information (the feature) to classify this person as belonging to the class of targets, then it would use the same features that were used to legally qualify the members of the class of targets also to classify potential targets as members of that class. Put simply, those legally qualifying the class of targets would say: 'I am qualifying these people as lawful targets because they are transporting munitions'. Those classifying them (whether a person, or an AI-DSS) would say: 'I am classifying this person as a member of the class of targets because he or she is transporting munitions'.

The second type of classification features are proxy features, whereby a target is classified not by the legally determinative features but by some other feature(s), the occurrence of which correlates sufficiently with the occurrence of the legally determinative features. For instance, take the country-specific uniform of combatants mentioned earlier. Combatants are military objectives who can be lawfully attacked not because they are wearing the country-specific uniform but because they are members of the enemy armed forces.[50] However, the country-specific uniform can be used as a proxy feature to conclude that someone is a member of the enemy armed forces because the correlation between wearing the country-specific uniform and being a member of the armed forces is sufficiently high. In a similar way, according to media reports, Israel used features such as 'being in a WhatsApp group with a known militant, changing cell phone every few months, and changing addresses frequently' as features that may indicate membership of an armed group such as Hamas

[50] See also ICRC, *Commentary on the Third Geneva Convention: Convention (III) Relative to the Treatment of Prisoners of War* (Cambridge University Press 2021) paras 983–85.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

or the PIJ.[51] However, these features (likely to be used in conjunction with others) did not make the respective residents of Gaza military objectives; rather, they were used to conclude that the residents showing these features were sufficiently likely to be Hamas or PIJ operatives.

Both types of classification feature can be used with an AI-DSS. In cases where an AI-DSS is used to detect the legally determinative features, the critical parameter from the perspective of precautions to verify would be the rate of false positives of an AI-DSS in detecting the legally determinative features (that is, indicating that a potential target shows the legally determinative features when in reality it does not). Assuming that the rate of false positives is greater than zero – as it would almost certainly be in reality – then the question arises of whether it is acceptably low from an IHL perspective, and whether or not a measure that may decrease it (for instance, more computational time, more pixels on the target) would be available and feasible. However, because classification by proxy features adds additional legal challenges, the focus of the remainder of this section will be on systems that use proxy features to classify potential targets.

5.1. The use of proxy features for the classification of potential targets

IHL requires that parties distinguish at all times between civilians and civilian objects, on the one hand, and military objectives, on the other.[52] In order to bring this distinction into reality, those selecting proxy features as classification features must therefore ensure that the correlation between the proxy features and the legally determinative features is sufficiently strong. In an ideal world, this correlation would be perfect and one could be certain that, for instance, a person wearing a country-specific uniform is a member of the armed forces and therefore a target.[53] In reality, however, no such certainty exists and there is always the possibility that someone showing the proxy features does not belong to the class of targets. Because the correlation between proxy features and legally determinative features is never perfect, one will only ever be able to say that a given person or object is a target with a certain probability, never with absolute certainty.

However, in some conflicts, under some circumstances, the probability that a person or object showing the proxy feature – the country-specific uniform, in particular – is not a target (that is, a member of the enemy armed forces) is so minor that it can practically be ignored. This would particularly be the case in prototypical international armed conflicts where all parties have the ability and the will to enforce the proper display of the country-specific uniform by their combatants, and only by them.[54] However, in the reality of many modern conflicts, the country-specific

[51] Abraham (n 10); see also Brigadier General YS, *The Human-Machine Team: How to Create Synergy Between Human and Artificial Intelligence that Will Revolutionize Our World* (ebookPro Publishing 2021) 71.

[52] AP I (n 5) art 48.

[53] It is true that one would have to consider protected members of the armed forces, medical personnel and combatants *hors de combat* and establish negative features – that is, features that when shown exclude that the target description is fulfilled. However, for the sake of readability, these exceptions will not be discussed explicitly here.

