White Paper

on

AI Ethics and Governance

"*Building a Connected, Intelligent and Ethical World*"

By

Prof. Dr. Christoph Lütge

Peter Löscher Professor and Chair of Business Ethics
TUM School of Governance
Director of the TUM Institute for Ethics in Artificial Intelligence

&

Research Team

March 2020

Executive Summary

In the second decade of the 21st century, artificial intelligence (AI) is already being used around the world and now shapes how societies and their institutions are maintained, organized and controlled. Ranging from face recognition use in London and Beijing to autonomous vehicles driving in San Francisco and Munich, and price prediction applications in financial markets, AI has become omnipresent in our everyday lives. Given its tremendous influence on society, politics, science and economics, the **interrelationship between ethics and AI** concerns enterprises, governments and individual consumers worldwide. Thus, the main challenges for developers of AI solutions and policy makers will be to (1) navigate the differing interests and beliefs regarding AI that are prevailing in our increasingly interconnected societies, (2) understand stakeholders' perceptions on AI and (3) develop principles for the mitigation of compliance and reputational risks.

Developing AI principles

In order to understand the challenges policy makers and corporate decision makers face when confronted with aspects of AI ethics and AI regulation, it is first important to have a structural understanding of the principles behind AI regulation. While stakeholders' perceptions on AI ethics vary and are still materializing in numerous international, European and national frameworks, the majority touch on the overarching principles of **beneficence, non-maleficence, autonomy, justice** and **explicability**. These principles form the basis for our comparative analysis.

Overall, on the international level, the *OECD Principles on Artificial Intelligence* are one of the most influential frameworks on AI, due to the multi-stakeholder approach and its adaption by the G20. Even so, many aspects of the Principles remain vague and require future exploration and explanation as organizations continue to navigate this topic.

Although there is an overwhelming consensus on AI ethics in general, the interpretation of AI principles is displayed on the international, European and national level. Many stakeholders have stressed the relevance of international frameworks such as the United Nations Sustainable Development Goals or the United Nations Guiding Principles on Business and Human Rights for managing AI. Constitutional norms pertaining to human dignity, liberty, non-discrimination and privacy, play another decisive role in understanding how to interpret the AI principles of autonomy, justice and non-maleficence in practice.

In addition to written law and formal rules, there are also informal rules and meta-ethical considerations that are often difficult for non-specialists to grasp. Despite their theoretical nature, meta-ethical considerations influence the practical narratives on AI ethics, such as in the case of autonomous driving. Moreover, the interpretation of principles such as autonomy or justice has its roots in the perceptions of powerful interest groups, including political parties and NGOs, and is embedded in different religious and ethical traditions and influenced by a society's historical experiences.

Finding Common Ground

Finding common ground on AI ethics requires integrating global AI principles into local perceptions and incorporating them into established legal traditions. Apart from mere compliance with already existing standards, organizations need to consider the dynamics between stakeholder perceptions and general motives.

As demonstrated, AI regulations can be generally subdivided into the five overarching principles of beneficence, non-maleficence, autonomy, justice and explicability. The interpretation of these more general principles, however, hinges on specific AI relevant guidelines and individual stakeholder interests. **Stakeholders vary, therefore, in how they interpret these global principles**.

In addition to the challenge of understanding, interpreting and communicating commitment to these principles, the problem arises of how these principles interact with one another. For example, justice and beneficence may come in conflict in terms of how minority interests are valued against the interests of the majority. While there is no automatic solution, companies can deal with these conflicts by focusing on the public discourse and looking for potential win-win situations in the further enhancement of AI capacities.

The interrelatedness of the Principles also appears in practical use cases, such as face recognition or predictive policing. Based on the analysis of current tendencies in legislation and the interests formulated by major stakeholder groups, we conclude that the use and collection of data must, in most cases, be based on consent and that companies developing invasive AI solutions must take care to clearly explain the measures taken to prevent cases of non-maleficence and injustice.

How are companies handling AI ethics and stakeholder involvement?

In order to get a holistic understanding of how companies are implementing these Principles in practice, we examined selected companies with different national backgrounds. We assessed them in terms of their handling of AI ethics and stakeholder involvement, prioritizing four aspects:

- Does the company define their AI principles?
- Does the company define processes and policies concerning AI ethical principles?
- Is the company reporting and explaining its AI solutions?
- Does the company participate in the social discourse and stakeholder dialogue?

