*Expertise*

# Machine Learning for Intelligence: Opportunities and Risks for National Security

ÉTIENNE VOUTAZ

**Abstract** L'intelligence artificielle (IA) est une technologie disruptive pour toutes les activités humaines. Elle ouvre des opportunités insoupçonnées, permet d'automatiser des processus qui étaient auparavant effectués manuellement et accélère le cycle du renseignement. Mais dans le même temps, ces mêmes outils sont utilisés par des acteurs malveillants (désinformation, cyberattaques et criminalité, déstabilisation politique, etc.).

Dans cet article, nous examinons l'impact de l'intelligence artificielle – plus précisément, de l'apprentissage automatique – dans le domaine du renseignement pour la sécurité nationale en explorant les opportunités et les risques. Nous nous penchons également sur les défis liés à son utilisation dans ce contexte et sur les méthodes actuelles d'IA basées sur l'état de la recherche. Des cas d'utilisation possibles dans le domaine du renseignement sont dérivés et analysés à la lumière des recherches actuelles en intelligence artificielle. Nous formulons enfin des recommandations et esquissons les défis à relever en vue des développements futurs de l'IA.

**Schlüsselbegriffe** Maschinelles Lernen; Künstliche Intelligenz; Nationale Sicherheit; Aufklärung; Risikomanagement
**Keywords** machine learning; artificial intelligence; national security; intelligence; risk management

**DR. ÉTIENNE VOUTAZ** is a scientist specializing in cybersecurity and data science at armasuisse Science and Technology, the Swiss Federal Department of Defence's research and technology division. He has a Ph.D. in mathematics and is affiliated with the Cyber-Defence Campus, which focuses on advancing Switzerland's capabilities in cyber defence and related fields. He is working in particular on new machine learning techniques to detect early warning signals of social instability, such as riots, wars, or revolutions and on the defence of critical infrastructures. He is also working on the management and governance of information systems, holds various certifications (CISA, CRISK, CGEIT) and an EMBA in this field.
E-Mail: etienne.voutaz@ar.admin.ch

**Artificial intelligence is a disruptive technology for all human activities. It opens up unsuspected opportunities, makes it possible to automate processes that were previously carried out manually and accelerates the intelligence cycle. But at the same time, these same tools are being used by malicious actors (disinformation, cyber-attacks and criminality, political destabilisation, etc.).**

**We examine in this work the impact of artificial intelligence – more precisely, of machine learning – in the field of intelligence for national security by exploring opportunities and risks. We also look into the challenges associated with its use in this context and into current AI methods based on the state of research. Possible use cases in the domain of intelligence are derived and analysed in light of current research in artificial intelligence. We finally give recommendations and sketch challenges in view of future AI developments.**

## 1 Introduction

The recent rise of Artificial Intelligence (AI), or more precisely Machine Learning (ML), is revolutionising all human activity by opening up new opportunities. The field of intelligence is no exception to the revolution initiated by the democratisation of machine learning methods. On the one hand, information and patterns that are difficult to access using conventional methods can be detected by machine learning models. On the other hand, intelligence cycles can be accelerated through automation, rapid access to essential information and decision-making support.

However, these opportunities can only be assessed by considering the risks associated with their use. The field of intelligence is no exception to this rule. In this article, we review the opportunities and risks associated with the use of AI in intelligence for national security. We also discuss challenges and offer a range of use cases with varying degrees of granularity.

The technical AI tools used in the intelligence field do not vary fundamentally from current standards applied in all business areas. Risks, on the other hand, must be known and managed in such a way as to comply with current legislation and ethical standards.

Ensuring national security, particularly through intelligence activities, is becoming increasingly complex. ACLED[1], for example, has measured a doubling of conflict levels worldwide over the last five years. The nature of these conflicts is becoming increasingly hybrid, with the AI at the heart of the issues, whether in an offensive or defensive context.

This review is not technical and intended for a wide audience.

The opportunities offered by artificial intelligence cannot be discussed without mentioning the risks associated with its use. Opportunity and risk are two sides of the same coin.

Section 2 examines the opportunities and risks associated with the use of machine learning in an intelligence framework. This part is very generic and the recommendations made are valid well beyond the field

of intelligence. This is because the techniques used in the domain of intelligence are not specific to this area.

Section 3 describes machine learning challenges related to its use for intelligence purposes.

Section 4 contains a number of use cases in the field of intelligence where machine learning can help to turn opportunities into reality. These use cases are highlighted in terms of the associated risks and the current state of research.

A conclusion is finally formulated in Section 5. It is valid for many applications of machine learning and can be summarised as follows:

- Research work is necessary in order to obtain secure, transparent, robust and trustworthy added value.
- It is essential to integrate the risks inherent to the use of machine learning into the company's risk management processes.
- In the context of intelligence, it is important to be able to add value not only with open-source data and large models, but also with own private data and adapted models.
- The acceleration in the life cycle of artificial intelligence models will put pressure on data management and engineering processes.
- The use of own private data for creating an added value through machine learning is only possible if good practices are applied to data management and engineering processes and techniques.

Artificial intelligence makes it possible to create information that is useful for intelligence, but also mis- or disinformation and malicious activities. These same tools can be used to detect activities harmful to national security, but are reactive. We can therefore expect an arms race, and we need to be able to act with agility while respecting the legislative framework, which must also be able to evolve with agility in order to cope with this new dynamic.

# 2 ML and Intelligence: Opportunities and Risks

## 2.1 The Intelligence Cycle

The intelligence cycle, also called intelligence process, is a model describing how intelligence is processed (see Figure 1). This model is abstract by nature and can be adapted for different purposes.

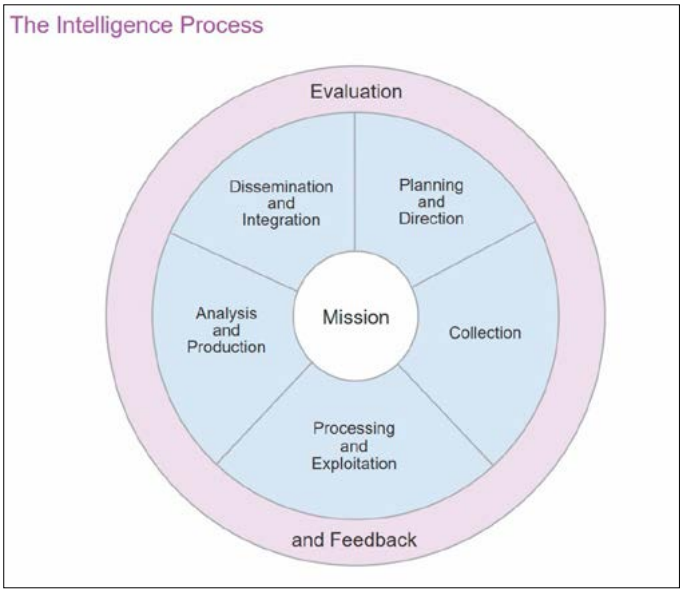


**Figure 1:** The intelligence cycle. The conceptual model consists of the six described steps[1]: direction, collection, processing and exploitation, analysis, dissemination and feedback. (Source: DoD, 2013)

**Direction:** Intelligence requirements are determined by a decision maker to meet his objectives.

**Collection:** In response to requirements, an intelligence staff develops an intelligence collection plan applying available sources and methods.

**Processing and exploitation:** Once the collection plan is executed and the data arrives, it is processed for exploitation.

**Analysis:** Analysis establishes the significance and implications of processed intelligence.

**Dissemination**: Finished intelligence products take many forms depending on the needs of the decision maker and reporting requirements.

**Feedback**: The intelligence cycle is a closed loop; feedback is received from the decision maker and revised requirements issued.

The proliferation of data and advances in ML now make it possible to automate certain tasks that were traditionally carried out by HUMINT (Human Intelligence). ML methods can be used to analyse large quantities of texts, images, videos, sounds, signals, networks and more, enhancing the speed and quality of the intelligence cycle.

