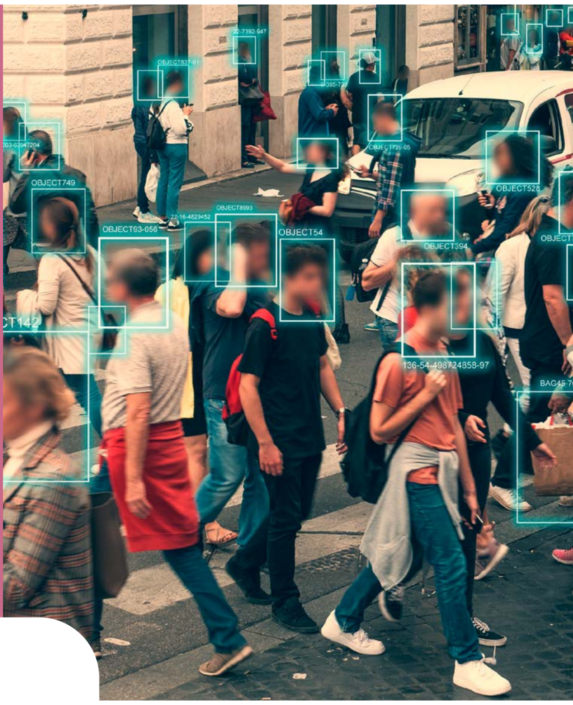
Lernende Systeme
GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

Criticality of AI systems in their respective application contexts

White Paper by Jessica Heesen,
Jörn Müller-Quade, Stefan Wrobel et al.
Working Group IT Security, Privacy, Legal and Ethical Framework
Working Group Technological Enablers and Data Science



Executive Summary

Artificial Intelligence improves processes and business models and helps to ensure the future viability of the economy and society. For example, AI systems can improve diagnoses and treatments in medicine, optimise route planning in mobility or enable a more precise match between needs and offers. At the same time, however, they also harbour risks and thus bring the issue of Artificial Intelligence (AI) and trustworthiness into play. After all, the risks that the use of AI systems can entail and the damage that can occur during these operations are diverse and often difficult to assess.

In order to bring safe and reliable applications into use, the European Commission, in its proposal on the regulation of AI systems in April 2021, advocated regulating AI systems according to their risk potential and classifying them into four risk levels, from minimal risk (no need for regulation) to unacceptable risk (prohibition of use). In the white paper „Criticality of AI systems in their respective application contexts – A necessary but not sufficient building block for trustworthiness", experts from the working groups IT Security, Privacy, Legal and Ethical Framework as well as Technological Enablers and Data Science deal with selected contents of this EU proposal: namely, which criteria can be used to determine in which cases the use of AI systems should be regulated from the outset and when this is not necessary. With this central question as a guiding principle, they want to give a „good" answer as to how AI quality can be ensured through regulation and how over-regulation can be avoided while at the same time promoting innovation, thus ultimately ensuring the protection of the subject. In doing so, they enrich the current political debate with further perspectives that take up the topic of criticality of AI systems with regard to their trustworthiness.

Ex-ante and ex-post measures

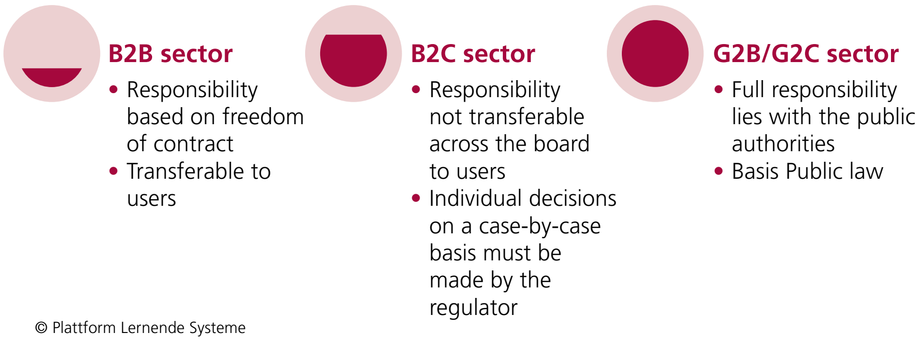
In principle, the authors recommend supplementing the criticality assessment of AI systems in advance (ex ante) with measures that take effect in retrospect (chapter 1). In terms of transparency and traceability as well as liability and compensation, these ex-post measures can only be promising if the groundwork for them has been established in advance. Effective, low-threshold and timely complaint and consumer protection regimes are listed as an example. This is because these strengthen and secure the data sovereignty of the data subject beyond consent at the beginning of the data processing process. The consideration of criticality should therefore take place in the sense of danger prevention rather than danger aversion, since risks are not only to be evaluated in purely technical terms, but also in socio-technical terms.

Responsibility, responsibility chains and liability

When considering the criticality of AI systems in certain application contexts, different dimensions of responsibility and concern should also be taken into account (chapter 2.3). This division of responsibility of the risk via liability regulations forms an important building block towards trustworthy AI systems. The aim of (corporate) liability is to share the technical risk in the best possible way among the different actors according to aspects of fairness and functionality.

The central question is who can better assess, better bear and, if necessary, eliminate a risk. For this purpose, the authors present a model for sharing responsibility in the case of damage caused by AI systems, which distinguishes on the one hand according to the addressee and on the other hand, if necessary, according to the criticality of the respective area of application. Thus, for liability issues in the B2B sector, the conditions of contract law should generally apply (except in socially critical areas), while in the B2C sector decisions should be made depending on the criticality. In the B2B sector, the user should in principle be responsible for the results delivered by AI systems in accordance with contract law. For applications in the public domain, the public authority should bear full responsibility for discriminatory or harmful consequences under public law. These considerations on responsibility and also liability should be considered not only nationally, but also throughout Europe.

