BDI

Statement

High-Level Expert Group on Artificial Intelligence – Draft Ethics Guidelines for Trustworthy AI

Bundesverband der Deutschen Industrie e.V.

Stand: 01.02.2019

Introduction: Rationale and Foresight of the Guidelines

Rationale: The BDI welcomes the approach of the High-Level Expert Group (HLEG) to define ethic guidelines for trustworthy AI. To foster support for the guidelines and facilitate their impact in common practice of AI development, deployment and use, the need/reasons for AI guidelines should be better motivated. The HLEG should carve out that the development of AI based on European ethical and societal values can build trust in artificial intelligence, facilitate a broader uptake of AI and can serve as a unique selling proposition for AI "made in Europe".

Aim/Scope: It is positive that the guidelines aim to support companies in implementing ethical principles and values in the development and application of AI. Although further adjustment is needed. The guidelines do not yet take sufficient account of the fact that the ethical boundary conditions of AI systems differ considerably depending on the field of application. Particularly for industrial applications ethical questions often play only a minor or very context-specific role. An insufficient differentiation regarding the criticality of AI applications may lead to undifferentiated red lines, which unnecessarily restricts Europe's competitiveness.

Endorsement mechanism: BDI appreciates the target of putting a method in place to enable all stakeholders to formally endorse and sign up to the guidelines on a voluntary basis. This can support transparency for users and foster trust. Due to the diversity of AI applications, however, a "one size fits all"-solution seems not to be feasible. The suggested technical (and non-technical) methods (as mentioned in Chapter 2) and detailed checklists (as described in Chapter 3) are too specific and not applicable for all use cases. Thus, a "holistic" framework including high-level principles, seem more appropriate for the commitment/endorsement proposal of the HLEG. The methods and checklists should not be included into the formal endorsement but added as best-practice examples instead. Furthermore, the endorsement process raises questions regarding its practicality. The guidelines do not make clear what consequences an endorsement has on the signatories, e.g. if signatories thereby fall under specific external governance or auditing.

Glossary: The definition of AI seems not to be appropriate for a political debate, since it does not differ between "weak" and "strong" AI. The choice of words such as "perceiving" or "reasoning" falsely suggests that AI systems are human-like, fully autonomous, thus potentially prejudiced, acting systems. Furthermore, the formulation "… and deciding the best action(s) to take to achieve the goal" should be replaced by "… and providing predictions or results, which might be implemented automatically where appropriate and traceable."Moreover, a definition of ethics should be included in the glossary, since the perception of ethics differs in different cultures.

Chapter I: Respecting Fundamental Rights, Principles and Values - Ethical Purpose

Definition of Ethical Purpose: BDI values the approach to derive responsible/trustworthy AI development, deployment and use of AI based on fundamental rights, ethical principles and values. But the guidelines should carve out more precisely the EU understanding of terms like "ethical purpose" and "wellbeing and the common good" (p. 5). The wording may indicate, that any AI should serve a higher ethical purpose or even the sole purpose of AI should be ethics.

Equality, non-discrimination and solidarity including the rights of persons belonging to minorities: AI development, deployment and use should adhere to the fundamental right of equal treatment, as set out e.g. in Chapter III of the EU Charter of Fundamental Rights. In this context, the document states that "equality of human beings goes beyond non-discrimination" (p. 7). To avoid misconceptions, conflicts with the fundamental right of individual freedom, or the notion of a 'levelling down', this statement should be revised to capture more precisely what it is supposed to mean in the context of AI.

Informed consent: The concept of an „informed consent" needs clarification. It remains unclear whether an "informed consent" is identical with the consent under the GDPR. It should be made clear that only the term used in the GDPR applies. Moreover, the GDPR offers further legal basis for data processing, such as processing necessary for the performance of a contract or for legitimate interest. Therefore, the notion of informed consent is given too much prominence and misleads it to be the only and best requirement to preserve autonomy.

Comments regarding the Ethical Principles and Values: The BDI agrees with the five principles and correlated values in general. However, some adjustments are necessary:

The Principle of Beneficence "Do Good"

Taking into account that AI should create added value to different stakeholders, economic interests have to be considered as legitimate interests of a company in order to promote economic growth. Adhering to principles such as traceability, transparency and self-determination might come to an economic cost. Therefore, the principle of beneficence should be complemented by a notion of proportionality.