## 2.2 The Intelligence Domains

A non-exhaustive list of intelligence disciplines for which ML can create opportunities is given below. It's worth noting, however, that the boundaries between these disciplines are becoming increasingly blurred; the nature and quantity of open source data is evolving, and there is now a great deal of open access data ranging across all intelligence disciplines (satellite images, military information, etc.).

**Imagery Intelligence (IMINT)**. Analysis of aerial and satellite imagery that can be completed by Geospatial Intelligence (GEOINT).

**Open Source Intelligence (OSINT)**. Analysis of data gathered from open sources, with the subdomain of Social Media Intelligence (SOCMINT).

**Signal Intelligence (SIGINT)**. Analysis of data gathered from interception of signals, with the special cases of Communication Intelligence (COMINT) (messages, voice) and Electronic Intelligence (ELINT) (electronic signals without speech or text).

**Cyber Threat Intelligence (CTI)**. Analysis and dissemination of data regarding potential or existing cyber threats.

**Advertising Intelligence (ADINT)**. Usage of targeted advertising. This discipline is widely used by major technology groups. This discipline is subject to attacks that enable a malicious entity with limited resources to extract personal information[2]. This discipline is not covered in this document because it does not concern national security.

## 2.3 Opportunities and Risks

### 2.3.1 Opportunities

We describe in **Table 1.** general opportunities offered by ML in the field of intelligence, and the most widespread techniques for seizing them. Opportunities cannot be assessed in isolation from the associated risks. Figure 2 shows that opportunities and risks are two sides of the same coin. Risks are treated in Section 2.3.2.

| Discipline | Opportunities | ML-Tools |
|---|---|---|
| IMINT | Object detection, transcription of image to text, detection of fake or AI-generated images or videos, localisation of objects | Anomaly detection, LLM, LVM and genAI, DL |
| OSINT | Analysis of all types of data available in open source, anticipation of conflicts, riots, terror attacks or wars, situational awareness, causal analysis, detection of mis- and disinformation | Anomaly detection, NLP, LLM and genAI, semantic analysis, sentiment analysis, graph-based methods, predictive analytics, DL, translations and transcriptions |
| SIGINT | Analysis of the electromagnetic spectrum, network analysis | Anomaly detection, DL, graph-based methods |
| CTI | Anticipation, classification, mitigation and attribution of cyber-attacks, cyber situational awareness | LLM, DL, NLP , graph-based methods, anomaly detection |

**Table 1:** ML opportunities in intelligence. Generic tools used to exploit the opportunities are also mentioned. Note that this is just a non-exhaustive snapshot. (Source: Author)

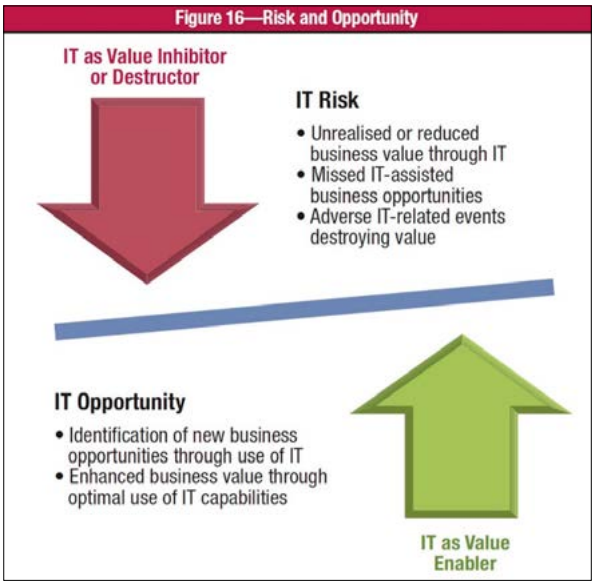


**Figure 2:** Opportunities and risks from ISACA Risk IT framework[3]. (Source: ISACA, 2020)

The boundaries between the intelligence domains described are becoming increasingly blurred, and multimodal analyses are becoming more and more necessary. For example, aerial images can be correlated with smart sensors (e.g., smartphones), data from social media and other sources. Concrete use cases are sketched in Section 4.

### 2.3.2 Risks

Using AI is not without risks, and it's important to identify and manage them. Like any disruptive technology, the biggest risk is certainly not using the technology wisely. The current trend is to ask: how can I use AI to solve this or that problem?

***"Using AI is not without risks, and it's important to identify and manage them. Like any disruptive technology, the biggest risk is certainly not using the technology wisely. The current trend is to ask: how can I use AI to solve this or that problem?"***

AI is just one technology among many; certainly a disruptive and promising one, but the question we should be asking ourselves is: I have a problem; can I formulate the problem clearly and how do I solve it? Perhaps AI is a good answer, but it's not necessarily the case.

NIST has recently published a risk management framework[4] to minimise potential negative impacts of AI systems, such as threats to civil liberties and rights, while also providing opportunities to maximise positive impacts. The notion of trustworthiness is at the heart of this document. It is presently very difficult and challenging to guarantee the trustworthiness of AI models[5]. Characteristics of trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed (see Figure 3).


**Figure 3:** Characteristics of trustworthy AI systems. Valid and Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable and Transparent is shown as a vertical box because it relates to all other characteristics[4]. (Source: NIST AI 100-1, 2023)

The field of intelligence is particularly sensitive to the risks indicated, if we accept the hypothesis that these tasks are regalian in a democratic state. The discipline is strictly legislated, and the following conditions merit particular attention:

- Privacy-enhancement and bias management: data protection does not tolerate personal profiling and models should be trained with data not exhibiting personal biases like gender, religion, origin, politics, etc.
- Accountability and transparency: it is a minimum requirement for regalian tasks.
- Explainability and interpretability: every use of ML conceals a certain amount of mystery. It's important to avoid 'black box' approaches of critical algorithms, and to evaluate results in detail and transparently. But it is worth remarking that human beings are also subject to cognitive biases (see Section 3.4).

- Safety, security and resiliency are general minimal requirements.

The ISO standards[6,7] provide requirement for establishing, maintaining and continually improving an AI management system, build and maintain AI systems over the entire life cycle.

Particular attention must be paid to the use of Generative Artificial Intelligence (genAI). It's tempting to use large generative models such as Large Language Models (LLM) in a variety of situations. These have the advantage of being pre-trained, but hide many risks. The training phase was carried out with a very large amount of data available on the web, the quality of which has not been verified (incorrect, biased information, misinformation). Practically the entire web was used or will soon be used for the training phase, and with the proliferation of AI-generated data, the risk of phagocytosis by training on such new data is real. We must never forget that genAI is not intelligent (AI in general either) but peremptory: it always delivers a result. There are however methods for mitigating these risks (fine tuning, Retrieval-Augmented Generation (RAG)). Risks and mitigation techniques related to the use of LLM are exposed by Kucharavy et al.[8].

NIST has also published a risk management framework for genAI aligned to the previously mentioned one[4]. Appendix B contains the main risks associated to the use of genAI.

We note that the EU AI Act[9] excludes national security initiatives from its scope: *"If, and insofar as, AI systems are placed on the market, put into service, or used with or without modification of such systems for military, defence or national security purposes, those should be excluded from the scope of this Regulation regardless of which type of entity is carrying out those activities, such as whether it is a public or private entity."* (Art. 24).

# 3 AI-Related Intelligence Challenges

In this section we present some general challenges related to the use of AI in intelligence.

## 3.1 Data Enrichment and Fusion

The proliferation of different types of data (text, image, sound, signal, voice) makes the issue of data fusion in an application context particularly interesting. Suppose an object or system can be detected by aerial or satellite imagery (a military system or critical infrastructure, for example). The question is whether other sources of data relating to this case can enrich the information derived from imaging and, if so, what fusion methods can be used to optimise the information produced.

In this case, we can imagine having the GPS coordinates of the system in question and, on this basis, analysing data corresponding to the space-time determined but from different modalities (video cameras, photos from smartphone, social media content, press articles, electromagnetic signals, etc.).

