International Institute of Humanitarian Law
Institut International de Droit Humanitaire
Istituto Internazionale di Diritto Umanitario

42nd ROUND TABLE ON CURRENT ISSUES OF INTERNATIONAL HUMANITARIAN LAW ON THE 70th ANNIVERSARY OF THE GENEVA CONVENTIONS

"Whither the human in armed conflict? IHL implications of new technology in warfare"

Sanremo, 4-6 September 2019

The SKYNET programme and the principle of distinction: why we should not let artificial intelligence lead the way

Andrea FARRÉS JIMÉNEZ
Humanitarian Policy Intern, Norwegian Refugee Council

Three years ago, a big data analytics for pattern recognition in intelligence data came to light through the Snowden leaked documentation. It was the SKYNET Programme, a machine-learning algorithm developed by the United States.

The aim of this programme was to analyse the cellular network metadata of millions of people in Pakistan, to identify couriers carrying messages between Al-Qaeda members, rating their likelihood of being terrorists. Whether the SKYNET Programme was actually used, and the exact characteristics of it, is unknown by the general public.

However, when analysing the leaked documentation explaining how the algorithm functioned, we came across an interesting case study to take as an example to analyse some of the challenges artificial intelligence (AI) can face to ensure compliance with the principle of distinction.

Therefore, the aim of my presentation is to take the case study of the SKYNET programme as the starting point, through which I would like to raise several general challenges of using AI as a decision aid system in relation to the compliance of the principle of distinction in the targeting process.

Starting with how the training of the algorithm was made, I am going to raise general concerns on some difficulties AI faces to ensure the protection of civilians. Also, I am going to tackle how the variants inserted can jeopardize cultural sensitivity and perpetuate biases.

I am also going to raise the practical problem of the scarcity of available and reliable data, and the problems which can result from that. Finally, I am going to discuss the psychological impact operators experience in human-machine partnerships, and how this can potentially impact in terms of IHL compliance.

Before starting the discussion on the SKYNET programme, I would like to talk about some background considerations relating to AI. AI is a branch of computer science that deals with the simulation of intelligent behaviour in computers. Especially during the last years, the development of AI has increased a lot. Currently, it impacts almost every aspect of our lives, for instance when we use our smartphones or when we get Amazon recommending to us what to buy or Netflix suggesting what to watch.

As an essential part of AI, algorithms are mathematical instructions which tell computers what to do. One widely used type of algorithms, also used for the SKYNET programme, are the machine-learning algorithms. Simplifying a lot, machine learning algorithms are systems which process massive amounts of data to identify autonomous rules or patterns. Through a learning process, they can end up drawing general conclusions from single pieces of information.

This learning process consists of 4 steps: training, testing, application and validation. From now until the end of the presentation, I will explain the concerns, both legal and practical, that arise throughout the machine-learning process the SKYNET programmers could have faced.

The first step to develop the SKYNET Programme was to train the algorithm. The training phase consists of feeding the computer with huge amounts of data labelled as a "ground truth". This "ground truth" is based on pattern recognition, using mathematical methods to find relationships in the sensory data.

At the outset, it is worrisome that as a "ground truth" of the training algorithm the obligation of refraining from targeting civilians receives no consideration, or at least this is what can be inferred from the leaked documentation. Ensuring that the civilian persons and objects are protected is crucial, as the principle of distinction requires parties to an armed conflict to distinguish between civilian persons and civilian objects, and combatants and military objectives, as only the latter can be targeted.

However, even if programmers wanted to make sure that the algorithm designed would comply with the principle of distinction, it seems that it is extremely hard to reach this result. This is because Article 50 of the Additional Protocol I to the Geneva Conventions describes the category of civilian population in a negative sense. This formulation entails an AI-related challenge which I consider difficult to overcome. Since a definition of the concept of the category of "civilians" is absent, it seems that it becomes very complex to translate this notion into a computer code.

I would argue that feeding the machine with a definition of what is a combatant, and that all what does not fit into that definition should not be targeted, may not be enough basis to ensure civilian protection. I believe that this is so, because this reasoning forgets a third category of people: civilians directly participating in the hostilities.

Inserting in an abstract way what is a civilian directly participating in the hostilities can entail various challenges. Those difficulties are due to the fact that, to assess if a civilian is DPH, the threshold of harm, direct causation and the belligerent nexus needs to be proved.

And to analyze these 3 characteristics, I think it is necessary to assess contextual information, "the big picture", a task only humans can undertake. For instance, commanders should assess "the tactical and strategic implications of a potential harm; the status of other potentially threatened individuals; the direct causal implications of someone's actions; or the sociocultural and psychological situation in which that individual's intentions and actions qualify as military actions".[1]

Therefore, contextual information, "the big picture", essential to properly identify civilians DPH, becomes a task only humans can undertake, and which the algorithm most likely would leave out of its assessment.

Finally, IHL states that the presence of military or civilians DPH among the civilian population does not deprive the population of the protection from an attack. I would argue that this provision favours the need for military commanders to issue context-based decisions, which reliance on AI computer-aiding would likely omit as well.

