MONITORING ARTIFICIAL INTELLIGENCE (AI):
SEIZING OPPORTUNITIES, MINIMISING RISKS

FX Powered by Reatch

POUR DEMAIN

Monitoring artificial intelligence: seizing opportunities, minimizing risks

Summary

This whitepaper presents a proposal for monitoring artificial intelligence (AI) in Switzerland. The proposal was developed during the parliamentary summer session 2022 in a workshop that facilitated an open exchange between members of the National Council, civil servants, researchers in relevant fields as well as business representatives. This workshop was organized in collaboration between the Franxini Project initiated by the scientific think tank Reatch, and Pour Demain, an AI-focused think tank.

Applications of AI are being used in more and more areas of life and will have a massive impact on society in the coming decades. On the one hand, they have the potential to significantly strengthen the common good, for instance by improving medical diagnoses. On the other hand, AI systems can cause material and/or immaterial damage to property or people.

The federal government can better understand the opportunities and risks of AI if it follows developments in the AI sector directly. To enable this, we recommend rapidly launching AI monitoring pilot projects to systematically capture the current capabilities and impacts of AI applications, thereby ensuring a timely and relevant evidence base for policy decisions.

The AI monitoring should achieve four **fundamental goals**:

1. **Identify opportunities**: Areas which can benefit from future AI applications become apparent.
2. **Classify risks**: Security risks posed by AI applications get detected.
3. **Enabling transparency**: It is revealed which everyday decisions are made by AI systems and not by people.
4. **Creating a basis for decision-making**: Society and politics receive a better basis for deciding on the use of AI.

The specific recommendations for action are:

- **Step-by-step approach via pilot projects** to assess and optimize the concrete benefits of AI for society and politics on the basis of initial results.
- **Pilot project 1 "Critical Infrastructures"**: An initial pilot project should be implemented in a safety-critical area such as healthcare, since the common good is particularly affected here.
- **Pilot project 2 "Meta-monitoring":** This project clarifies in which areas future monitoring is particularly necessary and which methods are suitable for this.

For the successful implementation of AI monitoring, a clear distribution of roles among the actors is crucial:

- **Administration**: Should receive sovereignty over monitoring projects to ensure that monitoring results are directly relevant to the political decision-making process.
- **Science**: Due to their expertise in AI and related fields, Swiss universities can implement monitoring projects effectively and in an interdisciplinary manner on behalf of the administration.
- **Private Companies**: Are set to benefit from monitoring through increased demand for AI applications in areas that have expanded application potential. At the same time, enterprises will be made more aware of risks, in a constellation similar to how the National Center for Cybersecurity (NCSC) raises awareness for cybersecurity risks.

1. Introduction

Opportunities and challenges of artificial intelligence

The capabilities of artificial intelligence (AI) are increasing rapidly. Both the Foundation for Technology Assessment TA Swiss[1] and the Geneva Science and Diplomacy Anticipator (GESDA)[2], which is co-funded by the federal government, emphasize that AI applications will radically change society in the coming decades. An interdepartmental federal working group reaches the same conclusion.[3]

AI is described by many experts as a fundamental technology: Like electricity 100 years ago, AI applications are currently beginning to permeate business and private life, and at a rapid pace. In fact, since the "machine learning" breakthrough a decade ago, there are more and more economic applications of AI systems. The potential for societal value is great; for example, in healthcare, image recognition algorithms can already detect cancers earlier than human doctors, thereby saving lives.

At the same time, however, the use of AI systems also harbours the potential for damage through accidents or misuse. Today, for example, it is not known how error-prone AI systems are in critical infrastructures such as power supply, or when application documents are evaluated by an AI system instead of a person.[4] In domains as dynamic as AI, it is correspondingly challenging for legislators to keep track of all relevant developments. Given the pace of progress of AI in recent years, this problem will become much more acute in the short to medium term.

Monitoring artificial intelligence

Under these circumstances, it seems very reasonable for the federal government to directly monitor the development of the AI sector to understand its impact on society. Continuous monitoring would help the federal government to better understand the opportunities and risks of deployed AI, and thus to fulfil its main objective: the promotion of the common good.