In this way, the system identified by imaging can be analysed in greater detail. A review of multi-modal fusion methods is given by Zhao et al.[10].

## 3.2 Anomaly Detection

A large number of questions can be modelled as anomaly detection problems. The aim is to spot abnormal and rare states in the mass of information available. Traditionally, anomalies are detected using time series. A signal (a time series) can be extracted from a specific problem and analysed. However, the extraction of this signal may be linked to a loss of information. It is therefore essential to be able to detect anomalies in a more general context.

Large ML models are not suitable for this type of analysis. The time dimension is crucial and not taken into account by these models (see Section 3.3).

Graphs (or networks) are used to model a number of intelligence-related problems. They serve to analyse data from social networks, newspaper articles or various communications. If the textual data sources are moreover open, the combined use of Natural Language

Processing (NLP) methods can also refine the analyses. Graphs are also used to model complex systems in which various entities interact.

Today, most communications are encrypted and the content of messages (e-mails, text or voice messages, etc.) is inaccessible. In this case, the only source of analysis left is the communications metadata, essentially in the form of non-attributed graphs.

By analysing such graphs over time (temporal graphs), it is possible to detect anomalies or changes in dynamic regimes without having access to the contents of communications. These methods also have the potential to improve the detection of anomalies and regime shifts in a SOCMINT context. The same techniques can be used to detect fraud in financial or cryptocurrency transactions[11,12].

The concept of anomaly detection is standard and well documented for time series in the euclidean space[13]. The concept of regime shift is also widely studied (change point detection[14], early warning signals[15]). To date, there has been little research into the generalisation of anomaly detection but more importantly regime shift for time series of graphs[16].

## 3.3 Concept Drift

LLM and Large Vision Models (LVM) are increasingly used for a wide variety of critical tasks for military and intelligence purposes including, e.g., the analysis of satellite images, visual surveillance of sensitive perimeters, or machine translation of text documents. Every model is trained on a subset of all (potentially) available data, which was acquired within a fixed and finite time-window. Most ML models, including foundation models (e.g. LLM, Diffusion-based models, etc.), incorporate in some way inductive reasoning. They generalise information extracted from a finite set of observations to the whole population (all possible realisations of samples).

This raises the question whether a model extracts information that generalises and stays valid for a long period of time. This problem is known as 'concept/data drift'. This phenomenon arises when environmental conditions, in which the deployed models operate, change over time (e.g. through seasonal changes).

Models employed in military and intelligence applications are confronted with precisely this challenge and as such are prone to suffer from concept drift. As a consequence, the statistical properties of the target variable of a model may change over time and information generated by the model may not be representative any more. This results in incorrect predictions of the model (concept drift) and in the worst case in unexpected/undefined behaviour of the system relying on the model.

Addressing this usually involves the re-training of the model (see Figure 4). This process can be very demanding in both data and energy requirements, more so when facing the foundation models (e.g. GPT, Stable Diffusion, etc.).

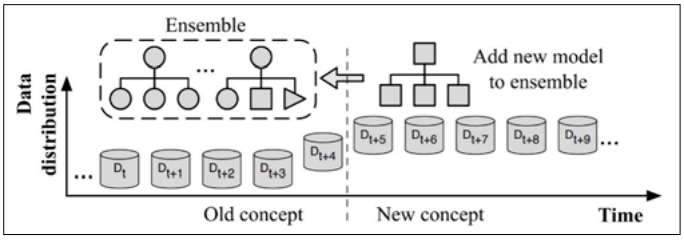


**Figure 4:** An example of concept drift mitigation. A new model is added to the ensemble when a concept drift occurs[17]. (Source: Garcia et al., 2025)

What is currently missing is a comprehensive benchmark for studying the impact of concept drift on modern ML models including foundation models. This benchmark could incorporate a framework based on well-defined datasets, problem domains, and evaluation criteria (e.g. time window) which would enable the assessment of modern models, in particular foundation models, with respect to their susceptibility to concept drift. Moreover, methods are needed to automatically detect concept drift during deployment of the model and strategies to mitigate concept drift in the most energy-efficient way possible.

As outlined, models employed in military and intelligence applications are typically confronted with concept drift during their deployment. In particular, any model which processes data from a sensor that in turn is sensitive to changing environmental conditions will suffer from concept drift. In computer vision, it has been shown that models used for object detection from thermal imaging in a surveillance setup suffer greatly

from seasonal changes (temperature, humidity). In NLP, it can be demonstrated that language models trained on retrospective factual data, will provide outdated suggestions for current events. This is particularly troublesome for applications in intelligence analysis.

Hence, the capability to evaluate ML models with respect to their susceptibility to concept drift is important to guarantee the most robust integration of these models in operational systems used for military and intelligence applications. Furthermore, the capability to detect concept drift during deployment is important to prevent generation of false or misleading data and in the extreme case avoid undefined behaviour of computer systems integrating such models. The state of the art of the drift-detection methods is given here[18, 19, 20].

### 3.4 Decision Support

The development of large AI models enables human beings to make decisions with the support of AI. This new phenomenon (AI-assisted decision-making) deserves special attention in intelligence. It is well established that humans are subject to various cognitive biases[21] (confirmation, anchoring, availability, hindsight biases, etc.). As the large AI models are trained with data produced by humans, there is little doubt that these biases persist in the AI models. The fundamental question is to assess the impact of AI in AI-assisted decision-making processes. The constituent spaces to capture different interactions in human-AI collaboration are exhibited in Figure 5.

***"The development of large AI models enables human beings to make decisions with the support of AI. This new phenomenon (AI-assisted decision-making) deserves special attention in intelligence."***

It is for example shown that a de-anchoring strategy can effectively improve collaborative performance when the AI model has low confidence and is incorrect[22]. It is interesting to note that the proposed mitigation method is a time allocation strategy, i.e. users are asked to take the time to evaluate the proposals resulting from the AI, whereas the use of AI in decision-making processes is often justified by the need to save time.

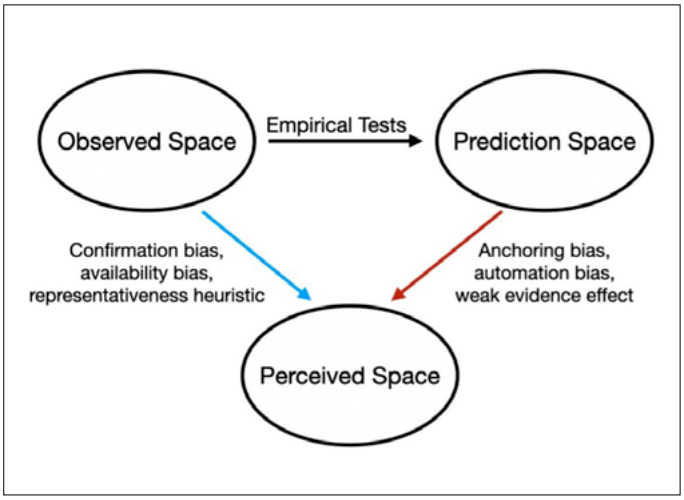


**Figure 5:** Constituent spaces and interactions in human-AI collaboration. The perceived space represents the human decision-maker, the observed space consists of the feature space and all the information the decision-maker has acquired about the task and the prediction space represents the output generated by an AI model[22]. (Source: Rastogi et al., 2022)

Four main ways cognitive biases affect or are affected by AI systems are identified by Bertrand et al.[23]:

1. cognitive biases affect how AI methods are designed,
2. they can distort how AI techniques are evaluated in user studies,
3. some cognitive biases can be successfully mitigated by AI techniques, and, on the contrary,
4. some cognitive biases can be exacerbated by AI techniques.

A different point of view is defended by Hagendorff et al.[24]. The authors argue for the implementation of human cognitive biases in learning algorithms and for more inspiration from human cognition.

The growing importance of AI in decision-making calls for major interdisciplinary work (psychology, linguistics, neurology, data science, ML).