The Principle of Non-maleficence: "Do no Harm"

BDI agrees, that AI systems should protect the dignity, integrity, liberty, privacy, safety, and security of human beings. AI applications are being developed by humans, and while it is understandable and correct, that the expectations are higher than towards humans, it must be understood, that high efforts and continuing improvements are necessary to reduce potential risks.

Use case specific, the necessary quality level and fault tolerance, fall back solutions in case of error, necessary effort for testing, monitoring/controlling have to be defined, always under consideration of the field of application, how autonomous the AI may act, the opportunities and possible risks, and which machine learning method is used.

Moreover, BDI supports the aim to avoid discrimination, manipulation or negative profiling in general. However, "negative profiling" is necessary in a certain context. Businesses must be able to segment customers and business partners based on certain predefined criteria to assess the creditworthiness of a person by using AI. This is also essential for insurances, financial institutions and e-commerce businesses.

The Principle of Autonomy: "Preserve Human Agency"

A general "right to decide to be subject to direct or indirect AI decision making" is impractical. A differentiation with regard to criticality and context is urgently necessary. According to the wording, every citizen could have the right to object to an uncritical use of AI in a longer process chain by which he or she is indirectly impacted (e.g. if an AI is used to calculate the time of garbage collection from a household). A general, non-context-specific right to opt-out would be highly impractical and would hinder the uptake of AI in administration and business processes.

It should also be considered that there are limits to the right to be subject to direct or indirect AI decision making. Suitable alternatives are not available in all cases. This applies in particular to the working environment mentioned in footnote 13. The formulation "…anyone using AI as part of his/her employment enjoys protection for maintaining their own decision making capabilities and is not constrained by the use of an AI system" should not be interpreted as an individual right to object to any AI implementation in the working environment. Today, AI is already an inherent part of the working environment for many professions (e.g. pilots are supported by AI in aircrafts). Employees should participate collectively in decisions around the implementation of AI systems in working environments through established bodies of representation.

The Principle of Justice: "Be Fair"

Instead of stressing "that AI systems must provide users with effective redress if harm occurs", the guidelines should emphasize that ultimately humans are responsible. Operators of AI should know and make clear who is responsible for which AI system or feature. As with other technologies and products, the people who design and deploy AI systems must be accountable for how their systems operate.

The Principle of Explicability: "Operate transparently"

BDI fully agrees with the AI HLEG, that explicability is a key success factor to increase the acceptance and trust in AI systems. For this, it is important to explain the function of AI in an understandable manner. Focusing on explaining the result, the base for decision making and the benefit of the system seems to be the key. In order to achieve a high explicability, non-AI specialists could be involved in the design process.However, transparency,

especially when using deep learning, still has its limits. But the limitations (e.g. accuracy, safety), the decision process, the algorithm and the defined quality criteria should be made transparent and documented within companies. Additionally, it needs to be clarified how potential neutral audits in critical contexts could be ensured, especially considering the limited pool of AI specialists.

Potential longer-term concerns: The probability of potential occurrences as mentioned by the HLEG are currently very low and well into the future. Therefore, we suggest focusing on realistic and existing challenges but remain attentive to future development of critical topics. However, Artificial Moral Agents (AMAs) should not per se pose a threat as long as these have been trained within a given and acceptable ethical framework. It is highly likely that AMAs being trained by re-enforcement principles (where the reward is adherence to the ethical principles) are near-future feasible (i.e., white swans). This is decidedly not a negative development and might be one of the few technology principles existing today that might actually work in terms of developing practical ethical AI.

Chapter II: Realising Trustworthy AI

Accountability: The HLEG rightly points out that the choice of accountability mechanisms is highly dependable on the use case, the field of application, the autonomy and many more factors. Regarding accountability, a highly differentiated approach should be targeted.

