What are good methods so the user also knows how an actor will be trusted

This research is based on the TrustInAI research, where we now determine for each principle which method(s) are good for it to use. There will not be a lot of research in this document, but methods that are thought about which can help achieve the principles. Although I would rather have that everybody in the DVerse network follow these 7 principles. I know it's not feasible to have everybody know and keep to it. It could also be that somebody else has a method that achieves the principle but is not included in this document. They should include it here so others can also implement those methods.

There is currently also a plan to have actors around that check if the principles are being followed, but that will be difficult in itself as it's quite difficult to have the source code. Docker-In-Docker can help with this, but there will still be challenges inside.

Human agency and oversight

AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights.

What are the methods for Human agency and oversight?

- **Human-in-the-loop**: This approach involves human intervention in the decision-making process of the AI system. The human operator actively participates in the AI system’s operation and can override its decisions if necessary.
- **Human-on-the-loop**: In this approach, the AI system operates independently, but a human operator monitors its operation and can intervene and override decisions when necessary.
- **Human-in-command**: This approach ensures that the AI system does not operate without explicit human approval. The human operator has full control over the AI system’s operation and decision-making process.

What are the characteristics of the methods for Human agency and oversight?

Human-in-the-loop:

- The maker can collaborate with domain experts so the data can be as accurate and of the highest quality as possible.
- You can also continuously monitor models for bias or errors and then make the necessary adjustments.
- You can add human annotation or review steps in the data preprocessing stage.
- Creating feedback loops between humans and AI models to continuously improve performance.
- Designing systems that allow for human intervention in certain situations. (Ex. Bot can't answer it.)
- Regularly audit and monitor AI systems for bias or errors.

Human-on-the-loop:

- The AI system operates independently, but its operation is continuously monitored by a human operator.
- The human operator can intervene and override decisions when necessary, but is not involved in every decision-making process.
- This approach allows for real-time monitoring and intervention, which can be crucial in high-stakes situations.
- It also allows for post-hoc analysis and adjustment of the AI system based on its performance and outcomes.

Human-in-command:

- The AI system does not operate without explicit human approval.
- The human operator has full control over the AI system's operation and decision-making process.
- This approach ensures that the AI system is completely under human control and cannot make decisions independently.
- It is particularly useful in situations where the stakes are high and human judgement is indispensable.
- It also ensures that the AI system is used responsibly and ethically, as the human operator can always intervene and control its operation.

Technical Robustness and safety

AI systems need to be resilient and secure.

What are the methods for Technical Robustness and safety?

- **Resilience and Security**: AI systems need to be resilient and secure. They should be designed to withstand various types of inputs and conditions, and should have security measures in place to prevent unauthorized access and manipulation.
- **Safety**: AI systems should be safe to use. This means they should have a fallback plan in case something goes wrong. This could include mechanisms for error detection, mitigation, and recovery.
- **Accuracy, Reliability, and Reproducibility**: AI systems should be accurate, reliable, and reproducible. They should provide consistent and correct output. This requires rigorous testing and validation processes.
- **Prevention of Unintentional Harm**: AI systems should be designed in such a way that unintentional harm can be minimized and prevented. This includes considering the potential misuse or unintended consequences of the AI system.

What are the characteristics of the methods for Technical Robustness and safety?

Resilience and Security:

- AI systems should be designed to withstand a wide range of inputs and conditions.
- They should have robust security measures in place to prevent unauthorized access and manipulation.
- Regular updates and patches should be applied to ensure the system remains secure against new threats.

Safety:

- AI systems should have a fallback plan or safe mode to revert to in case something goes wrong.
- They should include mechanisms for error detection, mitigation, and recovery.
- Safety considerations should be integrated into all stages of the AI system's lifecycle, from design to deployment and operation.

Accuracy, Reliability, and Reproducibility:

- AI systems should provide consistent and correct outputs.
- They should undergo rigorous testing and validation processes to ensure their accuracy and reliability.
- The results produced by the AI system should be reproducible under the same conditions.

Prevention of Unintentional Harm:

- AI systems should be designed to minimize and prevent unintentional harm.
- This includes considering the potential misuse or unintended consequences of the AI system.
- Regular risk assessments should be conducted to identify and mitigate potential harm.
- Ethical guidelines and legal regulations should be adhered to in order to prevent harm.

Privacy and data governance

Full respect for privacy and data protection, adequate data governance mechanisms must also be ensured.

What are the methods for Privacy and data governance?

