**Vice-Rectorate Research
and Innovation**

https://research.unibe.ch

b
UNIVERSITÄT
BERN

UniBE Guidelines for the responsible use of artificial intelligence (AI) in research and research on AI

Contents

1. Introduction

Artificial intelligence (AI) is the intelligence of machines or software, as opposed to the intelligence of humans or animals. It is also the field of study in computer science that develops and studies intelligent machines. "AI" sometimes also refers to the machines themselves.

AI technology is widely used throughout industry, government and science. Some high-profile applications are: advanced web search engines (e.g., Google Search), recommendation systems (used by YouTube, Amazon, and Netflix), understanding human speech (such as Siri and Alexa), self-driving cars (e.g., Waymo), generative or creative tools (ChatGPT and AI art), and competing at the highest level in strategic games (such as chess and Go).

The aim of these guidelines is to promote research in trustworthy AI. Trustworthy AI is characterised by three components: (a) it should be legitimate and thus comply with all applicable laws and regulations; (b) it should be ethical and thus guarantee compliance with ethical principles and values, and c) it should be robust, both technically and socially, as AI systems can cause unintended harm, even with good intentions. Each component in itself is necessary, but not sufficient to achieve the goal of trustworthy AI. Ideally, all three components work together harmoniously and overlap in their functioning. If, in practice, there is tensions between these components, a way to reconcile them must be found.

The following principles should guide all members of the University of Bern in their research and work on and with artificial intelligence. They are in line with the Swiss Federal Council's “Guidelines ‘Artificial Intelligence’ for the Confederation” from 2020 and the European Commission's 2021 proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.

The guidelines should be food for reflection before and during your work on and with AI. Research groups are invited to discuss how they could be implemented/respected in a given research project.

The UniBE's Digitalization Strategy aims, beside other things, at transforming research digitally (action field 5). Research in digital transformation has to be supported by both, new services and guidance. The present guidelines present a strong contribution to meet these goals.

They have been prepared by members of the KILOF (*Künstliche Intelligenz in Lehre, Organisation und Forschung [Artificial Intelligence in Teaching, Organisation and Research]*) Community of the University of Bern, in particular by the Focus Group Artificial Intelligence, with the support of the Research Management Office of the Vice-Rectorate Research and Innovation. The Digitization Commission (*DigiK*) of the University of Bern recommended the adoption of these guidelines.

These guidelines are reviewed yearly.

2. Guiding principles

2.1 Framework conditions for the development and application of AI.

The University of Bern creates the best possible conditions for research on and with AI so that we can keep up in this field internationally and in order to foster excellent research.

2.2 Putting people at the centre.

In the development and use of AI, human dignity and well-being as well as the common good should be paramount. Particular importance is attached to the protection of fundamental rights.

2.3 Values of the University of Bern.

Research with and on AI shall not conflict with the University's values such as equal opportunities, quality, sustainability, and internationalisation, but should rather help promote them.

Public research funding resources must be used efficiently and transparently.[1] Economic or ecological gains in effectiveness must be strived for.

2.4 Principle of sustainability.

AI-based systems, including algorithms and the needed infrastructure, should be developed, deployed and managed in accordance with the principle of sustainability to ensure long-term positive impacts on society and the environment on a global scale. As far as possible, it is recommended to consume energy in an economical and rational way (e.g., to limit energy consuming calculations to the necessary ones or to perform calculations over times when the electricity network is not overloaded, such as overnight or on weekends).

2.5 Principle of inclusivity and diversity.

AI-based research should be designed to be free from discrimination. AI algorithms must – whenever possible – ensure that discrimination and bias are not perpetuated via their use. Principles of inclusivity and diversity should be considered at various steps of AI development and deployment, including data collection, AI training, and evaluation of results. When elimination of bias is not possible, for example because of lack of available data, this should be explicitly mentioned via transparent reporting practices on the algorithms and datasets used. The principles of inclusivity and diversity should also be applied in the developer and researcher teams when selecting team members and collaborators. Likewise, access to AI-driven services must be equal and without discrimination.

2.6 Harm prevention and security.

AI systems should not cause or exacerbate harm or otherwise have a negative impact on humans. This includes the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure. They must be technically robust, and it must be ensured that they are not vulnerable to misuse or abuse. Particular attention should be paid to situations

[1] See e.g. SNSF regulations for funded projects: Open Access to Publications (snf.ch) and the UniBE's Open Access Policy: Service: Open Access - Universitätsbibliothek Bern UB (unibe.ch).

where AI systems can cause or exacerbate negative impacts due to unequal distribution of power or information, such as between employers and employees, businesses and consumers, or governments and citizens. Harm prevention also includes consideration of the natural environment and all living beings.

AI's capabilities must be limited as much as necessary to ensure security and avoid damage.

Undesirable use (e.g. military, unpeaceful, endangering fundamental rights, endangering democracy and market stability or ecologically dangerous) of the research should be excluded as far as possible, or made less likely or less attractive.

2.7 Principle of necessity.

When working with AI, researchers should evaluate the necessity of the AI model, especially if a risk for negative impact of any kind exists. Likewise, the suitability of the AI application for the intended use must be ascertained.

2.8 The principle of explainability.

AI-based research should be designed to be verifiable and accountable.

Processes must be transparent and the capabilities and purpose of AI systems must be openly communicated and decisions must be – to the greatest possible extent – explainable to those directly and indirectly affected by them. Without this information, a decision cannot be properly challenged.

An explanation of why a model produced a particular outcome or decision (and what combination of input factors led to it) is not always possible. These cases are called "black box" algorithms and require special attention. In these circumstances, other explanatory measures may be necessary (e.g. traceability, verifiability and transparent communication about the capabilities of the system) as long as the system as a whole respects fundamental rights. The degree to which explicability is necessary depends on the context and scope of the consequences of an erroneous or otherwise inaccurate result.