At the invitation of the Franxini Project and Pour Demain, experts from academia, the National Council, the administration and the private sector came together in a workshop during the 2022 summer session. The central question for the participants: Under what circumstances is AI monitoring necessary, and what are the roles of politics, science and society? Based on the findings of the workshop, we present here a concept for successful and agile monitoring, as well as two pilot projects. These are included as key recommendations since targeted monitoring pilot projects can create an up-to-date evidence base, which provides orientation for political decisions on AI.

Artificial intelligence: blessing or curse?
On the one hand, AI systems are able to solve problems that until recently were considered unsolvable for computers. Thus, they are both symbol of and catalyst for today's technological progress. On the other hand, these systems function in a highly complex and partly opaque way. Until we have a detailed understanding of how these systems work, we should carefully consider and observe their use in safety-critical areas.

[1] TA SWISS: Wenn Algorithmen für uns entscheiden: Chancen und Risiken der KI (2020)
[2] GESDA: Next-level AI
[3] SBFI: Herausforderungen der Künstlichen Intelligenz (2019)
[4] Franxini-Projekt: Algorithmen mit Vorurteilen (2022)

2. Goals

The goal of AI monitoring is to understand the current capabilities of AI systems, as well as to track their effects on society in a transparent and timely manner. This goal includes several aspects:

- **Classify risks:** Monitoring is a means to identify the damage potential of currently – or soon-to-be – deployed AI systems. It enables an analysis of the safety of any given system, for instance with regard to the observance of fundamental rights such as privacy or the principle of equality.
- **Identify opportunities:** Monitoring provides information on the economic and social applications of AI in different sectors (e.g. health, power supply, agriculture). This also renders visible whether there are sectors relevant for the common good in which AI could be applied productively, but where AI systems are currently not in widespread use. Particularly, the identification of barriers to AI deployment could also help to further establish Switzerland as a pioneer in the research and innovation of AI.
- **Enable transparency:** Monitoring renders visible where the population is (often unwittingly) interacting with AI. This makes it easier to evaluate where the AI infringes on personal rights and makes it easier to address doubts and concerns in the population. From a societal perspective, being able to explain decision processes is central, which presents a particular challenge with AI systems.
- **Create a basis for decision-making**: Monitoring provides the public and the federal government with constantly updated and accessible information that serves as a basis for cost-benefit considerations regarding the use of AI systems.

These goals are in line with the AI guidelines adopted by the Federal Council in November 2020.[5]

What do we mean when we talk about AI systems?
AI is an umbrella term for various technologies and methods to which scientific fields such as computer science, mathematics and neuroscience contribute. Machine learning (ML) applications form an important subfield of AI. Many technological breakthroughs of the last 10 years were enabled by ML. These breakthroughs depend on the availability of three factors: big data, high computing power, and algorithmic innovation. At its core, ML is capable of recognizing patterns and associations in large amounts of data using mathematical rules. ML applications often make use of artificial neural networks and in this sense are loosely modelled on the brain (this mechanism is also known as deep learning). Together with a linked learning algorithm, these networks can map complex input-output relationships, which are used, for example, in image or speech recognition.

[5] Der Bundesrat: Leitlinien "Künstliche Intelligenz" für die Bundesverwaltung verabschiedet (2020)

3. Recipes for monitoring AI successfully

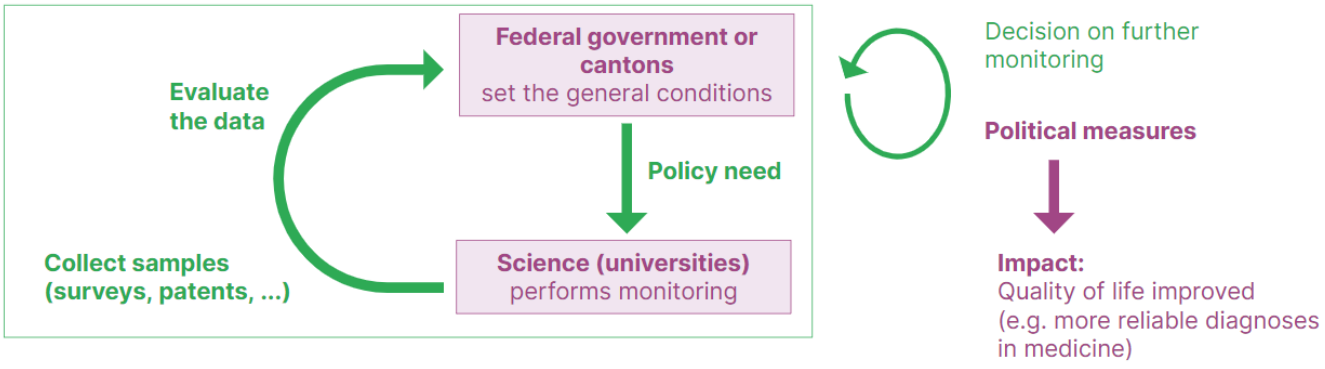
Successful AI monitoring needs to consider the following points regarding intentions, implementation, and actors.