Data Governance: The statement that "the datasets gathered inevitably contain biases, and one has to be able to prune these away before engaging in training" depends on the aim of a given policy or algorithmic model. Pruning a model to make it fairer for one group may inevitably create biases and unfairness for another group, in particular if different groups have different descriptive distributions and base rates. It thus makes more sense to identify bias / unfairness with data that reflects the real, imperfect world and then correct post-processing for bias and unfairness (which often would be relevant for minority groups in a machine learning setting).

Design for all: The requirements for „design for all" are far too general and not applicable to all AI solutions, especially with industrial applications. For example, AI included into automated cars (e.g. level 3) could still not be used by people without driver license, thus a use of the service/product will not be available for all ages. Furthermore, it is highly likely that there will be AI-based products and services that appeal to particular groups rather than universally to all humans, e.g., gender specific apps, age specific apps (and combinations thereof).

Governance of AI Autonomy: As mentioned by the AI HLEG, the use cases and the fields of application differ, thus the impact (benefit and risk) also differs immensely. Due to this fact, BDI proposes to always review what level of autonomy in decision should be applied (AI only as a source of information, AI as an assistant with final decision by user or AI acts fully automated without human involvement). Furthermore, it is essential to also review the level of autonomy in learning (may the AI learn on the market (retraining possible), with limited parameters (no safety relevant parameters) or no learning/evolvement on the market possible). And as a third dimension, the level of risk should be consindered (e.g. which persons or laws could be harmed and how). Such a structure could be very helpful to take a more differentiated view of ethical issues.

Transparency: Different Levels of Transparency will be necessary for different use cases and groups. The right level of transparency or explainability is important to strengthen the user's trust in AI applications. However, higher transparency will be necessary for developers and operators to ensure quality monitoring and continuous improvement. Furthermore, for use cases with potentially higher risks, higher levels of transparency are necessary.

(Non-)technical methods: As mentioned above, the proposed methods are not applicable for all use cases. Thus, they should only be considered as best practice and should not be included into the formal endorsement.

Additionally, as a non-technical method, all AI systems should come with a clear description of their limits, including the areas they are intended for and those, they are not intended for, as well as description of input data that the system cannot properly cope with (e.g. a system tailored to autonomous car control might not properly cope with autonomous truck control due to different dimensions and requirements). Moreover, another method/paragraph should be added on school education, vocational training, required curricula & skills for the new and working generations.

Chapter III: Assessing Trustworthy AI

BDI welcomes the efforts of the HLEG to offer guidance to steer developers, deployers and other innovators toward ethical purpose and technical robustness. However, the list of the HLEG is very inconsistent. The questions vary in their granularity and do not differentiate between the AI methods being used. The questions are not suitable as practical assistance yet, since they lack technical details and specification (which is crucial for real guidance).

General Comments

Costs of implementing trustworthy AI: To implement AI successfully in Europe it must safeguarded that additional costs and bureaucracy are minimized, e.g. the additional auditing services and storage of logs would demand additional development effort, operational cost for storage, processing power, licenses, as well as man power to monitor and maintain the auditing the system. The costs can be a high burden, especially for small and medium-sized enterprises.

Alignment with existing processes: The specific recommendations and guidelines should be more aligned to existing processes for data protection, product safety and security. Aligning it would mean to make a fit/gap analysis with existing procedures and integrate it into them in a lean manner. Within this context, the guidelines should point out that AI is not a completely new technology. Many industry companies have long term experiences with AI and already established well-functioning processes to minimise the risks. This might also help to calibrate societal concerns, presenting AI not only as disruptive and transformational but also as an incrementally developing technology.

Fostering R&I: Fostering R&I on achieving Trustworthy AI in EU should get a prominent position in the document, at the forefront of activities, including practical test cases (e.g. sandboxes) for various verticals.

Support companies to operationalise the guidelines: Companies will be responsible for operationalising these guidelines. A path to establishing measures for these companies should be described in detail.

Impressum
Bundesverband der Deutschen Industrie e.V. (BDI)
Breite Straße 29, 10178 Berlin
www.bdi.eu
T: +49 30 2028-0

Ansprechpartner
Clemens Otte
Stellvertretender Abteilungsleiter
Telefon: 030-2028-1614
c.otte@bdi.eu

BDI-Dokumentennummer: D1016