Through comparing companies, **we were able to analyze the relative weaknesses of major AI users and developers in terms of AI governance**. We found that the differences among companies were quite pronounced irrespective of the business sector. Three companies stood out as role models. Google offers the most precise principles for AI ethics. Microsoft and Daimler explain the benefits of AI solutions in a very transparent way and have strong collaboration with governments and NGOs.

Conclusion

Based on legal ramifications, stakeholder expectations and on the company comparison, we derived conclusions concerning the overall governance on AI. The overall conclusion, we can draw from our research, is that there is still a wide gap between theoretical AI principles and practical action guiding regulation. Moreover, we found that embedding **ethics within AI governance structures is crucial for uniting the different stakeholder perceptions prevailing in our societies**. The definition of core values and the adaptation of already existent legal frameworks to various cases is an important aspect of implementing practical AI solutions.

Based on our work, we were able to identify following priorities:

1. Policy makers need to distinguish between *hard* and *soft* frameworks for regulating different AI use cases.
2. Policy makers and stakeholder groups should foster global collaboration on defining minimum standards for AI.

3. Policy makers and corporate decision makers need to pay heed to potential conflicts within different principles of AI ethics and how relevant stakeholder perceptions and preferences interact with these conflicts.

In the light of these major recommendations, we have also developed concrete recommendations related to the different principles of AI governance. For example, we propose that beneficence criterion should be adapted to distinct fields of application, as it has different implications for the health sector, social media or autonomous driving. In the case of non-maleficence, policy makers have to define the exact meaning of non-maleficence for the respective use cases and distinguish between different cases of criticality. In contrast, some AI principles require more debate, as stakeholders have different conceptions of autonomy and justice. As a result, policy makers need to provide a definition of how they exactly understand “autonomy” and “justice“. However, there are also clear prohibitions on certain actions such as discrimination by gender, age or profession in dilemmatic situations. Although societies might have clear preferences for age discriminations when it comes to trolley dilemma like situations, it would constitute a violation of constitutional principles.

Explicability also plays an important role for AI regulation, as it is an ethical principle developed specifically for the context of the use of AI. Due to the black-box character of AI driven decisions, strengthening the rights of consumers with more transparency of AI solutions is important for guaranteeing human oversight and consumer sovereignty. The regulator needs to distinguish between cases where explicability is important ex ante (e.g. autonomous driving, public service) and ex post (e.g. estimation of creditworthiness in the private sector).

Given the high polarization and different perceptions on AI ethics in European societies, we recommend that policy makers should take the challenges posed by AI seriously.

Recommendations for Policy Makers and Corporate Agents

The immense influence of AI on society means that the interrelationship between ethics and AI has is a major concern for policy makers and corporate decision makers alike. However, the relationship between ethics and AI is not determined by AI itself, but rather by the way we develop, use and regulate it. In this white paper, we have looked at the different positions of stakeholders on AI regulations and elaborated on the common ground of AI ethics. The position of the various stakeholders concerning their overarching values in the AI ethics discourse is influenced by

perception and interests. This is partly driven by relevant entities reasons for existence. As a result of our analysis, we found that the prevailing values and norms represent a difficult task for decision makers, as some of the principles of AI ethics can be interpreted in conflicting ways.

Based on our review of various guidelines on AI ethics, secondary legal frameworks and the interests of different stakeholder groups, we have derived a set of conclusions and recommendations for policy makers, policy makers of different levels, stakeholder groups and industry. We first identify the following major priorities:

1. Policy makers need to distinguish between *hard* and *soft* frameworks for regulating different AI use cases.
2. Policy makers and stakeholder groups should foster global collaboration on defining minimum standards for AI.
3. Policy makers and corporate decision makers need to pay heed to potential conflicts within different principles of AI ethics and how relevant stakeholder perceptions and preferences interact with these conflicts.

General Conclusions

Many of the insights gain through this research lie beyond or across the identified principles of AI ethics. What we also find is that there is still a big gap between the theory of AI regulation and the actual implementation in laws and action guiding regulation. In this section, we outline these overarching findings, which help to narrow down the AI ethics principles for legislation, and distinguish between conclusions relevant for regulators [R] and industry [I].

