AI in early warning systems for nuclear threats

Karl Hans Bläsius, Jörg Siekmann

https://www.hochschule-trier.de/informatik/blaesius/ , http://siekmann.dfki.de/de/home/ ,

Trier, Saarbrücken, 20.7.2021, www.fwes.info/fwes-ki-21-1-en.pdf

See also www.unintended-nuclear-war.eu

Translated with www.DeepL.com/Translator and manually (slightly) corrected.

Summary

Securing the nuclear second-strike capability is the basis of the deterrence strategy that has so far deterred any potential attacker from launching a nuclear attack: "Whoever shoots first dies second". In order to be able to react even when the second-strike capability is threatened, the nuclear powers have developed and installed computer-supported early warning and decision-support systems with the aim of detecting an attack in good time so that they can activate their own nuclear launchers before the destructive impact. Such a strategy is known as a "launch-on-warning" strategy. Although the time required to make a decision when an attack is reported has fallen to a few minutes in recent years, the final decision is still left to humans - not least because of the error-proneness of such systems.

The end of the INF Treaty (INF stands for Intermediate Range Nuclear Forces) has led to a new arms race, also with hypersonic missiles, which now shorten this time span even further. There is therefore so little time for a human analyst to analyse and evaluate these alerts that artificial intelligence (AI) systems have recently been increasingly used for this purpose. However, even an AI system cannot provide reliable results in such applications because the underlying data is *uncertain*, *vague* and *incomplete*. Automatic recognition results are therefore only valid with a certain probability and can be wrong.

Due to the uncertainty of the data base, people also base their decisions on contextual knowledge about the political situation and the assessment of the "opponent", which is further complicated by the potential end of the "open skies" contract. For example, in an alarm case, the operating crew of the American Early Warning System decided that it must be a false alarm because the Soviet head of state was on a state visit to Washington at the time. A serious false alarm in the Soviet early warning system happened in 1983, only a courageous intervention by Commander Stanislav Petrov prevented the nuclear catastrophe.

Even in the case of machine decisions, contextual knowledge of the global political situation must be included in order to evaluate alarm messages, and this knowledge is also uncertain, vague and incomplete. The result of the analysis by an AI system is therefore always only correct within the framework of a statistical probability. The danger of an accidental nuclear war can therefore *not* be reduced by an increasing use of artificial intelligence methods in early warning systems. Due to the uncertain and incomplete data basis, neither humans nor machines can reliably evaluate incoming alarm messages in such a short time.

1 Military early warning and decision-support systems

Early warning systems serve to detect an attack by nuclear missiles on the basis of sensor data. An early detection of an enemy attack should enable countermeasures to be taken before a devastating impact. In case of an alarm message, there are usually only a few minutes available to check it and assess the situation. The situation assessment also depends on the global political situation. For example, in a crisis with mutual threats and the accidental coincidence with other events (e.g. cyber attacks), a false assessment may be made, i.e. the reports could be interpreted as an enemy attack that apparently justifies its own attack. This could lead to an accidental nuclear war.

Due to the increasing number of available sensors and monitoring systems, also in space, the available data and information increases disproportionately in a concrete decision situation. Thus, for the classification of sensor data and the evaluation of an alarm situation, more and more computer-aided procedures, especially of Artificial Intelligence (AI), are required to automatically make decisions for certain subtasks or to prepare human decisions.

The end of the INF Treaty has already led to a new arms race in which hypersonic missiles in particular have a high priority. With these new weapons, the warning times will be further reduced. There are already demands to implement autonomous AI systems for the evaluation and processing of alerts, as there may not be time for human decisions.

Decision-makers such as politicians and military personnel have the expectation that AI systems will be capable of making better decisions than humans in early warning systems, similar to what is expected – say - for autonomous driving.

This article deals with the question whether AI decisions in such early warning systems can be useful and whether these systems can be made safer by AI with regard to possible false alarms.

2 AI decisions

2.1 Classification

Many AI applications are about classification. The task here is to assign a given situation or object to one or more possible classes in a meaningful way. A given situation or object can be described by a series of characteristics (symptoms) and from a large number of given classes (diagnoses), one or more must be selected to which the object or situation fits.

The evaluation of sensor signals in early warning systems is also a classification task: Based on the signals of the sensors it has to be decided whether they indicate a possible attack. The subtasks include decisions about the type of attacking flying objects and the type of attack.

The results of such detection tasks always apply only with a certain probability, i.e., they may be wrong. In many cases, the detection results can also include a measure of security, which expresses how reliably the result is assessed by the automatic detection. However, results can also be false even if the automatic detection classifies them as very safe.

Corresponding findings can also be found in many other applications of automatic classification, such as OCR results (character recognition from images) or automatic invoice recognition. In such applications, invoices are automatically posted and paid without human intervention if the automatic system outputs a recognition result with a certain level of confidence. But even in these cases, incorrect postings and incorrect payment transactions occur.