Figure 1: Responsibility chains by area



B2B sector
- Responsibility based on freedom of contract
- Transferable to users

B2C sector
- Responsibility not transferable across the board to users
- Individual decisions on a case-by-case basis must be made by the regulator

G2B/G2C sector
- Full responsibility lies with the public authorities
- Basis Public law

© Plattform Lernende Systeme

Dimensions for evaluation: possibilities for control and decision-making

In addition to regulation based on criticality in advance, the creation of consumer protection regimes for possible cases of damage, the division of responsibility for risk via liability regulations, the authors also recommend focusing more on the control and decision-making possibilities of the users of AI systems when assessing criticality (chapter 3). For this purpose, they suggest dividing the criteria for assessing the criticality of an AI system in a specific application context into two dimensions. Namely, whether the recommendations or decisions of an AI system endanger human lives or legal assets such as the environment, and how much room for manoeuvre is left to humans in the selection and use of the application, for example to switch off certain functions: The higher the extent of the possible harm to human life and other high legal interests and the smaller the scope of the individual's options for action, the higher the criticality – and the need for regulation derived from this – and vice versa.

Figure 2: Criticality of AI systems against the background of their respective application context

Criticality

Higher criticality
- Possibility and extent of harm to human life and other high legal interests: **high**
- Extent of possibilities for action by individuals: **low**

Lower criticality
- Possibility and extent of harm to human life and other high legal interests: **low**
- Extent of possibilities for action by individuals: **high**

© Plattform Lernende Systeme

The explanations and starting points mentioned above are supplemented and discussed in more detail by experts from the Plattform Lernende Systeme in short interviews with further aspects from their respective areas of expertise (chapter 4). This thematic arc illustrates, on the one hand, the breadth and, on the other hand, the complexity of further possible answers to the question of the regulation of AI systems as a function of criticality: How are criticality and regulation connected? How is the concept of criticality to be understood? Or: What ethical demands arise with regard to responsibility or for technical action? to name just a few of the questions. This question-answer format clearly reflects the complexity of the criticality assessment of AI systems and at the same time shows that the topic cannot yet be considered comprehensively and conclusively.

The following figure provides a summary of the experts' suggestions for adjustments to make the EU regulatory proposal more precise and concrete:

Figure 3: Proposed adjustments to the regulatory proposal of the European Commission

EU regulation	Proposals for adjustment
Possibility and extent of harm to human life and other legal interests	
Severity of harm • Extent to which an AI system has caused harm or the risk of harm • Extent of the impact of the harm • Likelihood that the AI system will harm a high number of persons to be harmed • Likelihood that an AI system will cause more than one of the specifically defined harms	**Problem:** difficult to quantify → Consideration of Society as a whole and individual harms → Consideration of material and immaterial damages
Likelihood of harm • Number of people using the AI system	+ • Number of uses of the AI system • Persistence of the threat situation • Controllability of the threat situation (degree of connectivity)
Extent of freedom of action of individuals	
• Extent of dependence on potential Persons affected by a result • Extent of vulnerability of the potentially affected individuals vis-à-vis the user(s) of an AI system • Extent of reversibility of the outcome produced by an AI system • Availability and effectiveness of remedies (in Union law and in Member State law) • Extent to which existing Union legislation is able to prevent/minimize the risks posed by the AI system	+ • Opt-out/configuration possibilities of the users • Decision-making possibilities of the users • Market structure/plurality of service offering

© Plattform Lernende Systeme

The present analysis shows that the concept of risk or criticality assessment chosen by the European Commission to ensure the quality of AI systems provides a good orientation function for evaluation and regulation – if the concept is supplemented by some further classification criteria. According to the experts, it is essential to expand the concept and to define and specify further criteria in order to be able to measure and assess the risk potential of an AI system. Above all, they emphasise, it is crucial that the regulation of AI systems is always seen against its background of the respective application context. This is also in order to create a balance between openness to innovation on the one hand and protection of the subject on the other.

In addition to regulation based on criticality in advance, the creation of consumer protection regimes for possible cases of damage, the sharing of responsibility for risk via liability regulations, the control and decision-making possibilities of the users of AI systems in the evaluation of criticality also form an important – even if not sufficient – building block towards trustworthy AI systems. At the same time, the authors see a need for further research to overcome weaknesses (complexity reduction, lack of predictability and foreseeability of damage), which must also be taken into account in the future.

Imprint
Editor: Lernende Systeme – Germany's Platform for Artificial Intelligence | Managing Office | c/o acatech | Karolinenplatz 4 |
D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de |
Follow us on Twitter: @LernendeSysteme | Status: November 2021 | Photo credit: DedMityay/shutterstock/Title

This executive summary is based on the white paper *Criticality of AI systems in their respective application contexts – A necessary but not sufficient building block for trustworthiness,* Munich, 2021. The authors are members of the working group IT Security, Privacy, Legal and Ethical Framework and of working group Technological Enablers and Data Science of Plattform Lernende Systeme. The original version of this publication is available at: https://www.plattform-lernende-systeme.de/publikationen.html

SPONSORED BY THE

Federal Ministry
of Education
and Research

acatech
NATIONAL ACADEMY OF
SCIENCE AND ENGINEERING