### 3.5 Data Management and Engineering

It's worth mentioning the importance of data management processes. Without quality data, the full potential of AI cannot be realised. If we resort to using only large, pre-trained foundation models, we miss out on the real potential of AI. This is particularly true in the field of intelligence, where the added value generated on the basis of specific and non-open-source data needs to be exploited. Good practices in this area are a sine qua non for the full exploitation of AI and exposed in Appendix A.

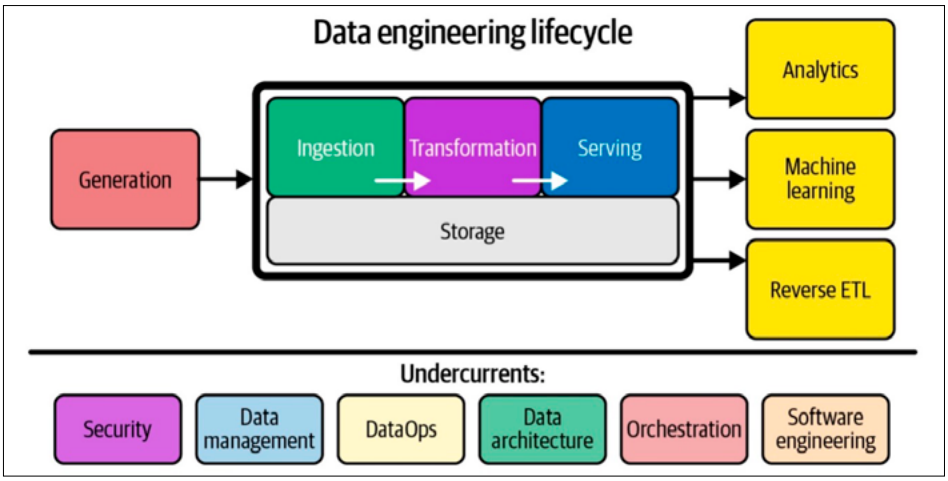


**Figure 6:** The data engineering life cycle consisting of ingestion, transformation, serving and storage[26]. (Source: Reis and Housley, 2022)

It should be noted that good practice in data management, regardless of its use in AI, remains valid: data quality approach and strategy, data profiling, cleansing, archiving and retention, data life cycle management, etc. The use of AI requires particular attention to these critical processes, because AI can only deliver good results with quality data.

Moreover, we must not neglect the data engineering stages required for its use by ML[25] (data cleaning and curation, segmentation, extraction, conversions, etc.) (see Figure 6).

The life cycle of AI models will be accelerated with Machine Learning Operations (MLOps), and data management and engineering will have to adapt to this pace. A paradigm shift will therfore be needed: data needs to be managed as a product or service, rather than as a necessary and mere database component for the purpose of AI-supported business processes.

# 4 Use Cases

In this section we describe some use cases in the field of intelligence. The aim is not to draw up an exhaustive list of the fields of application of ML for intelligence, but to illustrate its use with the help of cases directly related to the current state of research.

**Table 2** maps the use cases with the intelligence disciplines.

| Use Case | Intelligence Discipline |
|---|---|
| Section 4.1: Conflict Early Warning Systems | OSINT, GEOINT |
| Section 4.2: Dynamics in Arms Races of AI Image Generators | OSINT, SOCMINT, IMINT |
| Section 4.3: Battle Damage Assessment | IMINT, OSINT |
| Section 4.4: Cyber Threat Intelligence and Cyber Resilience | CTI |
| Section 4.5: Object Detection and Situational Awareness | SIGINT, ELINT, IMINT |

**Table 2:** Mapping of use cases with intelligence disciplines. (Source: Author)

## 4.1 Conflict Early Warning Systems

Anticipation and early detection of conflicts are important tasks to ensure national security. Anticipation and strengthening their early detection is an essential element of the vision and strategy of a state. The "Switzerland's Security Policy Report"[27] of the Federal Council of 24 November 2021 mentions for example the following objectives:

| Dimension | Component | Variable | Source |
|---|---|---|---|
| **Political** | Regime Type | Democracy | V-DEM |
| | Regime Performance | State Capacity | V-DEM |
| | | Repression | V-DEM |
| | | Corruption | V-DEM |
| **Security** | History of Conflict | Recent internal conflict | UCDP |
| | | Years since last conflict | UCDP |
| | Current Conflict Situation | Neighboring conflict | UCDP |
| | | Homicide ragte | IMHE |
| **Social** | Social cohesion and diversity | Female Empowerment | V-DEM |
| | | Ethnic Exclusion | EPR |
| | | Transnational ethnic ties | EPR |
| **Economy** | Development and distribution | GDP per capita log | World Bank |
| | | Income inequality | WID |
| | | Trade openness | World Bank |
| | | Oil Exports | World Bank |
| | Provisions and employment | Food security | FAO |
| | | Unemployment | World Bank |
| **Geography – Environment** | Environment | Drougths | SPEI/CSIC |
| | | Temperature Change | FAO |
| **Demographics** | Demographics | Population log | UN |
| | | Youth bulge | UN |
| | | Child mortality | World Bank |

**Figure 7:** GCRI variables and data sources[31]. (Source: Schvitz et al., 2022)

- Strengthening the early detection of threats, dangers and crises,
- Strengthening international cooperation, security and stability,
- Increased focus on hybrid warfare.

The data available in an open world is growing rapidly and its quality is improving. A task that used to be done manually with the help of human resources (HUMINT) can now be automated in a quantitative approach without ignoring human capabilities.

***"A task that used to be done manually with the help of human resources can now be automated in a quantitative approach without ignoring human capabilities."***

Social instabilities often originate on or are catalysed by social media, and abnormal activity there can be a harbinger of disruptive events in the physical world. This was demonstrated in the case of the 2011 Arab Spring[28]. It is for example possible to analyse early warning signals for socially disruptive events, like riots, wars, or revolutions using only publicly available data[29,30].

This is an important task for a state, as it has to guarantee the safety of its citizens abroad and carry out a risk analysis for state visits outside its borders. It is also important to be able to anticipate the flow of refugees arriving in a country so as to be prepared.

Traditionally, human resources are used to analyse conflict risks for the purposes of anticipation and early detection. The available data is qualitative and unstructured. However, the quality and quantity of this data is evolving and enabling a more quantitative approach known as Conflict Early Warning Systems (CEWS).

The Joint Research Center of the European Commission[31] defines for example:

- **A conflict frequency:** This variable takes the value of 1 if there is at least one ongoing conflict with more than 25 battle-related deaths in a country-year, and otherwise remains 0, and is used to model and predict the probability of conflict, and
- **An intensity of conflict:** This variable counts the total number of battle-related deaths in each country-year and is used to model and predict the intensity of conflict.

This is done using a set of data as displayed in Figure 7. On the basis of these data, various models could be developed and compared with the state of the art. A specific region, e.g., sub-Saharian Africa or Africa with country months, will be used for the development and validation of the models[32]. The predictions should cover a range from a few months to a few years. Particular attention should be paid to the reliability of the models (interpretability, robustness, security, transparency, traceability, trustworthiness).