The same applies to early warning systems. All results of automatic recognition apply with a certain degree of probability and can be wrong. For the usual applications this risk may be acceptable, but for an irreversible nuclear use with millions of deaths and incalculable health, ecological and economic consequences for mankind, this must be assessed differently.

2.2 Data Basis

Decisions in early warning systems are based on extensive data supplied by the sensors as well as on contextual information, such as threat analyses or analyses of the global political situation. Both in the case of human decision-makers and decisions by AI systems, such data and information are needed and must be summarized in a way that is relevant to the decision. However, this data basis is uncertain, vague and incomplete. Processing aspects of vagueness, uncertainty and incompleteness in AI systems are briefly described below.

2.2.1 Uncertainty, vagueness, incompleteness

Reliable knowledge

Example: If x is child of y and y is child of z, then x is grandchild of z

Such a rule can be assumed to be valid. Conclusions based on this again lead to valid, correct results, provided the premises were correct. The aspects of uncertainty, vagueness and incompleteness do not apply here.

Uncertainty

Example: If x is a car and y is the owner of x, then y is the current user of x.

Such a rule does not always apply, there can be exceptions. The user of a car could be a child of the owner. Also in companies, owner and user can be different. So, such a rule is uncertain, it does not always apply.

Another example: If a person x has fever and x has a cough,
and x had contact with y in the last 10 days,
and y has a proven corona infection,
then x also has a corona infection.

This relationship doesn't have to be valid, but it is valid with a certain probability.

vagueness

Example: if x is a heavy car, then x needs a lot of fuel.

The question here is: what does "heavy" mean, what does "much" mean.

In this example, the statements "x is a heavy car" and "x needs a lot of fuel" cannot simply be assigned to the truth values *true* or *false.* These properties are vague, and the truth value could be represented here as any value (real number) between 0 and 1, where 0 stands for false and 1 for true.

Incompleteness

The information needed as a basis for decisions is often incomplete. Since it is often not possible to obtain complete information, assumptions must be made that are typically valid or can be expected. On this basis, conclusions can then be drawn and decisions made.

Example: If x is a bird, then x can fly.

This is typical and applies in normal cases, but there are exceptions. An ostrich is a bird but cannot fly.

Artificial intelligence research has developed procedures to deal with these different types of knowledge and their degree of credibility, but the conclusions are then also not absolute, but only probable, as described below.

2.2.2 Automatic decisions using uncertain data

In practice, there are many interrelationships which are uncertain, i.e., which do not apply without restriction. Our normal everyday knowledge is vague, uncertain and incomplete. Nevertheless, in many situations (e.g. road traffic) conclusions are possible and possibly necessary. In the field of AI, different methods have been developed to treat uncertainties.

Especially important are methods of probabilistic reasoning. Here numerical values are used for the validity of formulas, which are then offset against each other when drawing conclusions. Different probability models differ in how formulas can be linked and how the probability values are then calculated.

Numerical values are also usually used to represent and process vague values.

In many cases, uncertainty or incompleteness can be treated in such a way that a "normal", "typical" rule application takes place first. Typical is that birds can fly, and that the owner of a car is also a user of that car. As long as nothing to the contrary is known and no contradiction arises, a corresponding conclusion can be drawn. In case of a conflict, appropriate measures must then be taken to resolve the conflict. There are also different methods for this type of conclusion in AI, especially logical procedures.

Regardless of the chosen method, the treatment of vagueness and uncertainty is quite complex, and the conclusions are also uncertain, i.e., they can be wrong. Wrong assumptions and wrong conclusions often lead to inconsistencies. In these cases, corrective action can be taken.

As long as no inconsistencies occur, it is automatically impossible to conclude that a conclusion is wrong.

2.2.3 Uncertainty in Early Warning Systems

The data base for decisions in early warning systems is also vague, uncertain and incomplete. This applies to both humans and machines. Errors in early warning systems are caused, for example, by special light effects of the moon or sun or by the detection of flocks of birds by radar systems.

With new technical possibilities, the variety of sensor data for detecting a missile attack will grow. The variety of object types that can be detected will also grow, e.g., due to an increasing number of objects in air space (drones) and in space (satellites, space weapons, defence system). In addition, collisions with space debris and burning up in the earth's atmosphere can cause sensor signals that are detected by early warning systems and are difficult to interpret. The uncertainty of data in early warning systems is therefore likely to increase.

Vague values such as brightness and size also play a role in the evaluation of sensor signals. Signals will also not always occur, so they may be incomplete. This may apply in particular to new steerable missile systems which may evade detection. Furthermore, systems such as "Kalaetron Attack" have been developed for electronic warfare, which should make it

possible to fend off detection by the enemy air defence.[1] In the event of an attack being reported, it cannot therefore be guaranteed that the data can be checked on the basis of several independent signal sources.

