More security through missing information in AI systems?

Karl Hans Bläsius, 4.5.2021, https://www.hochschule-trier.de/informatik/blaesius/

Link to this document: www.fwes.info/fwes-KI-fi-21-1-en.pdf

German Version: www.fwes.info/fwes-KI-fi-21-1.pdf

See also: www.unintended-nuclear-war.eu

Artificial intelligence (AI) techniques are increasingly being used in military early warning and decision-making systems to detect attacks with nuclear weapons, as it is becoming more and more difficult for humans to assess such alerts in the short time available. However, since the data basis in these applications is uncertain and incomplete, even AI systems cannot decide with certainty. These interrelationships are described in https://www.fwes.info/fwes-ki-20-1-en.pdf. Some passages of this article are taken from that article. The issue here is whether security can be increased if certain information is reserved for humans only and is not available to AI components.

Early warning systems

Early warning systems are based on sensors and very complex computer networks and are used to predict attacks with nuclear missiles. Based on such detection, one's own missiles could be launched for a counter-attack before one is devastated and a counter-reaction is made more difficult or prevented. In early warning systems, there can be false alarms, i.e. false attack reports, which can have very different causes. In phases of political detente, the risks are low that the assessment of an alert message will lead to a nuclear attack. The situation can change drastically if there are political crisis situations, possibly with mutual threats, or if other events occur in temporal connection with a false alarm that could be related to the alarm message.

Automatic decisions

The further development of weapon systems with higher accuracy and ever shorter flight times (hypersonic missiles) will increasingly require artificial intelligence techniques to make decisions automatically for certain subtasks. There are already calls in connection with early warning systems to develop autonomous AI systems that fully automatically assess an alert and, if necessary, trigger a counterattack, as there is no time left for human decisions. However, the data available for a decision are vague, uncertain and incomplete. Therefore, even AI systems cannot make reliable decisions in such situations. Automatic detection results therefore only apply with a certain probability and can be wrong.

Contextual knowledge

Due to the uncertain data basis, people also base their decisions on contextual knowledge about the political situation and the assessment of the opponent. If, due to the limited time available to assess alerts, decisions are to be made largely automatically, contextual knowledge about the global political situation would also have to be included here. However, such knowledge is also highly vague, uncertain and incomplete.

The assessment of the global political situation is the subject of a project called "Preview", which the Bundeswehr launched in March 2018 with the aim of predicting crises and wars on the basis of artificial intelligence methods. To this end, large amounts of data are to be analysed automatically. Internet sources as well as military and economic databases and also intelligence information will be evaluated. The type of data used covers a wide spectrum, including trade data, market prices, demographic developments, crime rates, opinions in social networks or data on political violence. Other countries (e.g. Sweden, USA) also have such AI-based systems for predicting crises and wars.

Even though such projects as the Preview project can be useful for the early detection of potential crises, e.g. in Africa, and there is currently no evidence of a connection with early warning systems for the detection of nuclear attacks, such a connection may occur at some point: If an early warning system reports a missile attack and this situation is assessed over several alert levels in the corresponding crisis meetings, it is quite possible that commission members also have access to such a system for war prediction. If this AI system predicts war in such a situation, this can have a significant influence on the assessment of the alert by the commission members.

Decision proposals

A professional assessment by humans of the decisions made by an AI-based system is practically impossible in the short time available. This is already true because automatic recognition is often based on hundreds of features. The AI systems usually cannot provide simple comprehensible justifications and even if recognition features are output by an AI system, they could not be verified in the time available. Humans are therefore usually left to believe only what an AI system provides. The increasing prevalence of AI systems in our everyday world also promotes trust in the decision-making competence of technical systems and it is to be expected that certain decision proposals or situation assessments will automatically be evaluated by humans as facts that are difficult to ignore.

Quote from a novel

In the novel Qualityland 2.0 by Marc Uwe Kling, there is the following dialogue on page 206 in connection with the triggering of World War 3 by an AI system:

> Henryk ponders. "Do you think the Third World War could have been prevented if a human being had been involved in the decision-making chain that triggered it?"
>
> "Depends on the human," says Peter, "Besides ... if you make decisions with the help of an A.I., on the basis of data that the A.I. shows you, so if you only see the world that the

A.I. has prepared for you, don't you almost inevitably decide in favour of what the A.I. suggests? Do you see what I mean?"

"You mean one needs additional input. Input that the A.I. doesn't have."

"Yes."

Solution approach: missing information for AI system

Perhaps this aspect is also relevant in connection with more and more AI in early warning systems. One approach could be, for example, that there are firm agreements that systems like Preview for predicting wars and crises are definitely not allowed to be used in early warning systems for detecting attacks with nuclear weapons. It is then clear that the assessment of the context in relation to the current world political situation can only be carried out by humans. Since the AI system does not have this knowledge.

If people in early warning systems are aware that only they have certain knowledge, but not the AI system, then it is also easier to oppose a decision by the machine.

This is not a technical aspect, but a psychological one. The self-confidence of humans vis-à-vis decisions made by a machine can be strengthened in this way. The human knows he can oppose the machine's decision, he cannot be held responsible for it because he has additional knowledge that the machine does not have.

It should not be very difficult to convince all nuclear powers of such an agreement, as this should seem sensible to all involved. Even if compliance with such an agreement can hardly be verified, there is a lot to be said for complying with it, since it also protects oneself.

Of course, such an agreement cannot replace urgently needed concrete disarmament agreements on nuclear weapons, but it could be a small step towards building trust between nuclear powers and strengthening awareness of possible dangers.

Translated with www.DeepL.com/Translator