Intentions (based on basic goals):
Monitoring...

1. identifies sector-specific potential for improvement and innovations that could be enabled through the deployment of AI systems.
2. places a strong emphasis on critical infrastructure, whose failure would have particularly severe consequences for people and their livelihoods.
3. observes in detail those sectors in which fundamental rights, such as the right to privacy, could potentially be violated.
4. creates a basis for further activities, such as conformity tests and labels, which put the quality/safety of the systems under the microscope.
5. builds on existing data (e.g., NCSC cyber incidents[6], CNAI list of AI systems deployed in the federal government[7], Swiss AI Research Overview Platform[8], GESDA Science Breakthrough Radar[9], Swiss AI Report[10]).

Actors and implementation:

1. Monitoring will be commissioned by a responsible administrative body so that the conclusions can be fed into the political process and so the federal government can develop its AI competencies.
2. Specialist organizations such as universities and universities of applied sciences, which have the interdisciplinary know-how, are included as implementation partners. It is appropriate that the core infrastructure (such as aggregated datasets, search tools or indexes) is hosted by the federal government (while the federal government ensures that the core infrastructure is accessible to external parties where necessary).
3. In the form of a dialogue process, monitoring takes into account all stakeholders (e.g. civil society) as stakeholders.
4. Monitoring is adapted to existing regulations in the relevant sector and involves the corresponding sector-specific organizations (such as Swissgrid for electricity).



[6] NCSC: Aktuelle Vorfälle
[7] CNAI: Projektdatenbank
[8] SAIROP: Swiss AI Research Overview Platform
[9] GESDA: Advanced Artificial Intelligence
[10] Mindfire: Swiss AI Report 2022

4. Step-by-step approach using pilot projects

Pilot projects provide helpful insights for the implementation of AI monitoring. They allow a flexible, iterative approach that enables adjustments based on initial results, while at the same time already informing decision-makers about concrete opportunities and risks of deployed AI systems.

The participants of the workshop specifically suggested pilot projects on critical infrastructures and a so-called meta-monitoring. The focus on critical infrastructures such as hospitals or energy service providers is explained by the fact that these have a very direct impact on the common good of society. The pilot project on meta-monitoring is intended to identify further useful AI monitoring projects and to clarify their methodological implementation in more detail.

Pilot project 1: Critical infrastructure

AI monitoring of critical infrastructure improves federal and public understanding of the capabilities and impacts of AI systems deployed in safety-critical areas. The following case study is included to address key points:

- **Where**: The pilot will focus on a specific area of critical infrastructure, such as hospitals, as part of the healthcare sector. In addition, the focus will be on one specific AI application.

  Concrete example: Samples are collected on the use of image recognition algorithms in Swiss hospitals. This provides information on the frequency with which these AI systems are used in the healthcare sector. Depending on the result of the measurement, it could be interesting to obtain further information or to draw a new sample from another specific AI system, or from another field of application.

- **How**: Pilot projects use existing data. This includes, for example, from the federal side, the list compiled by the Competence Center for AI at the Federal Statistics Office of AI systems used in the administration, or the NCSC statistics on cyber incidents. Cantonal data are also included.

  Concrete example: For image recognition in hospitals, other publicly available data can be used, such as publications, patents or the public communication of hospitals, companies and universities. Alternatively, data is also collected by means of surveys at hospitals.

