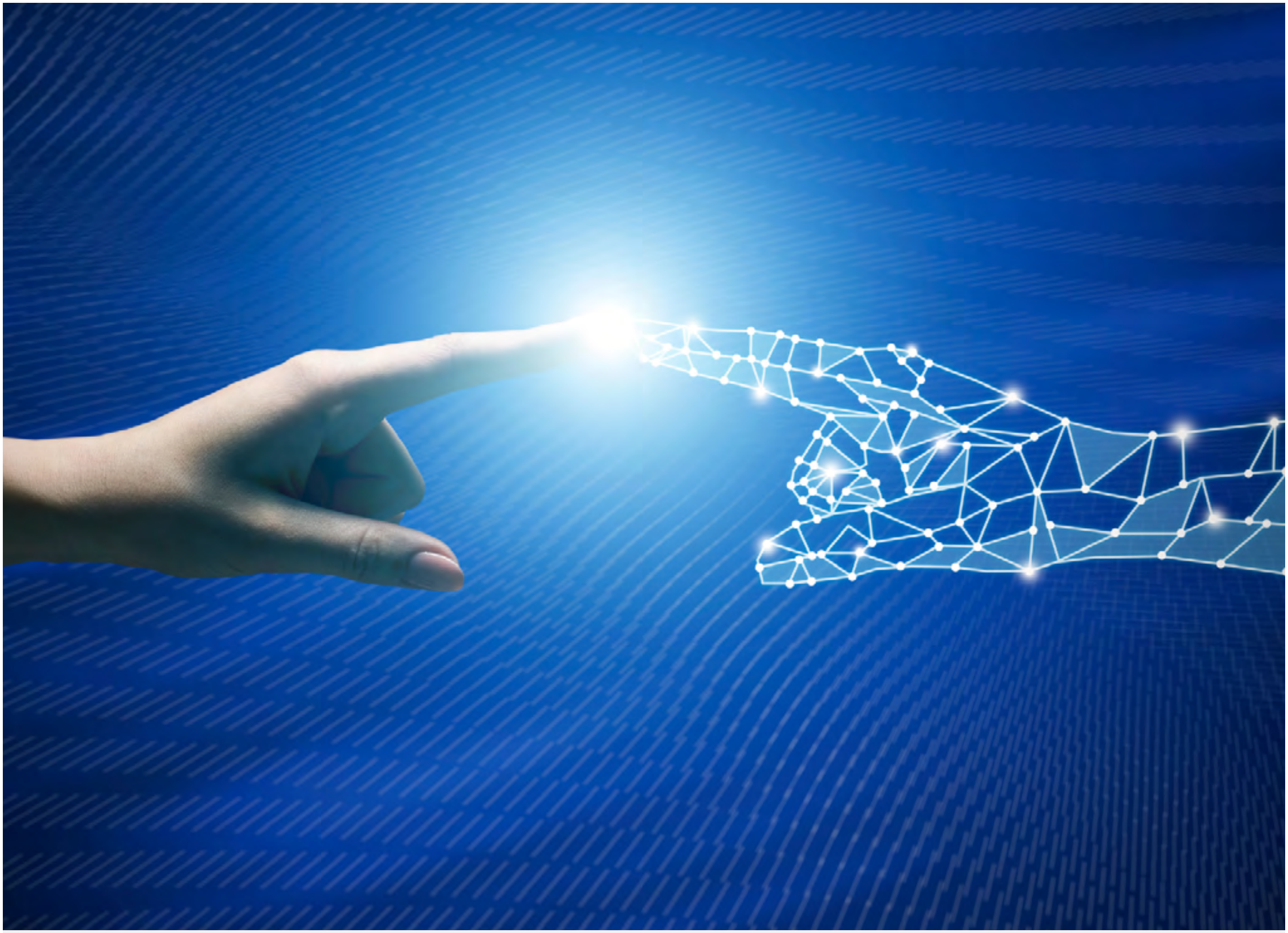

Trustworthy AI:
A Guide for Risk Assessment and Compliance

JPM | PARTNERS

Trustworthy AI: A Guide for Risk Assessment and Compliance

Publisher: JPM | Partners
Delta House, 8a Vladimira Popovića street
www.jpm.law
Authors: Ivan Milošević, Partner, Katarina Rosić, Senior Associate, Prof. Gojko GruborPhD
Design and prepress: JPM | Partners

Disclaimer:
The sole purpose of this publication is to provide information about specific topics.
It makes no claims to completeness and does not constitute legal advice.
The information it contains is no substitute for specific legal advice.
If you have any queries regarding the issues raised or other legal topics, please get in touch with your usual contact at JPM & Partners.