[54] See also Cooper (n 23) 124; ICRC, *Commentary on the Third Geneva Convention* (n 50) paras 983–85; Toni Pfanner, 'Military Uniforms and the Law of War' (2004) 86 *International Review of the Red Cross* 94, 102–103.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

uniform as a single proxy feature that correlates strongly enough with the legally determinative features, does not exist.[55] Quite the contrary, by mixing with the civilian population, and by blurring the lines between protected civilians and civilian objects and persons and objects that can be lawfully attacked, many members of armed groups or civilians who participate directly in hostilities may actively seek to prevent the display of reliable proxy features by which they could be classified and/or identified.[56] Nonetheless, the absence of a single, sufficiently strong correlating proxy feature does not necessarily prevent classification by proxy features. Instead, a proxy feature that on its own would not correlate strongly enough with the legally determinate features may do so in combination with other weakly correlating proxy features.[57] Such weakly correlating proxy features that more strongly correlate in some combination will be referred to in the following discussion as 'composite proxy features'.[58]

By way of a simple example, assume that ten proxy features were identified, each of which correlates weakly with the legally determinative features of a given class of targets.[59] A person who shows only one of the ten features would only be slightly more likely to belong to the class of targets than a person who does not show that feature (that is, the proxy feature correlates only weakly with the legally determinative features). However, a person who shows all ten features would have a much higher probability of belonging to the class of targets in question than a person who does not show any of the features (because the composite proxy feature correlates highly with the legally determinative features). Assume, for instance, that a person who shows all the proxy features belongs to a given class of targets with a probability of 90%. In other words, on average, nine out of ten persons who show all ten proxy features would also show the legally determinative features. However, where only some of the proxy features of a given composite proxy feature are shown, the probability that a person belongs to a class of targets decreases. For example, where a person shows some combination of only nine proxy features, the probability may decrease to 80%; for a different combination it may only be 75%; for a combination of eight proxy features it may be 70%, and so on. Some of these proxy features may also negatively correlate with the legally determinative features – for instance, a person who shows nine proxy features but not a tenth negatively correlated feature will belong to a class of targets with a probability of 90%. However, if the person shows

[55] See also F Kalshoven, *Reflections on the Law of War: Collected Essays* (Martinus Nijhoff 2007) 453–54.

[56] See, eg, ICRC, 'How Does Law Protect in War? Principle of Distinction: Introductory Text', https://casebook.icrc.org/law/principle-distinction; Jeffrey Lovitky, 'Israel – Hamas 2023 Symposium – Distinction and Humanitarian Aid in the Gaza Conflict', Lieber Institute West Point, Articles of War, 13 November 2023, https://lieber.westpoint.edu/distinction-humanitarian-aid-gaza-conflict.

[57] See, eg, Christian Heumann, Michael Schomaker and Shalabh, *Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R* (2nd edn, Springer International 2022) 267–68, 280–81; Larry D Schroeder, David L Sjoquist and Paula E Stephan, *Understanding Regression Analysis: An Introductory Guide* (Sage 2017) 21–22, 25–26, https://methods.sagepub.com/book/understanding-regression-analysis-2e.

[58] For an example of the use of the term 'composite feature' see Fergus Imrie and others, 'Composite Feature Selection Using Deep Ensembles', 36th Conference on Neural Information Processing Systems, December 2022, 1–2, https://proceedings.neurips.cc/paper_files/paper/2022/file/eab69250e98b1f9fc54e473cc7a69439-Paper-Conference.pdf.

[59] For a similar example see Elliot (n 7).

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

the tenth proxy feature (for example, being female) then the probability decreases drastically to 0.1%.