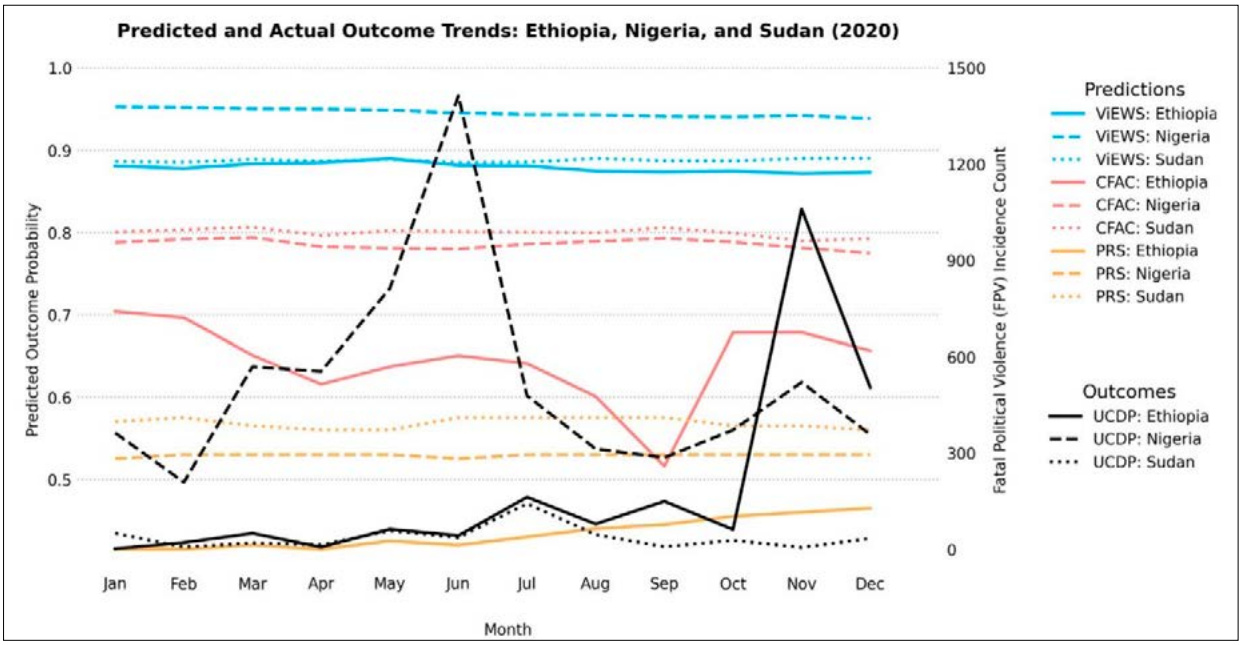


**Figure 8:** Comparison of the country-month predictions (conflict frequency as probability) of three CEWS (ViEWS, CFAC and PRS) for three countries in 2020. The outcomes represent the real and dynamic value of conflict deaths[32]. (Source: Rød et al., 2024)

A fundamental problem is defining the term "conflict" in both qualitative and quantitative terms. This term differs depending on the objectives to be achieved and influences the selection of data to make predictions. An approach that is consistent with the current state of the art should be preferred in order to be able to compare different models.

The source data used covers a wide range of public information. However, the relevance of this data in relation to the type of conflicts being analysed is unclear. It is not enough to include a particular type of data just because it is assumed to be related to conflicts. A causality analysis based on statistical or Convergent Cross Mapping[33] (CCM) methods is required. For example, if there is a cause-and-effect relationship between conflicts and per capita gross domestic product, the causality analysis will determine which of the two variables causally influences the other.

Another open question is the use of data from social media and news. Can these provide added value?

Artificial intelligence models are constantly being improved. Which ones are best suited for this type of prediction?

Experience has shown that in many cases, the intensity of conflicts is low over a long period of time before a sudden event occurs. These events are difficult to predict using current methods. Furthermore, the quality of short-term forecasts is currently insufficient and the static nature of the forecasts does not reflect the dynamics that can be observed in reality (see Figure 8). The use of social media or the early warning signals theory[15] could bring improvements here.

There is evidence that a region surrounded by conflict areas is highly likely to be conflict-prone itself. It would be interesting to analyse the potential of using graph structures to capture the spatial dimension of the problem.

## 4.2 Dynamics in Arms Races of AI Image Generators

Today disinformation, "fake news", and other forms of influence-related activities are widely acknowledged to be a form of warfare. In the public, especially social media, deepfakes and other synthetically generated media have the potential of disruption and destabilisation. AI image generators are possible threats for any political and social system – especially democracies, in which public opinion is a fundamental pillar of daily life. Historically, propaganda and disinformation have always been part of warfare, yet especially since the advanced development of AI, information warfare has risen to a problem that has to be taken seriously[34].

In the context of AI development, many speak of a new arms race of AI. To get an overview and fundamental knowledge to model and simulate influence-related activities through images in such arms races, it is important to explore the dynamics of arms races of im-

age generators and to explore the engineering requirements for modelling and simulation of that arms race.

Information warfare happens on the battlefield in a military context. Images showing vehicles or troops in certain locations, or performing certain manoeuvres to lure an adversary into an ambush or divert them from the actual combat zone can be generated. But destabilisation and disruption happens not only on the traditional battlefield, rather it shifts to the social, political, economic and juridical domain. Especially the arms race of AI image generators influences these domains yet is also dependent of them. More precisely, a reciprocal influence is apparent between technological actors such as algorithms, chip production, data, energy and other actors such as political decision-making and elections, societal movements and public opinions, financial investment and market development, and ethical and moral elements influencing the law.

***"But destabilisation and disruption happens not only on the traditional battlefield, rather it shifts to the social, political, economic and juridical domain. Especially the arms race of AI image generators influences these domains yet is also dependent of them."***

Because of the reciprocal influence between actors in certain domains, networks are thus formed. In other words: chips, algorithms, data, energy are actors and together they form a network of interdependence in the domain of technology and allow to study how these actors are connected to one another. How does the mechanism of social and political disruption really work? To what extent are economic, juridical and technological actors involved? By influencing each other actors thus form together networks which again operate in so-called domains (law, politics, economy, etc.).

It is not only important to identify content used for disinformation campaigns, but also the actors producing it and the techniques and tactics used. However, this is an arms race: given the rapid development of AI, the methods that enable this identification become obsolete very quickly and are circumvented.

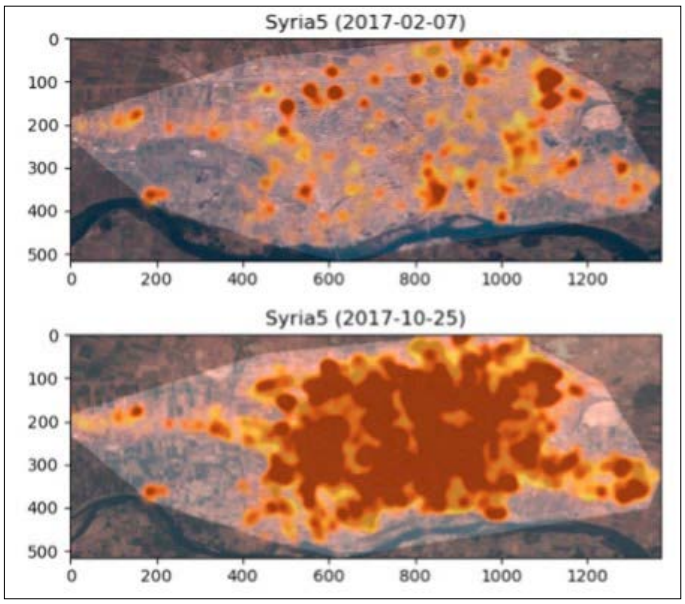


**Figure 9:** Damage density evolution for Raqqa, Syria[35]. (Source: Gazel-Anthoine, 2022)

## 4.3 Battle Damage Assessment

In times of conflict, the destruction of buildings and infrastructure in war zones warrants special attention because it is an indicator of harm directed towards the civilian population. The quantitative process which is aimed at evaluating the damage inflicted on a particular settlement over time is called Battle Damage Assessment (BDA) (see Figure 9).

The main data source for BDA are satellite images taken before and after the event (i.e. pre-event and post-event images) but does not exclude the use of social media or other public sources. Intelligence drawn from BDA can be used for various purposes including the study of conflict development, planning of humanitarian relief efforts, human-rights monitoring and media reporting. BDA is usually posed as an image segmentation and/or image detection problem, in which either intact/destroyed infrastructure is segmented and/or detected. The type of satellite images can vary from low- to high-resolution images, and from optical to Synthetic Aperture Radar (SAR) images.

Meanwhile, ML methods are widely used to quantitatively assess satellite images of conflict zones. In contrast to the assessment of damage caused by natural disasters, BDA has been an understudied problem. Consequently, there is an opportunity to improve on either the performance and/or usability of the currently available methods.