The effects of lack of information and false assumptions are illustrated by an incident during the Cuba crisis in 1962, when a Russian submarine, which was in international waters off Cuba, was surrounded and attacked by the American navy. The Americans wanted to force it to surface and had informed Moscow about it. What the Americans did not know:

- The batteries of the submarine were almost empty, the air conditioning had failed and the temperature on board was over 45 degrees.
- Many crew members were on the verge of carbon dioxide poisoning and passed out.
- The submarine has had no contact with Moscow for days.
- The submarine had a nuclear weapon on board which could be used under certain conditions without further clearance from Moscow.

Due to the attacks, the Russian crew believed that the war had already broken out and had to decide whether to use the nuclear weapon on board. The captain of the submarine considered the situation of the submarine and the crew to be hopeless and decided to fire the nuclear torpedo. The torpedo officer agreed to the firing. On this boat three officers were responsible for the decision about the use of nuclear weapons, because the fleet commander was also present. Only if all three agreed, a deployment was allowed. The third officer, Vasily Archipov, refused to give his consent to the launch, thus possibly preventing a nuclear war.

Other documented false alarms also show that the data displayed in early warning systems is uncertain, i.e., it could be wrong. In the event of an alarm, the available information must be evaluated. However, the available information is usually not a complete description of a given situation. Important information may be missing, i.e., for the evaluation of a threat situation assumptions have to be made, which may also be wrong.

2.3 Context Knowledge

In peacetime and phases of political relaxation, the risks that the assessment of an alert leads to a nuclear attack are relatively small. In such situations, human decision-makers assume false alarms in case of doubt. However, the situation can change drastically when political crisis situations arise, possibly with mutual threats, or when further events occur in a temporal context with a false alarm. For this, causes are searched for in a valuation, i.e. attempts are made to find causal connections. If such causal connections are found and are

[1] Behördenspiegel, May 2020, page 45, https://issuu.com/behoerden_spiegel/docs/2020_mai

logically plausible, there is a great danger that they will be assumed to be valid, i.e. that the alarm message will be assumed to be valid even if it is an accidental coincidence in time of independent events.

If the global political situation and other contextual information is not used by automatic decision components of an early warning system, then false alarms are always dangerous, even in peacetime.

If the AI components of early warning systems also use such contextual knowledge for their decisions, then this data basis is also highly vague, uncertain and incomplete.

The assessment of the global political situation is the subject of a project called "Preview", which the German Armed Forces launched in March 2018 with the aim of predicting crises and wars on the basis of artificial intelligence methods. For this purpose, large amounts of data are to be analyzed 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. The AI platform Watson is to be used for this purpose, among others. Other countries (e.g. Sweden, USA) also have such AI-based systems for predicting crises and wars.[2]

Even though such projects like the Preview project may 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 and decision-support systems, such a connection may occur: If an early warning system reports a missile attack and this situation is assessed over several alert levels in the relevant crisis meetings, it is quite possible that commissioners will also have access to such a system for predicting war. If this AI system predicts war in such a situation, this can have a significant impact on the assessment of the alert by the Commissioners.

[2] Süddeutsche Zeitung, 23.7.2018, page 5 and 9.10.2018, page 16

3 Proposals for Decision

Many people demand that decisions to kill people should not be made automatically, but that such a decision should only be made by a person. Such demands mainly concern autonomous weapon systems, but they must also apply equally to counter-reactions to alarms in early warning systems.

Even if such a requirement is complied with, people generally do not have any real opportunity to make a decision because of the short time span involved. The information on which an automated decision proposal is based is too complex to be checked in the short time available (only a few minutes).

A professional evaluation of the decisions made by an AI-based system by humans is practically impossible in the short time available. This is already true because automatic recognition is often based on hundreds of features. AI systems usually cannot provide simple, comprehensible reasons and even if recognition features are issued by an AI system, they could not be checked in the time available.

Therefore, people usually can only believe what an AI system provides.

4 "Blessing" or loss of control with semi-automatic

AI systems are increasingly accessible to the general public through intelligent voice assistants from the major technology groups and through numerous assistance systems in modern motor vehicles. This means that in many applications, AI systems are capable of making better decisions than humans. This is expected to be the case, for example, for autonomous driving; there is even speculation here as to whether autonomous vehicles could prevent traffic accidents. A prerequisite for this is extensive learning data based on many tests, even under real conditions. Even though the number of accidents per distance travelled is significantly lower than with human-operated vehicles, accidents do happen.

However, different conditions apply to AI decisions in early warning systems. For the recently successful "deep-learning" approaches, the problem is that "learning data" for detection tasks in early warning systems are only available to a very limited extent. Testing in real situations is hardly possible. AI-based detections can also be realized on the basis of a few examples, but it is not possible to foresee all variants and exception situations that may occur. Therefore, incorrect classification results can occur.