In reality, composite proxy features would be likely to take into account more proxy features, and would be aggregated in a far more complex manner than indicated here. Furthermore, it is likely that not one but a number of different composite proxy features will be used. However, the example shows the challenges that the use of composite proxy features may pose from the perspective of distinction under IHL: what is the probability that a person or object that shows some or all elements of the composite proxy feature belongs to the class of targets, and what probability is enough to conclude that a potential target does belong to the class of targets?[60]

The reported use of Lavender in the Gaza war exemplifies this. Hamas and PIJ operatives would not distinguish themselves visually from the civilian population by a uniform or some other 'fixed distinctive sign recognizable at a distance'.[61] Hence, a single proxy feature by which a resident of Gaza could be classified as either a civilian or a Hamas/PIJ operative was likely not to have been available. According to the *+972 Magazine* report:[62]

> [Lavender] analyzes information collected on most of the 2.33 million residents of the Gaza Strip … [and] then assesses and ranks the likelihood that each particular person is active in the military wing of Hamas or PIJ. According to sources, the machine gives almost every single person in Gaza a rating from 1 to 100, expressing how likely it is that they are a militant.
> Lavender learns to identify characteristics of known Hamas and PIJ operatives, whose information was fed to the machine as training data, and then to locate these same characteristics – also called 'features' – among the general population, the sources explained. An individual found to have several different incriminating features will reach a high rating, and thus automatically becomes a potential target for assassination.

The identification of characteristics of known Hamas or PIJ operatives is, in the terminology established for the purpose of this article, the selection of proxy features for the class of targets 'Hamas or PIJ operatives'. In turn, the composite proxy feature is the combination of the selected proxy features, aggregated in some unknown way, and the rating between 1 to 100, which Lavender assigned to the residents of Gaza, expresses the probability for each resident to be a Hamas or PIJ operative.[63]

5.2. Precautions in the choice of the composite proxy feature

Classifying potential targets with composite proxy features in practice will result in a certain rate of false positives greater than zero which, in turn, has the potential – if not detected by the human operator – to result in civilian deaths or injuries

[60] See also Andersin (n 7) 7.
[61] GC III (n 26) art 4(2)(b); see also Lovitky (n 56).
[62] Abraham (n 10).
[63] ibid.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

and/or the destruction of civilian objects. A measure that can decrease the rate of false positives could hence be a possible precautionary measure to verify.

The first possible precautionary measure to verify that can be taken in this context is to establish the rate of false positives that results from the use of a given set of composite proxy features. With that, the basis for other important determinations can be made. Based on the established rate of false positives, it can be decided whether the use of a method of warfare that classifies potential targets with the given composite proxy features, together with human operators with their rate of detecting false positives, can sufficiently distinguish between lawful targets and civilian objects according to the minimum standards of IHL. Where the use of the given composite proxy features results in an extremely high rate of false positives (and is not corrected by the human operator), the use of such a method may amount to an indiscriminate attack.[64] Where the composite proxy feature results in a rate of false positives that is not inherently unlawful but can potentially be improved with some measures, it becomes a question of whether taking those measures is feasible.

With the rate of false positives established, the humanitarian benefits of possible further precautionary measures to decrease the rate of false positives can then be quantified and their feasibility determined. For instance, assume for a given AI-DSS that a possible precaution was a composite proxy feature that is more complex to calculate but would have a lower rate of false positives. Assume that this precaution could decrease the rate of false positives by half (humanitarian benefit) but would also delay the use of the method for target classification and require more computational time (military costs). For composite proxy features with a high rate of false positives – say, two out of ten potential targets classified are false positives – the humanitarian benefits of halving the rate of false positives would be high, whereas for a system with an already lower rate of false positives – say, one in 100 – it would be considerably lower. Consequently, in the former example the humanitarian costs may outweigh the military costs; in the latter – depending on the military costs – they may not.

5.3. Precautions in the choice of the sufficient probability

A target may show all the proxy features of the respective composite proxy feature, in which case the correlation between the composite proxy feature and the legally determinative features is the highest, and the rate of false positives the lowest. However, a target may show just some of the underlying proxy features, in which case the correlation between the composite proxy feature and the legally determinative features is lower, and the rate of false positives higher. This is likely to be what the rating of 1 to 100 represented in the reported use of Lavender. A rating of 100 may have indicated that a person showed all underlying proxy features, and thus the probability was the highest that the person showed the legally determinative features, whereas a rating of 1 may have indicated that the person showed none of them, and consequently the probability that the person showed the legally determinative features was the lowest. From a humanitarian perspective, the preference is clear: a lower rate of false positives will decrease risks for the civilian population;

[64] AP I (n 5) art 51(4).