- AI governance requires overarching principles. The 5 principles developed by the AI4People Forum (Beneficence, Non-maleficence, Autonomy, Justice and Explicability) form, in our view, the baseline for further discussions on AI ethics in Europe.[I,R]
- The further enhancement of AI requires ethical principles. At the same time, the reaping of ethical benefits of AI also depends on a high-tech infrastructure. [I,R]
- The key task of regulators and policy makers will be to tackle cases where the different AI principles conflict with each other. Exemplary cases are the conflict between beneficence and explicability.[R]

- Given the diversity of AI use cases, a “one size fits all” approach is not applicable. We recommend, therefore, to define criteria for distinguishing between “hard” and “soft” rules.[R]:
    - The involvement of AI in “critical decisions” requires hard rules.
    - The involvement in rather uncritical situations requires soft rules (ethics codices).
    - Regulators need to define different levels of criticality. Criticality - in the sense of the necessity to establish legal ramifications - needs to take irreversibility, individual rights and trust in social institutions into account.
    - The aspect of time criticality (e.g. autonomous driving) requires a special focus in this case, due to its close linkage with irreversibility.
- Legislation needs to include both “top-down” and “bottom-up” approaches in order to connect the inputs from industry to the interest of society.[R]
    - The second AI4People paper[1] has dealt with the implications of AI governance. One important finding of this paper is that “the complex set of provisions regulating the production and use of AI for autonomous vehicles scarcely overlaps with that of AI appliances for smart houses, for finance, etc. The same holds true in terms of governance,” with stresses the relevance of context dependency.
    - We argue that policy makers should combine the approaches of top-down and bottom-up in the process of finding the apt legal framework.
- The European Union and national governments should embrace the efforts of international organizations, such as the ITU, to standardize AI regulations [R][2]:
    - The pivot to digital sovereignty should contribute neither to a widening of the digital divide nor to technological disintegration (case by case).
    - It may be necessary to define areas in AI research and development more open to international cooperation and exchange, and areas that are more restricted due to concerns revolving around digital sovereignty.
    - The goal of international AI governance should be to establish a competitive environment, based on fair principles and low market entry barriers.

[1] Compare: Pagallo et al. (2019).
[2] We encountered different stakeholder perspectives on the exact degree of international cooperation. The Beijing principles as well as the G20 Human Centered-AI Principles focus on the realization of international cooperation and joint governance mechanisms, while the EU has been recently focusing on strengthening digital sovereignty.

Beneficence

The vast majority of stakeholder and frameworks acknowledge that AI should be in the service for the common good and the benefit of humanity. Nevertheless, it often remains unclear what is explicitly meant by beneficence and how it applies in different use cases. Thus, we find that:

- Policy makers and regulators need to acknowledge the relevance of AI ethics and regulation to reap the benefits of AI. Corporate decision makers require clear rules for further advancing research and development of AI (e.g. autonomous driving). [R,I]
- The beneficence criterion should be adapted to the distinct fields of application. Beneficence has different implications for the health sector, social media or autonomous driving.[R,I]
    - In the case of autonomous driving, it makes sense to coordinate efforts to come closer to Vision Zero[3]
    - The same might apply to the health sector and cancer detection, where AI should enhance the accuracy of diagnosis.
- In most cases, policy makers should leave enough space for enterprises to define their own vision on AI, as long as it does not contradict other AI principles or legal frameworks. [R]
    - An exemplary case for how to integrate beneficence into corporate decision-making is the AI for Good Strategy of Microsoft, which illustrates how companies can link the further development of AI with ethical goals. Another example of beneficial AI is the chatbot Raaji.[4]
- Governing bodies should outline key areas in which they seek to further enhance AI capacities.[R][5]

Non-maleficence

While the principles that AI should work against the risks arising from technological innovations is fundamental, it remains vague. Therefore, we find that:

[3] The goal of vision zero („reduction of road fatalities to zero“) has been integrated in the German Ethics Code on Automated and Connected Driving („The primary purpose of partly and fully automated transport systems is to improve safety for all road users“).
[4] UNICEF. (2018). Design for girls, by girls - Period. https://www.unicef.org/innovation/U-Report/design-for-girls-by-girls-pakistan
[5]The G20 have explicitly stressed the connection between the UN SDGs and the UN SDGs (Compare: Chapter 2.1.6).