Moving on to the "ground truths" which were inserted in the training of the SKYNET algorithm, more issues of concern arise. For instance, the more than 80 properties entered as relevant data in the SKYNET Programme to help rating people as couriers assumed that their behaviour differs from the

[1] Peter Asaro, "On Banning Autonomous Lethal Systems: Human Rights, Automation and the Dehumanizing of Lethal Decision-making," Special Issue on New Technologies and Warfare, *International Review of the Red Cross* 94, no. 886 (Summer 2012); 789.

rest of the population and those variants included factors as turning off the phone or swapping SIM cards, understood as attempts to evade mass surveillance.

The assumption that couriers portray a distinct behaviour in relation to the use of their phones can be by itself problematic. This shows that the process of data interpretation is not neutral, as the biases of the programmers can be reflected in how and which information is introduced to the machines.

To gain a more accurate insight on which variants are actually relevant or not, situational awareness and cultural sensitivity is needed by those operators. A good example of this need is in conflicts like Yemen or Afghanistan, where civilians carry weapons for self-protection. In these cases, programmers in the US should be aware that carrying guns or not is not a valid criterion to distinguish civilians from combatants, contrary to what someone working in a robotics laboratory in the US may assume at first.

For example, there have been cases of signature drone strikes where civilians were targeted after feeling forced to provide shelter and food to militants in their homes. In cases like this, it has been debated that partially due to such loss of situational awareness, the duress the civilians experienced was not considered.

In conclusion, I would argue that cultural sensibility is an important component for an algorithm to be accurate, so AI programmers should receive specific training on that.

Another challenge I would like to point out is at the scarcity of available and reliable data. Machine learning algorithms need massive amounts of data to infer accurate patterns. This means that feeding the machine with information of dozens of known couriers is not enough.

This is so because the least information available, the more the machine can produce false positives and false negatives. False positives occur when someone is mistakenly identified as a terrorist or combatant, and false negatives refer to the contrary. Besides, the reliability on the information these databases provide depends on their accuracy, which is challenged as sometimes those are not updated or contain mistakes.

As the last point concerning the training of data, I would also like to highlight that in the leaked information the data is not trained to identify *hors de combat*. This could lead to IHL violations if the eventual targeting decision does not consider this option, because assuming that somebody is a combatant without the possibility of contemplating their surrender is an IHL violation.

Once the training phase is done, the testing period starts. This second step consists of inserting another data set to check whether the machine can properly generalize. If it can do so, then it is put into operation, for the third step, which is the application phase. During this stage, the machine analyses a wide range of information to find the patterns that match with the training set. After this procedure, the validation phase starts, in which programmers assess the machine's performance.

The information available of how the testing, validation and application phases were made in the SKYNET programme, spotlight some general concerns.

First, the question of the scarce data available. As mentioned, for machine learning algorithms to work, massive amounts of data need to be inserted. If all the relevant information regarding known couriers' database is inserted in the algorithm for the training phase, it means that there is no other separate and different dataset available to corroborate how accurately the algorithm works. In this sense, the training phase would get totally invalidated.

In relation to the validation phase, some worrisome concerns arise, as we can observe the psychological impacts operators are likely to experience when working through a human-machine partnership. As the leaked information reveals, what appears to be shown as a proof of accuracy is that the person who got the highest likelihood of being a courier is, in fact, a well-renowned Al Jazeera journalist, meaning, a false positive.

A way to explain these low validation standards is the demonstrated psychological impact technology has on human operators. Indeed, partnerships human-human and human-machine do not work in the same manner.

Human performance is affected by automated systems, having an impact on the "loss of situational awareness, complacency, skill degradation, and decision biases."[2] Operators, and humans in general, have a tendency to over rely on the outcome the computer produces (who has not reluctantly followed a suggested google maps route, even if in fact he or she thought that there were alternative faster ways to reach the desired destination?). Indeed, humans tend to delegate too much to automation. Eventually, this means that the decision-makers are less attentive and healthy scepticism over what the decision aid suggest is erased.

This can have negative implications on some IHL obligations. For instance, it can lower the presumption of the civilian status in case of doubt.

[2] Mary L. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," *American Institute of Aeronautics and Astronautics* (2004): 2.

Humans tend to over rely in the outcomes of machines, downplaying hesitations and suppressing doubt over what computers suggests. The practical implication of this reaction is jeopardizing the application of the mentioned IHL presumption. Nevertheless, a way to mitigate those impacts human-machine partnerships experience, could be with specific training.

To conclude, I would like to highlight some general challenges taken from the SKYNET example regarding the compliance of the principle of distinction and the reliance of AI as a decision support aid.

Modern asymmetric and urban warfare entail high levels of controversy regarding who can be lawfully targeted. With the fog of war getting thicker and thicker, commanders and politicians are naturally inclined to search for tools to get guidance on whom can they lawfully target. However, AI should not be a substitute of combat-experienced human judgement. The principle of distinction is highly complex, contextual, and requires a kind of analysis only human minds are suited to fully undertake.

The SKYNET Programme spotlights some relevant AI-related challenges, such as the risk of a lack of ensured protection of civilians, or a problematic selection of “ground truths” to feed the algorithm with. The scarcity of reliable data, and the psychological impacts human-machine partnerships have on the human operator, does not seem to help either in lifting the fog of war.

Therefore, when AI is used as a decision aid, what the algorithm leaves out, or what data is considered relevant, is information which must be kept in mind by the military commander, before issuing any targeting decision, so all the necessary considerations to comply with the principle of distinction are taken into account. Indeed, in my view, ensuring the respect of the principle of distinction necessarily requires human’s assessment, which AI is not “intelligent” enough to replace.

Thank you.