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

therefore, the highest possible congruency between the target and the composite proxy feature is desirable.

In an ideal world, all targets would therefore show the composite proxy feature fully; hence, one would classify a person or object as belonging to a class of targets only when it showed the composite proxy feature fully. However, in reality, targets are unlikely to be uniform. To take the example of the reported use of Lavender, some Hamas or PIJ members will not have changed their mobile phones frequently, some will not have changed their address frequently, and others may not have been in a WhatsApp group with known militants. Given non-uniform targets, classifying only persons or objects as targets that show the composite proxy feature fully will result in a very high rate of false negatives (that is, actual targets not classified as belonging to the class of targets).[65] By way of illustration, assume that where a person or object shows the composite proxy feature fully, it belongs to the class of targets with a 99% probability; however, only 50% of the targets show the composite proxy feature fully. In other words, the rate of false positives can be expected to be 1 in 100; however, the rate of false negatives would be 1 in 2 (that is, every second actual target would not be classified as such because it does not show the composite proxy feature fully). Now assume that it was determined that because of the high rate of false negatives, it would be sufficient to classify a person as a target if they showed at least 90% of the underlying proxy features of the composite proxy feature. Because there are more targets showing 90% of the underlying proxy features than targets that show 100% of them, the rate of false negatives decreases, and more of the actual targets are classified as such. In turn, because more civilians or civilian objects show 90% of the underlying proxy features, the rate of false positives increases, and more civilians or civilian objects are falsely classified as belonging to a class of targets. This is the dilemma: reducing the rate of false positives is desirable from a humanitarian perspective but, all else equal, results in an increased rate of false negatives, which is militarily undesirable.[66] These are the factors that a commander must quantify and balance against each other to determine whether an increase in the required congruency with the composite proxy feature is feasible.

5.4. The precautionary measure of adding or introducing human applicable classification features

The purpose of the human operator from the perspective of precautions to verify is the reduction in the rate of false positives resulting from the use of the AI-DSS. In other words, the human operator must identify persons or objects classified by the AI-DSS as belonging to the class of targets that are in fact civilians or civilian objects, and reject their classification as targets. The more false positives the human operator detects, the fewer civilians or civilian objects will be falsely classified as targets. The rate at which the human operator detects false positives resulting from the AI-DSS will be referred to as the 'detection rate'. Where, therefore, the role of the human

[65] See also Andersin (n 7) 10.

[66] See also Leon Gordis, *Epidemiology* (5th edn, Elsevier 2014) 114–19; Charlotte Baker, *Epidemiology* (Open Education Initiative, University Libraries at Virginia Tech 2023) 102–107; Zhi-Hua Zhou, *Machine Learning* (Springer Singapore 2021) Ch 2.3.2, https://link.springer.com/10.1007/978-981-15-1967-3.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

operator is to detect false positives, the role of precautions in the classification is to increase the detection rate.

There are at least two distinct ways in which the human operator can detect false positives. First, the operator can assess the same data available to the AI-DSS without additional classification features. For instance, the operator could assess the reliability of the specific data used about the person or object in question, or trace how the AI-DSS came to the determination that the person or object belongs to the class of targets. Given the likely complexity and the number of proxy features involved and their underlying data, it is unlikely that a human would be able to recreate the entire process that led to the classification of a certain person or object as a target, in particular within the time constraints typical for military operations. More realistic would be the application of what could be called 'human judgement' or a 'human plausibility test'. Rather than applying additional classification features or re-creating the classification process, in a human plausibility test the operator would look at the proposed target and possibly some of the underlying data, and query 'Does this look right?'. This kind of verification will be discussed below (Section 6.2.1).

In the second way of detecting false positives, the human operator applies additional classification features that the AI-DSS has not yet used to classify the possible target (human applicable classification features). An example of this is the reported use of Lavender in which a human operator verified that the proposed target is male.