- The legislator has to define the exact meaning of non-maleficence for the respective use cases and distinguish between different cases of criticality.[R]
    - In terms of criticality, we would emphasize on the aspects of irreversibility, individual rights and societal trust.
- In order to prevent harm for third parties, policy makers need to establish clear accountability structures and push forward quality seals for realizing safety and security gains. Some cases might require industry based regulations, such as autonomous driving. [R][6]
- International standards frameworks play a critical role in developing a common understanding for minimum required standards. [R,I]
    - This relates to the discourse on common regulation initiated by the Beijing Principles and the guidelines of the OECD.
- Policy makers need to adapt consumer protection and human rights frameworks (such as the UN Guiding Principles on Business and Human Rights) to incorporate dilemmas related to AI and elaborate on the influence of algorithm driven decisions on systemic aspects of governance, such as rule of law or freedom of speech.[R]

Autonomy

In the context of AI, autonomy refers to the need to strike a balance between the decision-making power we retain for ourselves and that which we delegate to AI. This concept lies at the core of human-centered AI and is of particular importance to the European context. However, room for interpretation still exists. Thus, we find that:

- Stakeholders have different concepts of autonomy. As a result, policy makers need to provide a definition of how they understand “autonomy” and clarify whether it is understood in a *narrow sense* or in a *broad sense*.[R]
    - For example, the Catholic Church has emphasized a rather broad understanding of human autonomy, as has the Opinion of the German Data Ethics Commission.
- Policy makers need to define concrete lines for the question of when human decisions should/can/must be replaced by AI.[R]

[6]The importance of accountability has also been identified in most AI frameworks, such as the German Data Ethics Code for Automated and Connected Driving.

    - The Singapore framework provides important overview input for the debate on how to establish criteria and frames these criteria within the severity/criticality debate.
- Again, a “one size fits all” approach is not applicable here due to the different implications of irreversibility and systemic relevance, such as confidence of individuals in AI solutions.[R]
    - For instance, some companies have already develop internal approaches: Daimler has issued own principles on AI regulation with an emphasis on the importance of self-determination in autonomous driving, while Siemens has stressed the role of human oversight as one of mitigation principles on AI.

Justice

While a more elusive principle, it is largely agreed upon that AI should promote justice and eliminate all types of discrimination. The recent development of the use of AI in legal systems and compliance can be seen as a further starting point for instrumentalizing AI in order to enhance justice and accountability. At the same time:

- Legislation has to define the baseline of justice and has to develop concepts for different types of justice, such as fair and equal access.[R]
- The implications of the justice principle differ in the private and public sector. The standards for fairness and justice need to be stricter in the case of the public sector.[R]
- The legislator has to implement procedures for guaranteeing compliance to structural principles such as democratic accountability, rule of law and the right to remedy.[R]
- The mitigation of biases also requires industry standards, due to the fact that all industry types have specific environments.[R,I]
- In dilemmatic situations, regulators should not allow for the discrimination by gender, age or profession. Although societies might have clear preferences for age discriminations when it comes to trolley dilemma like situations, it would constitute a violation of constitutional principles.[R,I]

Explicability

Explicability as an ethical principle was developed specifically for the context of the use of AI. In essence, AI decisions must be understandable and interpretable. Moreover, we interpret this principle as including issues of traceability and auditability. Companies have already developed

several aspects of explicability in their processes.[7] In general, the debate has shifted from a focus on full transparency of codes and algorithms to a focus on the explicability of procedures. The relevance of this concept applies first and foremost to AI use in the public sector. Given these conditions, we find that:

- Strengthening the rights of consumers in regard to more transparency of AI solutions is important for guaranteeing human oversight and consumer sovereignty.[R]
- The regulator needs to distinguish between cases where explicability is important ex ante (e.g. autonomous driving, public service) and ex post (e.g. estimation of creditworthiness in the private sector).[R]
- The European legislator should focus on streamlining reporting standards concerning AI. This could be based on the further enhancement of already existing frameworks (for instance: CSR Guidelines of the European Union).[R]
- In the case of accidents, it needs to be clear why the accident happened. A comparable example would be the use of black boxes in airplanes, which aids investigative units in uncovering the reasons for an accident, allowing them to make corrections to the system in the future.[R,I]
- Research and development of AI systems with a special focus on displaying explaining factors of AI process results should be encouraged. [R,I]

Research Statement: The White Paper on AI Ethics and Governance is based on the findings of a Research Project that was supported by Huawei. The recommendations, suggestions, and conclusions in this document are the outcome of systematic academic research and not influenced by any funding.

[7] Google has actively pushed forward the dissemination of technology related to explainable AI, such as xAI. Microsoft, for instance, is involved in the OpenAI project. Facebook is also working on solutions such as captum [software to explain machine learning] to allow for decisions made by AI to be explained. NVIDIA engineers, for example in the case of drive PX, implemented an opening of the AI black box, developing a way to get a Drive PX vehicle to explain its driving style visually.