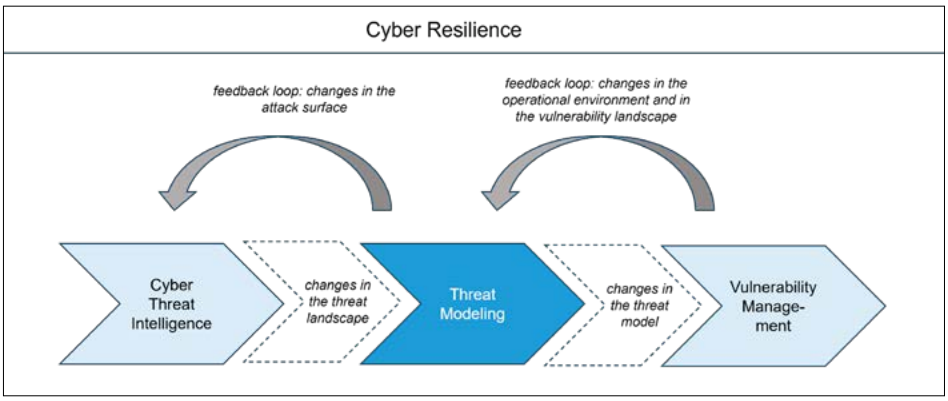


**Figure 10:** Cyber resilience internal dependencies. Feedback loops are necessary to guarantee a holistic approach to cyber resilience. (Adapted from Podlesnik and Mihelič[36].)

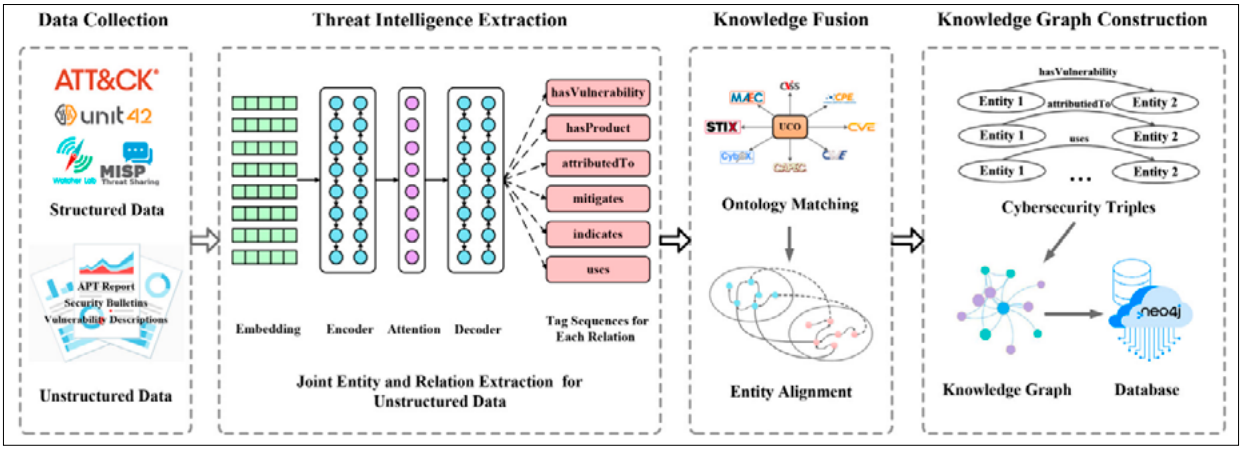


**Figure 11:** Framework for threat intelligence information extraction and fusion[38]. (Source: Guo et al., 2023)

## 4.4 Cyber Threat Intelligence and Cyber Resilience

CTI has been demonstrated to be an effective element of defensive security and cyber protection. Automated methods are needed today in order to stay current with the magnitude of attacks across the globe. It provides information about potential threats and enables organisations to make informed cybersecurity decisions. It involves collecting, analysing, and disseminating information about emerging cyber threats, allowing organisations to stay informed and responsive.

CTI can be embedded in the more general concept of cyber resilience. Cyber resilience is about preventing or mitigating cyber-attacks but also about maintaining critical functions and ensuring the ongoing operation of systems and services in the face of adversity. It refers to an organisation or system ability to withstand, adapt to, and recover from disruptive cyber events. It could be defined as the system ability to recover or regenerate its performance after a cyber-attack degrades its performance.

Threat Modeling (TM) is a proactive approach to identifying and analysing potential threats to a system. It focuses on identifying and mitigating vulnerabilities before they can be exploited. Vulnerability management is an information security continuous monitoring capability that identifies Common Vulnerability and Exposure (CVE) on devices that are likely to be used by attackers to compromise a device and use it as a platform from which to extend compromise to the network.

The three concepts of TM, CTI, and cyber resilience may be intrinsically linked and can enhance the cybersecurity posture of organisations (see Figure 10). TM helps organisations identify and mitigate vulnerabilities, while CTI provides the necessary information to understand and anticipate potential threats. TM focuses on internal systems and vulnerabilities while CTI focuses on external threats. Cyber resilience, in turn, enables organisations to anticipate, adapt, and respond to cyber incidents, ensuring the continuity of critical functions and services.

A vast amount of data, both structured and unstructured, is available in the field of cyber resilience. Much of it is open-source. AI can help to merge all this data in order to detect relationships not visible to the naked eye, malicious actors with identical behaviour (attribution), suspicious clusters, etc. NLP and graph-based methods (knowledge graphs) enable to process human

language and extract meaningful threat intelligence from unstructured text sources, such as news articles and reports[37] (see Figure 11). The challenge is to use AI to produce information and automate cybersecurity processes.

The various devices on a computer network can be telemetered and send a quantity of information to operators. AI can help humans to categorise this data, detect anomalies and identify new threats. Human manual work cannot be replaced, but AI enables humans to sort things out beforehand, otherwise they can only look for a needle in a haystack.

Threat information must be actionable, current and credibly validated if they are to be ingested into computer operated defence systems. False positives degrade the value of the system. A variety of methods have been developed to create learning models that can be integrated with firewalls, rules and heuristics. In addition more work is needed to effectively support the limited number of expert human hours available to evaluate the prioritized threat landscape flagged as malicious in a Security Operations Center (SOC) environment[39].

### 4.5 Object Detection and Situational Awareness

A military command needs to analyse the situation in real time. To do this, various sensors can be used to send back information in the form of text, images or video streams. In this context, AI can be used to recognise objects such as vehicles, troops or aircraft, and to classify them as friend or foe[40].

Based on this situational analysis, AI models can then propose different scenarios depending on the mission and objectives.

While current ML models can detect objects very well in good conditions, identification becomes more difficult if conditions change (weather, environment, camouflage, etc.). The use of synthetic images can be useful in this context[34]. In addition, it is currently very difficult to identify abstract concepts such as violence in images.

The electromagnetic spectrum is becoming increasingly crowded. On the one hand, this makes reconnaissance (SIGINT) more and more challenging and, on the other hand, radio systems must have more flexibility and cognitive capabilities in order to be able to dynamically select the available frequencies. In both cases, the correct recognition of received signals is crucial.

The proliferation of drones, or Unmanned Aerial Vehicule (UAV), has raised significant safety concerns due to their potential misuse in activities such as espionage, smuggling, and infrastructure disruption. Drone detection and classification systems that operate independently of UAV cooperation can be effectively performed using DL[41].

## 5 Conclusion

### 5.1 Summary

Machine learning offers great opportunities in the field of intelligence, as it does in many other areas. The technology is disruptive and deserves the utmost attention. We have presented various opportunities in relation to current research topics. These opportunities also represent risks and cannot be realised without a number of preconditions.

On the one hand, the use cases presented show that a certain amount of research work is necessary in order to obtain secure, transparent, robust and trustworthy added value. However, we must not lose sight of the fact that machine learning is just one technology among many, and that it is not the solution to every problem.

On the other hand, the opportunities cannot avoid the risks associated with the use of this technology. It is essential to integrate the risks inherent in machine learning into the company's risk management processes.