To the degree that human applicable classification parameters are available and able to decrease the rate of false positives, they are possible precautionary measures to verify that the target is a military objective. Whether their use is feasible and therefore legally mandated is, however, a different question. The humanitarian benefit of additional human applicable classification parameters would depend, in particular, on the rate of false positives that result from the AI-DSS itself, and after the existing human applicable classification parameters (if any) are applied. In particular, the higher the rate of false positives, the higher the humanitarian benefit of having an appropriate additional classification feature. On the other hand, in terms of military costs, having a human operator applying additional classification features will inevitably increase the time needed to verify the proposed target and introduce a certain rate of false negatives (that is, the rate at which proposed targets are rejected even though they are actual targets). As a consequence, either fewer targets can be verified by the same number of persons, or the same number of target verifications would require more human operators, and fewer actual targets can be classified as such.

6. The classification of the targets

At this stage of the process, potential targets are classified using the classification method and classification features. This stage can be subdivided into two consecutive stages. First, the AI-DSS applies the classification parameters to its data set and classifies potential targets. Second, a human operator applies the human applicable classification features (if any) and conducts a human plausibility test (if mandated), and consequently either approves or rejects the proposed targets.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

6.1. Classification by AI-DSS

Where a person or an object is classified by the AI-DSS as belonging to a class of targets, even though it did not show the composite proxy feature (or not to the degree determined by the operator), then a false positive occurs as a result of malfunctioning of the software and/or hardware that conducted the classification (but not of the classification features).[67] To the degree that such malfunctions, if they occur, are an intrinsic characteristic of the classification method chosen (that is, they cannot be prevented), they should be considered in the choice of the classification method (see Section 4). To the degree that they can be reduced by modifications to the software and/or hardware conducting the classification (but not the classification features), these are precautionary measures specific to the classification step by the AI-DSS. If such precautionary measures cause military costs (for example, delay in the employment of the AI-DSS in question), then the question of whether they must be taken by the operator of the AI-DSS is one of feasibility.

6.2. Confirmation by the human operator

After an AI-DSS has classified a given person or object as belonging to a class of targets, this classification will be either confirmed or rejected by the human operator. Two possible precautions to verify have been described above: that of the human plausibility test and that of additional human applicable classification features, both of which can potentially be combined.

6.2.1. The human plausibility test

The human plausibility test has been described above as a test in which a human operator would apply what could be described as human judgement, or even 'gut feeling'.[68] While it is difficult to describe how such a test could be conducted, it can be assumed that to have any prospect of detecting a false positive, a human operator would need to access and assess at least some of the data previously used by the AI-DSS to classify the person or object as a target. Whether such a human plausibility test can be effective in detecting false positives would be likely to depend on factors such as the time available to conduct the test, the selection and training of the human operators, their work hours and the understandability of the AI-DSS in question.[69]

6.2.2. Additional human applicable classification features

Human applicable additional classification features are described above as features that the AI-DSS has not yet used to classify the possible target but are applied by a human operator.[70] The effectiveness of human applicable classification features depends, on the one hand, on the features themselves and whether the addition of

[67] False positives occurring as a result of the classification features are covered in Section 5.

[68] Section 5.4.

[69] See also Ruben Stewart and Georgia Hinds, 'Algorithms of War: The Use of Artificial Intelligence in Decision Making in Armed Conflict', *Humanitarian Law & Policy Blog*, 24 October 2023, https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict; Andersin (n 7) 31.

[70] Section 5.4.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

these features can effectively reduce the rate of false positives. On the other hand, it depends on the rate of false positives of the human operator in determining whether the human applicable classification feature is shown by the target. For instance, assume, in line with the reports about Lavender, that a human operator must verify whether the proposed target is male (in which case the human applicable classification feature would be 'male'). How likely would a human operator be to fail to detect that a proposed target was not male and, as a consequence, incorrectly classify the person as a target? The rate of failure to correctly determine the human applicable classification feature would depend on several factors, including the difficulty of determining the feature in question, as well as human or personal factors such as work hours, fatigue, skill and experience.