- **Who:** The administration should commission the monitoring. This has the advantage that questions with political and social relevance are addressed from the outset.

  Concrete example: In the case of image recognition in hospitals, the Federal Office for Public Health or the cantonal health directors' conference could give the order. The implementation would be done by an existing team in the administration, which already carries out similar work. Alternatively, there is expertise at universities that could implement this pilot project (e.g., the Distributed Information Systems Lab at EPFL, the Center for AI at ZHAW, or the ETH AI Center). It should be ensured that all relevant stakeholders are involved in the project (e.g., the business community and civil society). The success of the monitoring project can be evaluated by an external evaluation provider.

Pilot project 2: Optimize use and implementation of monitoring (meta-monitoring)

The meta-monitoring project clarifies in which subject areas - beyond critical infrastructures - future monitoring is particularly necessary and which methods are suitable for monitoring. Thereby, it creates the basis for future pilot projects and makes it possible to strategically plan the institutionalization of monitoring. The following questions are the focus here:

- **Where**: With regard to which topics is monitoring particularly necessary? Possible areas include self-driving cars, surveillance systems, social media and weapons. Meta-monitoring should assess both the societal impact of future AI developments in a sector, and the expected benefits of monitoring to identify the most impactful future projects.

  Special attention must also be paid to neglected topics (while, for example, the consequences of AI systems for assessing social media content are relatively widely discussed, the use of AI in energy providers is hardly present in the public debate). Meta-monitoring finalizes a catalogue of criteria in a first step and then, in a second step, applies it to prioritize key topics and areas.

- **How**: Meta-monitoring develops methods and metrics that can later be used to monitor applications. The aim here is to identify methods that can assess the AI ecosystem in Switzerland particularly effectively and efficiently. These include surveys of IT managers, media research, systematic information searches on the web (scraping), and possibly network scans.

  In addition, the meta-monitoring takes into account successful strategies from other areas (e.g., early detection of drought in agriculture) and from international AI projects (e.g., AI Index).

- **Who**: Ideally, meta-monitoring should be initiated by a government agency so that decision-makers can benefit directly from the findings. This could be done, for example, through a mandate from the federal government's Competence Network for Artificial Intelligence (CNAI).

  It would also be conceivable, however, to initially carry out the pilot project without government involvement and to bring the administration on board on the basis of convincing results. In this case, implementation would be carried out by an external expert organization (university institute, NGO, consulting office). In addition to such classic implementation assignments, more creative formats could also be considered, such as project competitions and hackathons (collaborative processing in a diverse group over a short period of time).

5. Swiss policy developments

With the introduction of AI guidelines for the federal administration[11], there exists a basis for AI monitoring. In this context, the relevant actors include, among others, the Competence Network for Artificial Intelligence (CNAI), which is currently focusing on terminology and AI applications in the administration under the leadership of the Federal Statistical Office (FSO). Further thematic work is carried out by the Federal Office of Communications (OFCOM), which monitors the implementation of AI guidelines in the administration and organizes the annual Plateforme Tripartite. Overall, however, the majority of the federal government's AI activities have so far been limited to the administration.

In the first half of 2022, the first motions were submitted in the National Council for the introduction of AI monitoring. The focus of monitoring should not rest exclusively on the federal administration, as the majority of the Swiss AI landscape lies outside the administration. If the manageable additional resources are made available, experts from the scientific community are ready to implement first pilot projects.

6. Distribution of roles among actors

Administration

While a big part of the AI monitoring implementation can be outsourced, the federal government (or the cantons) should have sovereignty over the projects, which they are set to commission. This ensures that the monitoring:

- is based on the needs of policymakers, i.e., is tailored to and useful for political decision-making.
- contributes to a broad technical understanding of this key technology that has a positive impact outside of monitoring (e.g., closer exchange with research/industry, more expertise in the government's deployment of AI).
- does not lead to dependency on external actors.

Science

Swiss universities enjoy a high degree of trust among the population and, with their AI expertise, are very well positioned to effectively carry out monitoring projects (as implementation partners). In particular, technical expertise on the functionality of AI systems is required, as well as the competence to provide concrete policy assessments from the collected data. However, the goals and scope of monitoring are not defined by the scientific institution itself, but are specified by policymakers as a framework condition.