Current trends give the illusion that large, pre-trained models can solve all problems. As we have described, this is an illusion. It is certainly important to work with large models and open-source data, if only to assess the potential of an ill-intentioned lambda actor.

However, the added value of the application of machine learning, in the context of intelligence, can only

be achieved with not just public but quality own data. This potential for using own data is only possible if good practice is applied to data management and engineering processes.

Two opposing forces are at work here: on the one hand, silos need to be broken down and data made accessible in the form of products or services, and not simply as a medium for business applications in databases; these silos also need to be overcome in terms of management and governance. On the other hand, experience shows that quality data with integrity and timeliness can only be managed as close to the business as possible.

The challenge is to find the right compromise between centralisation to eliminate silos and decentralisation to ensure quality. And in the context of intelligence, data confidentiality, privacy and security is a stumbling block that must not be overlooked.

### 5.2 Future Development

In the future, we can expect to see an arms race in the use of AI for intelligence purposes. Defence and detection processes are reactive, while threats are intensifying in step with technological developments. The dynamic developments in AI make it possible to escape the defence processes quickly. The challenge is to be able to adapt to this dynamic with agility.

It is nevertheless undeniable that the AI will speed up the intelligence cycle while increasing its quality. The key point will be human-AI teaming, i.e., use AI as an augmentation of human intelligence. But these same tools can be used for both defensive and offensive purposes. It is therefore vital to remain attentive and active so as not to suffer the asymmetry of a technological deficit.

***"It is nevertheless undeniable that the AI will speed up the intelligence cycle while increasing its quality. The key point will be human-AI teaming, i.e., use AI as an augmentation of human intelligence."***

There is currently some pressure to legislate on the use of AI. We can therefore expect to have to use these tools within a stricter or clearer framework, particularly with regard to data protection.

It is presently very difficult to guarantee the trustworthiness of AI models, which poses an ethical and legislative challenge for their use. However, we can expect to see greater maturity in this area in the future. Various measures, such as the Swiss AI Initiative[2], are designed to create value within a reliable framework.

As Prof. Sarah Summers from the Zurich University mentions[42]: *"Even if an AI system makes a 'correct decision', the question remains as to whether this can be explained or justified. (...) This challenge emphasises the importance of legal regulation. In liberal democracies, the commitment to the rule of law and the idea that the state must follow certain procedures in the prevention, investigation and punishment of criminal offences is of central importance. It is not only the correctness of the decisions made that matters, but also how they are made."*

A balance will have to be found between regulation, agility and security objectives. Malicious actors act beyond any legislative or ethical constraints.

## Acknowledgement

The author would like to thank his colleagues at armasuisse S+T for their contribution. In particular Albert Blarer for Section 4.1, Raphael Meier for Sections 3.3, 4.2 and 4.3, Ljiljana Dolamic for Section 3.3, Christof Schüpbach for Section 4.5 and Gérôme Bovet for various discussions.

The author reports there are no competing interests to declare and no funding was received. ◆

### References

1 DoD. 2013. "Joint Intelligence." Joint Publication 2-0.
2 Vines, Roesner and Kohno. 2017. "Exploring ADINT: Using Ad Targeting for Surveillance on a Budget - or - How Alice Can Buy Ads to Track Bob." *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society*. https://api.semanticscholar.org/CorpusID:45344516.

3 Risk IT Framework. 2020. ISACA.
4 NIST AI 100-1. 2023. "Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1.
5 Kowald et al. 2024. "Establishing and Evaluating Trustworthy AI: Overview and Research Challenges." *Frontiers in Big Data* 7. https://doi.org/10.3389/fdata.2024.1467222.
6 ISO/IEC 5338:2023. "Information Technology – Artificial Intelligence – AI System Life Cycle Processes.".
7 ISO/IEC 42001:2023. "Information Technology – Artificial Intelligence – Management System.".
8 Kucharavy et al. 2024. *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*. Springer Nature. https://library.oapen.org/handle/20.500.12657/90897.
9 European Commission. 2024. "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)." *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/reg/2024/1689.
10 Zhao, Zhang and Geng. 2024. "Deep Multimodal Data Fusion." *ACM Comput. Surv.* 56 (9). https://doi.org/10.1145/3649447.
11 Kim et al. 2024. "Temporal Graph Networks for Graph Anomaly Detection in Financial Networks." https://arxiv.org/abs/2404.00060.
12 Han et al. 2024. "MT²AD: Multi-Layer Temporal Transaction Anomaly Detection in Ethereum Networks with GNNAD: Multi-Layer Temporal Transaction Anomaly Detection in Ethereum Networks with GNN." *Complex & Intelligent Systems* 10 (1): 613–26. https://doi.org/10.1007/s40747-023-01126-z.
13 Boniol, Liu, Huang, Palpanas and Paparrizos. 2024. "Dive into Time-Series Anomaly Detection: A Decade Review." https://arxiv.org/abs/2412.20512.
14 Burg, van der Gerrit and Williams. 2022. "An Evaluation of Change Point Detection Algorithms." https://arxiv.org/abs/2003.06222.
15 O'Brien et al. 2023. "EWSmethods: An R Package to Forecast Tipping Points at the Community Level Using Early Warning Signals, Resilience Measures, and Machine Learning Models." *Ecography* 2023 (10): e06674. https://doi.org/10.1111/ecog.06674.
16 Ekle and Eberle. 2024. "Anomaly Detection in Dynamic Graphs: A Comprehensive Survey." *ACM Transactions on Knowledge Discovery from Data* 18 (8): 1–44. https://doi.org/10.1145/3669906
17 Garcia et al. 2025. "Concept Drift Adaptation in Text Stream Mining Settings: A Systematic Review." *ACM Trans. Intell. Syst. Technol.* 16 (2). https://doi.org/10.1145/3704922.
18 Lu et al. 2018. "Learning Under Concept Drift: A Review." *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/tkde.2018.2876857.
19 Hinder, Vaquet and Hammer. 2024a. "One or Two Things We Know about Concept Drift – a Survey on Monitoring in Evolving Environments. Part a: Detecting Concept Drift." *Front. Artif. Intell.* 7. https://doi.org/10.3389/frai.2024.1330257.
20 – – –. 2024b. "One or Two Things We Know about Concept Drift – a Survey on Monitoring in Evolving Environments. Part b: Locating and Explaining Concept Drift." *Front. Artif. Intell.* 7. https://doi.org/10.3389/frai.2024.1330258.
21 Swiss Federal Intelligence Service. 2024. "Cognitive Biases." Handbook.
22 Rastogi et al. 2022. "Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making." *Proc. ACM Hum.-Comput. Interact.* 6 (CSCW1). https://doi.org/10.1145/3512930.
23 Bertrand, Belloum, Eagan and Maxwell. 2022. "How Cognitive Biases Affect XAI-Assisted Decision-Making: A Systematic Review." In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 78–91. AIES '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3514094.3534164.
24 Hagendorff and Fabi. 2024. "Why We Need Biased AI: How Including Cognitive Biases Can Enhance AI Systems." *Journal of Experimental & Theoretical Artificial Intelligence* 36 (8): 1885–98. https://doi.org/10.1080/0952813X.2023.2178517.
25 Heck. 2024. "What about the Data? A Mapping Study on Data Engineering for AI Systems." In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, 43–52. CAIN '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3644815.3644954.
26 Reis and Housley. 2022. *Fundamentals of Data Engineering*. O'Reilly Media.
27 Swiss Confederation. 2021. "The Security Policy of Switzerland: Report of the Federal Council." 21.070.

https://www.sepos.admin.ch/en/security-policy-report-2021.

28 Meral and Meral. 2017. "The role of social media in Arab Spring." *Electronic Journal of New Media* 5 (January): 26–34. https://doi.org/10.17932/IAU.EJNM.25480200.2021/ejnm_v5i1003.

29 Shamsaddini et al. 2023. "Early Warning Signals of Social Instabilities in Twitter Data." https://arxiv.org/abs/2303.05401.