6.2.3. Precautions to address bias

Bias has been cited frequently as a central issue with the use of AI-DSS.[71] Two kinds of bias are particularly relevant. The first is bias caused by the data and/or the algorithm itself, where the AI-DSS replicates bias that – inadvertently or not – was contained in the data and/or programmed into the algorithm.[72] For instance, consider software that has been trained (that is, it established proxy features or composite proxy features) with a data set in which the targets belong predominantly to ethnic group A and the civilian population to a different ethnic group, B.[73] With such training data, the AI-DSS may identify the ethnicity A (their typical appearance, style of dress, and so on) as a proxy feature for the classification of the targets. If such an AI-DSS were then used in contexts where the civilian population also belongs to the ethnic group A, it may be more likely to falsely classify civilians of the ethnic group A as targets.

The second kind of bias relates to the phenomenon where humans place too much confidence in the proposals of the AI-DSS and hence fail to recognise failures of the machine (automation bias).[74] The human operator who confirms or rejects targets proposed by the AI-DSS can therefore be both a measure to remedy bias (by correcting proposals by the AI-DSS that are biased) and the source of bias (particularly because of automation bias).

Bias caused by the underlying training data and/or algorithm can potentially be addressed in the choice of classification method and/or the classification features (that is, by choosing classification methods and features that result in less bias). With regard to the role of the human operator, possible steps to address bias include the selection of appropriate operators, the training process (in particular, with a

[71] ICRC and Geneva Academy of International Humanitarian Law and Human Rights (n 1) 16; ICRC (n 6) 5.

[72] Blanchard and Bruun (n 3) 16–17; Ingvild Bode and Ishmael Bhila, 'The Problem of Algorithmic Bias in AI-Based Military Decision Support Systems', *Humanitarian Law & Policy*, 2 September 2024, https://blogs.icrc.org/law-and-policy/2024/09/03/the-problem-of-algorithmic-bias-in-ai-based-military-decision-support-systems/#:~:text=Both%20specific%20examples%20of%20bias,recognized%20as%20a%20particular%20problem.

[73] For a similar example see Andersin (n 7) 29.

[74] ICRC and Geneva Academy of IHL and Human Rights (n 1) 27; Bo and Dorsey (n 6); Andersin (n 7) 13; Bode and Bhila (n 72).

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

focus on identification of targets proposed as a result of bias) and work conditions (including time to review proposed targets, working hours, and so on).[75]

For both kinds of bias, the possible precautionary measures, to the degree that they have military costs, can be considered under the feasibility proviso.

7. Precautions in identifying and locating the target

Once a person or object has been classified as belonging to a class of targets, the location of this target must be determined. In many combat situations this may be done simultaneously, using the same methods as target classification; for instance, a person can be visually classified as an enemy combatant by wearing a country-specific uniform. At the same time, the person's location is visually determined. However, where targets are classified without knowing their location, and are subsequently located prior to the attack – as was reportedly the case with Lavender and 'Where's Daddy?' – then classification and identifying the locality of the target become two distinct steps.[76]

However, locating a person or object previously classified as a target must intrinsically encompass a step to identify that target. After all, it would logically be impossible to state that person X is located at position Y without determining that the person at location Y is, in fact, person X. Intrinsically, therefore, targets who have been classified without knowing their location must have their identity determined in conjunction with their location.

It is likely that the methods used to identify and locate a target are different from those used to classify them. Furthermore, it is also possible that different methods are used for the identification and for the location of a target. For example, a state may classify targets with a Lavender-style AI-DSS by using a digital footprint, identify them with surveillance cameras, and locate them with mobile phone triangulation. However, for the purpose of this article, identification and location using an AI-DSS will be considered together.