Introduction

In our previous practice, we have assisted clients in the course of implementation of AI systems ("AIS"), i.e., assessment of the impact of risks of AIS, especially in the medical and research sector. While working on these projects, we came to the conclusion that many companies overlook obligations or are not even aware of risks for safety, health and fundamental human rights which are tied to the implementation and use of high-risk AIS and can result in significant compliance risks under the applicable legal framework.

As the use of artificial intelligence grows, organizations must adopt robust frameworks to ensure AIS are trustworthy, effective, and aligned with human values.

In Serbia, the Government adopted Ethical Guidelines for the Development, Implementation and Use of Reliable and Responsible Artificial Intelligence (the Guidelines), as a part of the Strategy for Development of Artificial Intelligence in the Republic of Serbia for Period 2020-2025. The Guidelines are obligatory for public entities and are recommended to be used by other entities that are developing or using AIS.

In order to address potential risks associated with the use of AIS we developed an AI Risk Assessment Model ("Model"). This Model is designed to help deployers identify information and documents required for high-risk AIS risk assessment.

Model

The Model contains a self-assessment questionnaire which provides support to organizations developing, placing on the market, putting into service and/or using AIS to assess whether AIS meets conditions for reliability, safety, responsibility, confidentiality and ethical acceptability.

Based on the results of the questionnaire, deployers shall define and implement adequate technical, organizational and personnel measures, proportional to the assessed risk, to mitigate unacceptable risk factors to an acceptable level, i.e., to ensure that the risks do not harm the rights and freedoms of natural persons.

Moreover, in addition to assessing the overall relative[1] risks and compliance of high-risk AIS, this questionnaire also contains a set of questions aimed at assessing the risk of harm to personal data processed by an AIS.

Typically, the provider (developer, vendor) assesses risk during the design and development phases, while deployers assess relative risk during the implementation phase. The questionnaire is suitable for use in all socio-economic areas and sufficiently adaptable for specific areas and sectors.

[1] Risk is always relative due to zero (0) risk is impossible and maximal (1) risk is certain event - not risk at all.

Self-evaluation of Ethical Principles, Reliable, Safe, and Responsible AIS

An AIS is ethically acceptable, technically reliable, safe, and "accountable" when it is aligned with requirements such as the Law on Information Security, the Law on Personal Data Protection, and ethical principles and values for the use of AIS explained in the Guidelines. Most of the questions should be addressed by the deployer of the AIS. However, the deployer should also obtain certain information from the AIS provider.

The role and place of AIS in processing activities

The first set of questions relates to the introduction of AIS into processing activities. Within this section, deployers need to disclose whether they received information and/or documentation about the AIS from the provider:

i. confirming whether users and data processing scenarios, complete range of variability and specifics of personal data processing, in the context of relevant business sector, have been applied in the training of AIS;

ii. information and/or documents about capabilities and functionalities of AIS, all or some relevant usage scenarios of AIS, operational infrastructure and configuration contributing to reliable and responsible use of AIS, limitations of AIS, segments within which AIS is not designed for use, accuracy and proper functioning of AIS and a description of the extent to which such results can be expected for general use in scenarios that were not initially considered, limits to which further development of AIS (self-learning) is expected without direct human influence.

Human Agency and Oversight

The AIS should serve as a reliable support in decision-making processes and must be subject to continuous human monitoring and oversight.

One of the seven non-binding ethical principles for AI [2] , which are intended to help ensure that AI is trustworthy and ethically sound is – human agency and oversight. According to the guidelines, human agency and oversight means that AIS are developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans.

These sections of the questionnaire deal with:

i. effects AIS can have on human behavior, i.e., the effect of AIS that are guiding, influencing or supporting humans in decision-making processes; and
ii. level of human oversight over the AIS.

These questions examine:

a. the influence of the AIS on decision-making, with or without human intervention;
b. risk analysis and assessment for the rights and freedoms;
c. the perceptions and expectations of those developing, maintaining, and using the AIS;
d. the individuals affected by the AIS;
e. their trust and acceptance of the AIS;
f. the (in)dependence of the AIS in decision-making

[2] 2019 Ethics guidelines for trustworthy AI developed by the independent AI HLEG appointed by the EuropeanCommission.