An accident due to a wrong decision of an AI system, for example during autonomous driving, can also claim individual human lives or, in an industrial context, lead to production stop or loss of income. However, the consequences are limited, and their impact can be limited by making short-term changes to the programs. In an early warning system resulting in nuclear war, however, it would lead to irreversible consequences, with millions of people killed in extreme cases and billions more being deprived of their livelihoods by the nuclear winter that follows.

In the past, there have been repeated threats to launch one's own missiles fully automatically by computer in the event of an attack, with no possibility of human intervention. This suggests that corresponding software components were developed, even if there was no intention to use them in a fully automatic version.

The two crashes of the Boeing 737 Max 8 aircraft on 29.10.2018 and 10.3.2019 show that incorrect and unfavourable programming can make it impossible for humans to oppose the decisions of the machine. Although the two pilots had behaved correctly, they could not prevent the crashes and the death of all passengers. They were not able to correct the wrong decisions of the machine.

If, in connection with early warning systems, one or more of the nuclear forces have software components in use that can (partially) automatically support a counter-reaction when an attack is reported, it cannot be ruled out that, due to errors in design or implementation, these software components may carry out actions that cannot be stopped by humans, similar to the two aircraft crashes. Just as in the case of aircraft crashes, actions in early warning systems are carried out under enormous time pressure within a few minutes.

If such semi-automatic components exist in early warning systems, they will probably be tested, for example, by simulation. Such tests and simulations could also get out of control due to program errors. The Chernobyl catastrophe was triggered by such a test.

5 Comparison of decisions by man or machine

Two examples

Example 1:

In January 2020 the USA had killed the Iranian General Soleimani with a drone attack. In retaliation, Iran attacked American positions in Iraq a few days later. Shortly afterwards a Ukrainian airliner was accidentally shot down in Iran. The operating crew came to the conclusion that the flying object could be an attacking cruise missile. The wrong decision was made mainly because the crew had expected war or an attack by the USA.

In this situation a machine might have made a better decision. Because the pure facts, such as the size of the radar signal, would probably have spoken against a cruise missile. Perhaps a machine could have taken more information, such as flight plans, into account in the short time available. The operating crew had probably overestimated the political context.

Example 2:

A satellite of the russian early-warning system reports five incoming ICBMs on September 26, 1983. Since the correct function of the satellite was established, the Russian officer on duty, Stanislav Petrov, should have passed on the warning message according to regulations. However, he considered an attack by the Americans with only five missiles to be unlikely and decided, despite the data available, that it was probably a false alarm, thus preventing a catastrophe with nuclear strike and counterattack. The incident occurred during an unstable political situation: The retrofitting with medium-range missiles was pending and a few weeks before the Soviets had accidentally shot down a Korean passenger plane over international waters. It is possible that a machine would have been more likely to assess the attack as real and to initiate counter-reactions based on the facts. Petrov had emotionally hoped for a false alarm, did not want to be responsible for the millions of deaths of people and decided accordingly.

Ever shorter decision-making periods

Decisions in early warning systems must be made in very short time spans and a new arms race will further reduce these times.

If, in the event of an alarm, a person still has possibilities to evaluate relevant features, he may conclude that there is not enough time to rule out a false alarm. Consequently, he should not trigger a counter-reaction for safety's sake.

For a machine, the time will be sufficient to create a basis for a decision. If an attack message is classified as valid with a certain probability, an automatic counter-reaction could be initiated, even if it is a false alarm. For a machine, there will be enough time to make a decision, even if it is a false one.

6 Conclusion

Neither humans nor machines can make error-free decisions in early warning systems in such a short time, because the data basis is uncertain, vague and incomplete and cannot be checked by humans in the short time available.

It is neither possible to say that in case of doubt humans make the better decision, nor is it true that in case of doubt machines make the better decision.

In the case of human decisions, the result may depend, among other things, on who is on duty at that time and what the current basic attitude of those on duty is.

In the case of decisions by machines, the result can depend, among other things, on the choice of features with priorities by the programmers or an existing database as a learning basis. At development time, when characteristics and priorities are defined by programmers or a learning basis is established, it is not possible to estimate the consequences for certain alarm situations.

It must not be the case that the survival of the whole of mankind depends on the decision of a single person or a machine. Therefore, the approach of using early warning systems to detect nuclear attacks at an early stage and possibly launch a counterattack before the enemy missiles strike is fundamentally unacceptable, regardless of whether humans or machines ultimately decide. This problem cannot be solved by using AI technologies.

Further information on the topic "Nuclear war by accident": www.unintended-nuclear-war.eu, there are also references to other articles, such as www.fwes.info/fwes-21-1-en.pdf