30 Oliveira et al. 2024. "Graph Language Model (GLM): A New Graph-Based Approach to Detect Social Instabilities." https://arxiv.org/abs/2403.17816.

31 Schvitz et al. 2022. "The Global Risk Index 2022: Revised Data and Methods." European Union. https://dx.doi.org/10.2760/041759.

32 Rød, Gåsste and Hegre. 2024. "A Review and Comparison of Conflict Early Warning Systems." *International Journal of Forecasting* 40 (1): 96–112. https://doi.org/10.1016/j.ijforecast.2023.01.001.

33 Deng, Jinxian et al. 2023. "Causalized Convergent Cross-Mapping and Its Approximate Equivalence with Directed Information in Causality Analysis." *PNAS Nexus* 3 (1): pgad422. https://doi.org/10.1093/pnasnexus/pgad422.

34 Meier. 2025. "Threats and Opportunities in AI-Generated Images for Armed Forces." https://arxiv.org/abs/2503.24095.

35 Gazel-Anthoine. 2022. "Deep Learning-Based Damage Detection for Monitoring of Armed Conflict Using Open-Access Satellite Imagery." Master's thesis, ETHZ.

36 Podlesnik and Mihelič. 2024. "Relationship Between Threat Modelling, Cyber Threat Intelligence, and Cyber Resilience: A Systematic Literature Review." *Varstvoslovje Journal of Criminal Justice and Security* 26: 1–13. http://www.dlib.si/?URN=URN:NBN:SI:doc-1XWT8G3X

37 Ma et al. 2025. "TIMFuser: A Multi-Granular Fusion Framework for Cyber Threat Intelligence." *Computers & Security* 148: 104141. https://doi.org/10.1016/j.cose.2024.104141.

38 Guo et al. 2023. "A Framework for Threat Intelligence Extraction and Fusion." *Computers & Security* 132 (1). /https://doi.org/10.1016/j.cose.2023.103371.

39 Haass. 2022. "Cyber Threat Intelligence and Machine Learning." In *2022 Fourth International Conference on Transdisciplinary AI (TransAI)*, 156–59. https://doi.org/10.1109/TransAI54797.2022.00033.

40 Parrish, Corns and Schreiner. 2025. "Analyzing Identification Friend or Foe (IFF) of Armored Vehicles: A Comparative Approach with Transfer Learning and Pre-Trained Models." *Industrial and Systems Engineering Review* 12 (1): 53–69. https://doi.org/10.37266/ISER.2025v12i1.pp53-69.

41 Glüge, Nyfeler, Aghaebrahimian, Ramagnano and Schüpbach. 2024. "Robust Low-Cost Drone Detection and Classification Using Convolutional Neural Networks in Low SNR Environments." *IEEE Journal of Radio Frequency Identification* 8: 821–30. https://doi.org/10.1109/JRFID.2024.3487303.

42 Summers. 2025. "DSI Insights: Polizeiarbeit mit KI – Herausforderung für die Justiz." https://www.inside-it.ch/dsi-insights-polizeiarbeit-mit-ki-herausforderung-fuer-die-justiz-20250415.

43 NIST AI 600-1. 2024. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST Trustworthy and Responsible AI". https://doi.org/10.6028/NIST.AI.600-1.

## Appendix A: ISO Standards

The documentation and management of data should include the following topics, processes and controls, among others[6,7]:

- **Data for the development and improvement of AI systems:** The organization shall define, document and implement data management processes for the development of AI systems.
- **Procurement of data:** The organization must define and document details about the sourcing and selection of data used in AI systems. In particular, the date the data was last updated or changed (e.g., date in metadata); for ML, the categories of data (e.g., training, validation, test, and production data); categories of data (e.g., as defined in ISO/IEC 19944-1); procedures for labeling the data; intended use of the data. Procurement processes need to be adapted to manage the procurement of data, not just systems and services.
- **Quality of data for AI systems:** The organization must define and document data quality requirements and ensure that the data used to develop and operate the AI system meets these requirements. Data quality should be reviewed on an ongoing basis so that the AI-generated models can be continuously

checked (security, bias, data poisoning, interpretability, effectiveness, robustness, etc.).

- **Data provenance:** The organization must establish and document a process for recording the provenance of the data used in its AI systems over the lifecycles of the data and the AI system. This applies in particular to the storage and disposal of data, potential problems and biases.
- **Data preparation:** The organization must define and document its criteria for selecting data preparations and the data preparation methods to be used. Targeted:
  - Required data and data sets are identified, sampled and procured.
  - The training data and, if applicable, the validation data are prepared, formatted and made available to the ML models.
  - Test data is prepared for testing or validation.
  - Data is prepared for manual analysis in order to gain a better understanding and support the AI data and model development processes.
  - Automated processes, if available, for extracting, transforming and loading the data are identified.
  - Any recording and use of personal information in the data is done in accordance with applicable laws and legal requirements.
  - Artifacts (e.g. metadata) for traceability, documentation, maintenance of data and automated process, including configuration management, are created.
  - The data is decommissioned in good time.
  - Multimodal data is managed.

## Appendix B

The NIST AI 600-1[43] identifies the following risks related to the use of genAI:

1. **CBRN Information or Capabilities:** Eased access to or synthesis of materially nefarious information or design capabilities related to chemical, biological, radiological and nuclear (CBRN) weapons or other dangerous materials or agents.
2. **Confabulation:** The production of confidently stated but erroneous or false content (known colloquially as "hallucinations" or "fabrications") by which users may be misled or deceived.
3. **Dangerous, Violent, or Hateful Content:** Eased production of and access to violent, inciting, radicalizing, or threatening content as well as recommendations to carry out self-harm or conduct illegal activities. Includes difficulty controlling public exposure to hateful and disparaging or stereotyping content.
4. **Data Privacy:** Impacts due to leakage and unauthorized use, disclosure, or de-anonymization of biometric, health, location, or other personally identifiable information or sensitive data.
5. **Environmental Impacts:** Impacts due to high compute resource utilization in training or operating genAI models, and related outcomes that may adversely impact ecosystems.
6. **Harmful Bias or Homogenization:** Amplification and exacerbation of historical, societal, and systemic biases; performance disparities between subgroups or languages, possibly due to non-representative training data, that result in discrimination, amplification of biases, or incorrect presumptions about performance; undesired homogeneity that skews system or model outputs, which may be erroneous, lead to ill-founded decision-making, or amplify harmful biases.
7. **Human-AI Configuration:** Arrangements of or interactions between a human and an AI system which can result in the human inappropriately anthropomorphizing genAI systems or experiencing algorithmic aversion, automation bias, over-reliance, or emotional entanglement with genAI systems.
8. **Information Integrity:** Lowered barrier to entry to generate and support the exchange and consumption of content which may not distinguish fact from opinion or fiction or acknowledge uncertainties, or could be leveraged for large-scale dis- and mis-information campaigns.
9. **Information Security:** Lowered barriers for offensive cyber capabilities, including via automated discovery and exploitation of vulnerabilities to ease hacking, malware, phishing, offensive cyber operations, or other cyberattacks; increased attack surface for targeted cyberattacks, which may compromise a system's availability or the confidentiality or integrity of training data, code, or model weights.
10. **Intellectual Property:** Eased production or replication of alleged copyrighted, trademarked, or licensed content without authorization (possibly in

situations which do not fall under fair use); eased exposure of trade secrets; or plagiarism or illegal replication.

11. **Obscene, Degrading, and/or Abusive Content:** Eased production of and access to obscene, degrading, and/or abusive imagery which can cause harm, including synthetic child sexual abuse material (CSAM), and nonconsensual intimate images (NCII) of adults.
12. **Value Chain and Component Integration:** Non-transparent or untraceable integration of upstream third-party components, including data that has been improperly obtained or not processed and cleaned due to increased automation from genAI; improper supplier vetting across the AI lifecycle; or other issues that diminish transparency or accountability for downstream users.

## Endnotes

1 https://acleddata.com/conflict-index/
2 https://www.swiss-ai.org/