Additionally, this also helps to self-assess necessary oversight measures through lifecycle governance mechanisms such as:

i. Human-in-the-loop (HITL) approach – human intervention is enabled at all stages of AIS decision-making;
ii. Human-on-the-loop (HOTL) approach – human intervention is enabled during development and monitoring of the AIS operation; or
iii. Human-in-command (HIC) approach – capability to oversee the overall activity of the AIS (including economic, social, legal, and ethical impacts) and the ability to decide when and how to use the AIS, including not using it in certain situations.

It is necessary to determine whether the AIS is self-learning or allows certain levels of human intervention, and if so, to what extent, in which operational phases, and under what relevant circumstances.

High-risk AIS should comply with the requirements prescribed byRegulation (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) - “AI Act”, inter alia, they shall be designed and developed in such a way that they can be effectively overseen by natural persons during the period in which they are in use (Article 14 of the AI Act).

Human oversight shall aim to prevent or minimize the risks to health, safety and fundamental rights that may emerge when a high-risk AIS is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse.

Technical robustness and safety of AIS

A fundamental requirement for achieving trustworthy AIS is its (i) ability to deliver services that can justifiably be trusted and (ii) robustness when facing changes.

Technical robustness requires that AIS is developed with proactive risk assessment from the earliest design phase, behave reliably and as intended, and minimize potential unintentional and unpredictable harm.

This section of the questionnaire addresses four key areas:

i. protection against threats and misuse;
ii. security;
iii. accuracy and precision;
iv. reliability, fallback mechanisms, and reproducibility of outputs. Information/documentation confirming these areas shall be requested from providers.

In accordance with Article 15 of the AI Act, high-risk AIS shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness and cybersecurity, and that they perform consistently in those respects throughout their lifecycle.

Privacy, personal data protection, and data governance

Privacy and data protection are tightly linked to the principle of harm prevention—specifically, risk to the data and rights and freedoms of individuals.

This section of the questionnaire is intended to help to self-assess the impact of the AIS on privacy and data protection, which are fundamental rights that are closely related to the fundamental right to the integrity of the person, which covers the respect for a person's mental and physical integrity.

Preventing privacy violations and protecting individuals' data rights requires proper data governance, which includes data quality, confidentiality, integrity, and availability; data relevance based on the context in which the AIS is developed and applied; access control protocols; the AIS's capability to process data in a privacy-preserving way that respects data subject rights.

Transparency of AIS

One of the most important elements of trustworthy AIS is transparency, which is defined as:

- the extent to which the AIS discloses processes/parameters related to its functionality, or
- the property that enables understanding of how and why the AIS made a specific decision or acted in a certain way, given its environment and operational context.

Transparency is important for at least three reasons: 1) autonomous AIS can make errors or cause harm—transparency helps explain why; 2) AIS must be understandable and trustworthy to users, as much as possible; and 3) without transparency, accountability is not possible.

A key feature of intelligent systems is autonomy. An autonomous AIS is defined as a “system capable of making decisions in response to certain input data or stimuli, with varying degrees of human intervention depending on the level of system autonomy.”

AIS demonstrates some level of autonomy in interacting with its environment. While robots have long been used in various business systems, future application of AIS in fields like healthcare, pharmaceuticals, law, and personal data processing will depend on traceable decision-making processes, interpretable outputs, user interaction methods, and the presentation of results to end users.

To build trustworthy and reliable AIS, transparency must include three elements:

i. traceability – the ability to track AIS behavior;
ii. explainability – the ability to understand how AIS functions; and
iii. open communication about the AIS's limitations.

Moreover, one of the requirements established in Article 13 of the AI Act for high-risk AIS is that they must be designed and developed to ensure a sufficient level of transparency. This transparency should enable deployers to interpret the system's output and use it appropriately. Additionally, high-risk AIS must be accompanied by instructions for use—provided in an appropriate digital format or other means—that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers.

Diversity, non-discrimination, and fairness

In order to achieve trustworthy AIS, it is important to enable diversity in its outputs throughout the entire lifecycle of the AIS.

AIS may have shortcomings due to poor or incomplete data governance, leading to unintentional (in)direct biases and discrimination of certain social groups—potentially worsening biases or marginalizing vulnerable populations.

AIS must be user-focused and built in a way that allows everyone to access AIS-based products or services, regardless of age, gender, ability, or characteristics. Accessibility is especially important for persons with disabilities. Article 10 of the AI Act emphasizes the need for data governance to prevent biases that could negatively impact fundamental rights or lead to discrimination.