2.9 Principle of validity.

The principle of validity especially applies in research using AI as current AI algorithms have limited generalisability: AI models should only be used for contexts for which they have been developed.

2.10 Transparency, comprehensibility, and reproducibility.

The inherent limits to the comprehensibility, reproducibility and predictability of the "behaviour" of the AI solution – as well as how the experimental design or the project account for the associated fuzziness/quality deficiencies – must be made transparent.

At a minimum, the answer to the following questions must also be made transparent:

- Privacy: What information about oneself or one's connections must a person reveal to others, under what conditions and with what safeguards? What things can one keep to oneself and not be forced to reveal to others?

- Accuracy: Who is responsible for the authenticity, fidelity and accuracy of information? Who is responsible for errors in information and how can the aggrieved party be compensated?
- Ownership: Who owns the information? What are the just and fair prices for its exchange? Who owns the channels through which information is transmitted? How should access to this scarce resource be allocated?
- Accessibility: What information does a person or organisation have the right or privilege to receive, under what conditions and with what guarantees?

2.11 The principle of acceptability.

The principle of acceptability should apply to any human interaction with an AI system: Humans must be enabled to adopt an approving or disapproving position vis-à-vis the AI system. Therefore, it must at all times be transparent for the human user whether he or she is in an interaction situation with the AI system or not. It must also be clear when and how he or she can leave this interaction situation. The consequences of leaving must equally be clear. Furthermore, it must be possible to explain the processes within the system to the human being if necessary.

2.12 Digital sovereignty/possibility of not to use AI.

Research promoting AI solutions that contribute (possibly indirectly) to lock-in effects and other forms of dependence on certain technologies or that create an inevitability of surveillance that can no longer be circumvented in everyday life should be avoided.

2.13 Responsibility.

In order to be able to clarify responsibilities in the event of damage, an accident or a violation of the law, liability must be clearly defined when using AI. It must not be possible to delegate responsibility to machines; responsibility rests with humans.

2.14 Data protection, security and privacy.

If sensitive data is processed with AI, researchers must be aware that the data may be reused by the provider for their own purposes. Researchers should be aware of this lack of control and the potential risks when processing unpublished or other sensitive research data with AI.

In principle, no or only completely anonymized personal data should be processed with AI. The direct or indirect (re-)identification must be prevented when using anonymised personal data.

When personal data are processed with AI, compliance with relevant data protection laws must be ensured. Consent must have been given by the data subjects to their use. When obtaining consent, the information sheet and the presentation of the type and objectives of the AI solution must be formulated in a sufficiently comprehensible manner.

Sufficient security measures must be taken for archiving (IT security, preventive measures against sabotage, data theft, etc.).

2.15 Documentation.

At a minimum, the characterisation of the data (type, scope, origin), the documentation of the software and training methods, the specification of the hardware, the documentation of relevant guidelines, etc. and, in the case of so-called “high-risk systems”, also the automatically generated documentation of the calculation process should be documented.

3. Support and advice at the University of Bern

If you have questions, the following contact points at the University of Bern will be able to help you:

- General questions: Research Management Office, Vice-Rectorate Research and Innovation
- Support and training in data science, machine learning, AI and research IT-related matters: Data Science Lab
- Open Science: Open Science Team at the University Library
- Sustainability: Office for Sustainable Development
- Data protection: Legal Services Office
- Harm prevention: Risk Management Office
- Military or dual-use research: Export control contact point

4. Contact information

To contact the authors of these guidelines, please address yourself to:

- Sarah Schlunegger, Research Management Office, sarah.schlunegger@unibe.ch
- Dr. Hannah Brodersen, Research Management Office, hannah.brodersen@unibe.ch

5. Resources

These guidelines are based on the following sources:

- Marie-Christin Barton & Jens Pöppelbuß: Prinzipien für die ethische Nutzung künstlicher Intelligenz. HMD Praxis der Wirtschaftsinformatik. 10 March 2022.
- Bundesministerium für Wirtschaft und Klimaschutz (D): Ethische Leitlinien für Künstliche Intelligenz. 1 September 2021.
- Bundesrat: Leitlinien „Künstliche Intelligenz" für den Bund. Orientierungsrahmen für den Umgang mit künstlicher Intelligenz in der Bundesverwaltung. 25 November 2020.
- Anne A. H. de Hond et al.: Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review | npj Digital Medicine (nature.com). January 2022.
- Deutsches Forschungszentrum für Künstliche Intelligenz: Handreichung zum Thema Ethik am DFKI.
- European Commission: Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. 21 April 2021.
- Max Haarich: Ethik im Zeitalter der Künstlichen Intelligenz. Ein Überblick internationaler Vorgehensweisen und Handlungsempfehlungen. 18 January 2019.
- Anna Jobin et al.: The global landscape of AI ethics guidelines. Nature machine intelligence 1: 389-399. September 2019.
- Seong Ho Park *et al.*: Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. Korean Journal of Radiology 22(3): 442-453. March 2022.
- Unabhängige Hochrangige Expertengruppe für künstliche Intelligenz. Ethik-Leitlinien für eine vertrauenswürdige KI. April 2019.

- Maarten van Smeden *et al.*: Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease | European Heart Journal | Oxford Academic (oup.com). May 2022.
- Zentrum verantwortungsbewusste Digitalisierung: Zur forschungsethischen Begutachtung von KI-Forschungsprojekten. Handreichung zur Unterstützung der Arbeit von Ethikkommissionen an Hochschulen. October 2022.