One possible format to support monitoring would be an interdisciplinary exchange between experts from different disciplines. This creates a feedback loop, and problems identified in this space can then be addressed in separate research projects.

Private sector

For the business community, monitoring is attractive for identifying areas where growth potential for AI applications exists. Thus, monitoring generates a demand for this expanding technology. At the same time, the economy will be made more aware of risks, similar to today's cyber security awareness, to which the National Center for Cyber Security contributes integrally.

Monitoring projects analyse products that are commercially available. It is not the intention to gain privileged access to company information beyond what normal customers can expect.

[11] Bundesrat: Leitlinien Künstliche Intelligenz für die Bundesverwaltung (2020)

7. Challenges

The following challenges must be taken into account when implementing monitoring projects:

- **Delimitation**: Until now, no clear definition of artificial intelligence exists in Switzerland, and a delimitation from other automation processes is often difficult in practice. A clear description of the applications to be investigated in advance is necessary. Definitions of the OECD or the EU could be applied.

- **Technology neutrality**: It remains open politically whether it is justified to conduct technology-specific monitoring for artificial intelligence. The focus on a single technology carries the risk that other relevant developments will not appear on the radar, which in turn makes forward-looking planning more difficult. At the same time, the assessments of scientists, TA Swiss and GESDA, among others, as well as the exponentially growing global investment volumes in AI over the past 10 years speak in favour of a focus on AI.

  Thanks to focus and specialization, AI-specific monitoring can deliver better results than general technology monitoring. Of course, activities in the field of AI do not exclude early detection in other technologies, but should be understood as complementary.

- **Legal framework**: The existing legal mandates of the individual administrative bodies provide the starting point for monitoring. This already provides sufficient scope for pilot projects. For broader monitoring, legislative initiatives may be necessary.

Conclusion

If the federal government implements AI monitoring projects and takes measurements of the AI sector, this brings key benefits: policymakers gain a clearer understanding of the strengths and weaknesses of Swiss AI applications. In addition, early signs of technological challenges are provided, which enable policymakers to act with foresight.

As an implementation partner, the scientific community has a central role to play, especially due to its broad AI expertise and know-how in developing methods and metrics that can be used in monitoring applications.

Through AI monitoring, the private sector gains insight and access to new sales markets. In addition, there is a growing awareness of the potential risks that companies face (e.g., in the case of faulty algorithms in central production processes).

Finally, society benefits from increased transparency in the use of AI and from wider deployment of beneficial applications that, in the best case, for instance in healthcare, can save lives.

Outlook

The Franxini Project and Pour Demain will continue the dialogue between representatives from national politics, science and administration in the second half of 2022 to further elaborate the details of the pilot projects. If you are interested in collaborating, please contact the authors.

Authors

Lead authors

David Marti
Program Manager AI
Pour Demain
david.marti@pourdemain.ch

Luca Schaufelberger
Co-head Franxini Project
luca.schaufelberger@reatch.ch

Contributing authors:

Patrick Stadler, Coordinator Pour Demain
Amir Mikail, Team Member Franxini-Projekt
Lucius Arn, Event Manager Franxini-Projekt
Max Lauber, Team Member Pour Demain

About Pour Demain

Pandemics, climate change, new technologies: we address these key challenges of our time with pragmatic initiatives - for the benefit of a secure future for our children, grandchildren and their descendants.

Pour Demain is a non-profit association that promotes effective and science-based national policies: with the greatest possible positive impact on present and future generations, in Switzerland and beyond. Pour Demain: today for tomorrow!

Further information: www.pourdemain.ch
Contact: info@pourdemain.ch

About the Franxini Project

The Franxini project builds bridges between science and politics by promoting the social and political participation of scientists, as well as mutual understanding and trust between politics and science. Through direct contact with decision-makers, researchers have the opportunity to understand what kind of scientific work will benefit them most. Politicians get to know the function and mode of operation of scientific work better through personal interaction with researchers.

The Franxini Project was initiated by the scientific think tank "Reatch! Research. Think. Change." and is supported by Mercator Foundation, Gebert Rüf Foundation, cogito foundation and other partners.

Further information: www.franxini-projekt.ch
Contact: franxini@reatch.ch