Accessibility and universal design

AIS should be user-centric and designed to enable everyone to use AIS-based products or services, particularly in business-to-consumer (B2C) sectors.

Accessibility for persons with disabilities—present across all social groups—is of particular importance.

AIS providers should move beyond the "one-size-fits-all" approach and consider principles of universal design to accommodate the widest possible range of users, aligned with relevant compliance standards.
This promotes equal access and participation in human-AIS interactions across all aspects of life.

Stakeholder participation

To build trustworthy AIS, it is recommended to consult stakeholders who may be directly or indirectly affected by the system throughout its lifecycle.

It is useful to seek continuous feedback, even post-deployment, and to establish long-term mechanisms for stakeholder engagement—such as providing information, consultations, and participation in AIS implementation processes.

Social and environmental well-being

In line with the principles of fairness and harm prevention, it is important to evaluate the AIS's impact on society and the environment throughout its lifecycle. AIS's presence in education, work, or entertainment can affect individual behavior, social relationships and social dynamics.

AIS impacts must be continuously monitored and reevaluated. Support should be given to research and development of AIS that positively contributes to environmental sustainability.

This section of the questionnaire helps to self-assess the impact of the AIS on the environment. AIS must work in the most environmentally friendly way possible. Additionally, it also helps self-assess the impact of AIS and its use in a working environment on employees, their relationship with coworkers and employers, as well as the impact of an AIS from a societal perspective, taking into account its effect on institutions, democracy and society at large.

Accountability

Accountability is closely tied to proper planning, monitoring, and risk management during AIS deployment.
It involves procedures for determining liability and remedying any harm resulting from AIS use.

While others in the AIS development chain may also bear responsibility, designers and developers hold a high level of accountability during design, development, training, and decision-making stages. This doesn't mean they shouldn't work in multidisciplinary teams—on the contrary. Human logic and judgment remain crucial throughout the AIS lifecycle, even as the system is expected to make objective and logical decisions.

Humans define the algorithms, select training data, determine successful outcomes, and ultimately decide when and how to use the AIS. Thus, all involved parties and companies funding AIS development are responsible at any stage of the lifecycle for evaluating its impact on the environment in which it operates—whether risks are introduced intentionally or accidentally.

This section of the questionnaire helps to self-assess the existing or necessary level that would be required for an evaluation of the AIS by internal and external auditors.

Stakeholder participation

To build trustworthy AIS, it is recommended to consult stakeholders who may be directly or indirectly affected by the system throughout its lifecycle.

It is useful to seek continuous feedback, even post-deployment, and to establish long-term mechanisms for stakeholder engagement—such as providing information, consultations, and participation in AIS implementation processes.

Social and environmental well-being

In line with the principles of fairness and harm prevention, it is important to evaluate the AIS's impact on society and the environment throughout its lifecycle. AIS's presence in education, work, or entertainment can affect individual behavior, social relationships and social dynamics.

AIS impacts must be continuously monitored and reevaluated. Support should be given to research and development of AIS that positively contributes to environmental sustainability.

This section of the questionnaire helps to self-assess the impact of the AIS on the environment. AIS must work in the most environmentally friendly way possible. Additionally, it also helps self-assess the impact of AIS and its use in a working environment on employees, their relationship with coworkers and employers, as well as the impact of an AIS from a societal perspective, taking into account its effect on institutions, democracy and society at large.

Conclusion

The implementation of AIS raises significant challenges and responsibilities in terms of ensuring that the AIS is secured, trustworthy, effective and aligned with human values.

Our AI Risk Assessment Model is a tool designed to help identify, evaluate, and mitigate overall ethical risks, as well as data protection risks throughout the AIS lifecycle.

Moreover, it helps individuals/organizations to identify areas for improvement and encourages them to undertake necessary measures to overcome identified risk factors. By filling in the questionnaire, one gets an insight into the already established measures and identifies the measures that should be implemented for the purpose of building a reliable, safe, confidential and ethically acceptable AIS.

Authors

Ivan Milošević

Partner

E:ivan.milosevic@jpm.law

Katarina Rosić

Senior Associate

E:katarina.rosic@jpm.law

Prof. Gojko Grubor Ph.D.

JPM | PARTNERS
8a Vladimira Popovića,
Delta House
11070 Belgrade, Serbia
T:+ 381/11/207-6850
E: office@jpm.law

www.jpm.law