The reported use of the AI-DSS 'Where's Daddy?' provides a possible example of a system that was used to determine the target's location. According to the *+972 Magazine* article, 'Where's Daddy?' alerted its operators when a target reached home, where it would be attacked.[77] Whether this software was used solely to locate the target, or whether it also supported the identification of targets, is unknown. The alleged location of the target at their home and the subsequent attack there may require particular examination, as the choice of the location of the attack would appear to maximise the collateral damage caused, rather than to decrease it. This is, however, a question that is governed primarily by the obligation to take all feasible precautions to minimise collateral damage according to Article 57(2)(a)(ii) AP I and its customary IHL equivalent. From the perspective of precautions to verify, however, the question is primarily about the accuracy (that is, the rate of false positives) of software like 'Where's Daddy?' in locating and identifying a potential target. In

[75] See also ICRC and Geneva Academy of IHL and Human Rights (n 1) 19–20; Emelia Probasco and others, 'AI for Military Decision-Making: Harnessing the Advantages and Avoiding the Risks', Center for Security and Emerging Technology, April 2025, 21.

[76] See also US Joint Chiefs of Staff (n 24) II-25-II26.

[77] Abraham (n 10).

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

other words, how likely is the software to alert its operators that a particular target reached their home when, in fact, it did not?

The identification/location of a person or object as one that has been classified as a target follows a very similar pattern to the classification of targets, and allows for similar precautions to verify. As with the choice of the classification method, the choice of the identification/location method involving an AI-DSS may have some inherent limitations and could result in a higher rate of false positives (here, a person identified as belonging to a class of targets when the person actually belonged to the class of non-targets) compared to a method without AI-DSS. Similar to the classification features, the choice of the identification/location features affects the rate of false positives. In particular, as with the classification parameters, it is possible to use some features that are directly linked to the target in question – such as a target person's facial features – or some proxy features (or composite proxy features) that correlate with the identity and/or the location of the target in question (such as a particular mobile phone).[78] In the former case, the rate of false positives depends, in particular, on the degree of congruency that a detected face has with the digital copies with which it is compared.[79] Determining a higher degree of congruency would decrease the rate of false positives, but increase the rate of false negatives. The same is true for cases where a proxy feature is used to identify and locate a target. In addition, the rate of false positives is also affected by the choice of the proxy features and/or the composite proxy feature. For the identification stage itself, the role of the human operator, as with the classification, is to detect false positives; however, it is also a source of false negatives. In turn, measures that decrease the rate of false positives also have the potential to increase the human operator's rate of false negatives.

8. Conclusions

This article identifies AI-DSS specific precautions that can – and under some circumstances must – be taken to verify that targets to be attacked are lawful military objectives. While the human operator remains central, for both legal and ethical reasons, there are numerous possibilities to reduce the risk of harm to civilians or civilian objects that are independent of the human operator. These precautions can be taken at various stages in the process of using an AI-DSS – specifically legal qualification, classification, identification and location – all of which are within the scope of the obligation to do everything feasible to verify. Of particular importance is the choice of classification and/or identification/location features, especially the accuracy with which a composite proxy feature can classify or identify/locate a person or object as a target. However, possible precautions to verify targets typically incur military costs, and states may take these into account in deciding the feasibility of precautions to verify.

Precautions to verify therefore play a vital role in reducing the risks that arise from the use of AI-DSS in armed conflicts. However, they are not a silver bullet.

[78] The *+972 Magazine* article makes some indications that mobiles phones were used to locate persons, however, without providing any details: Abraham (n 10).

[79] Andersin (n 7) 10–13.

https://doi.org/10.1017/S0021223725100071 Published online by Cambridge University Press

While they have the potential to minimise the harm done to civilians, they cannot entirely prevent it. Furthermore, precautions to verify cannot overcome, and potentially not even mitigate, some of the ethical concerns and/or broader concerns about the normalisation of armed violence that may be associated with such systems.

Acknowledgements. The author is grateful to Dr Natalia Jevglevskaja, University of New South Wales, and Associate Professor Ted Pavlic, Arizona State University, for their comments and suggestions.

Funding statement. Not applicable.

Competing interests. The author declares none.

Cite this article: Renato Wolf, 'Applying Precautions in Target Verification with AI Decision Support Systems' (2025) *Israel Law Review* 1–24, https://doi.org/10.1017/S0021223725100071