- **Respect for Privacy and Data Protection**: AI systems should fully respect privacy and data protection rights. This includes complying with all relevant laws and regulations.
- **Data Governance**: Adequate data governance mechanisms must be ensured. This includes taking into account the quality and integrity of the data, and ensuring legitimized access to data.
- **Data Quality and Integrity**: AI systems should use high-quality data that accurately represents the problem space. The integrity of the data should be maintained at all times.
- **Legitimized Access to Data**: Access to data should be legitimized. This means that data should only be accessed by authorized individuals or systems, and only for legitimate purposes.

What are the characteristics of the methods for Privacy and data governance?

Respect for Privacy and Data Protection:

- AI systems should comply with all relevant privacy and data protection laws and regulations.
- They should ensure that personal data is processed in a lawful, fair, and transparent manner.
- They should implement appropriate security measures to protect personal data from unauthorized access, use, disclosure, alteration, and destruction.

Data Governance:

- AI systems should have robust data governance mechanisms in place.
- These mechanisms should ensure the quality and integrity of the data used by the AI system.
- They should also ensure that data is accessed and used in a legitimate and authorized manner.

Data Quality and Integrity:

- AI systems should use high-quality data that accurately represents the problem space.
- The integrity of the data should be maintained at all times.
- The AI system should be designed to handle errors in the data and to mitigate their impact on the system’s operation and outputs.

Legitimized Access to Data:

- Access to data should be legitimized, meaning that it should only be accessed by authorized individuals or systems.
- The AI system should implement robust access control mechanisms to prevent unauthorized access to data.
- The AI system should also ensure that data is used only for legitimate purposes and in a manner that respects privacy and data protection rights.

Transparency

The data, system, and AI business models should be transparent.

What are the methods for Transparency?

- **Transparent Data, System, and AI Business Models**: The data, system, and AI business models should be transparent. This means that all aspects of the AI system, including its inputs, processes, and outputs, should be understandable and accessible to relevant stakeholders.

- **Traceability Mechanisms**: Traceability mechanisms can help achieve transparency. These mechanisms can track and document the decision-making process of an AI system, which can help stakeholders understand how a particular output or decision was reached.
- **Explainability**: AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. This means that the workings of the AI system should be interpretable and understandable to humans.
- **Awareness of Interaction with an AI System**: Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

What are the characteristics of the methods for Transparency?

Transparent Data, System, and AI Business Models:

- All aspects of the AI system, including its inputs, processes, and outputs, should be understandable and accessible to relevant stakeholders.
- The data used by the AI system should be transparent, meaning that its source, quality, and processing methods should be disclosed.
- The AI business model should also be transparent, meaning that stakeholders should understand how the AI system is used in the business context.

Traceability Mechanisms:

- Traceability mechanisms can track and document the decision-making process of an AI system.
- These mechanisms can help stakeholders understand how a particular output or decision was reached.
- They can also help identify and correct errors in the AI system's operation.

Explainability:

- AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned.
- This means that the workings of the AI system should be interpretable and understandable to humans.
- The AI system should provide clear and understandable explanations for its decisions.

Awareness of Interaction with an AI System:

- Humans need to be aware that they are interacting with an AI system.
- They must be informed of the system's capabilities and limitations.
- This awareness can help humans make informed decisions and use the AI system effectively and responsibly.

Diversity, non-discrimination and fairness

Unfair bias must be avoided.

What are the methods for Diversity, non-discrimination and fairness?

- **Avoidance of Unfair Bias**: AI systems should avoid unfair bias. This could have multiple negative implications, from the marginalization of vulnerable groups to the exacerbation of prejudice and discrimination.
- **Fostering Diversity**: AI systems should foster diversity and be accessible to all, regardless of any disability.
- **Non-discrimination**: AI systems should not discriminate against any individual or group. This includes ensuring that the AI system does not perpetuate harmful biases or stereotypes.
- **Fairness**: AI systems should be fair. This means that they should treat all individuals and groups equally and impartially.

What are the characteristics of the methods for Diversity, non-discrimination and fairness?

Avoidance of Unfair Bias:

- AI systems should be designed and trained in a way that avoids unfair bias.
- This includes using diverse and representative data for training, and regularly testing the system for bias.
- Any identified bias should be corrected, and the system should be continuously monitored and updated to ensure that bias does not creep in over time.

Fostering Diversity:

- AI systems should be designed to be inclusive and accessible to all, regardless of any disability.
- This includes considering diversity in the design and development process, and ensuring that the system is usable by a wide range of users.
- The system should also promote diversity by treating all users fairly and equally.

Non-discrimination:

- AI systems should not discriminate against any individual or group.
- This includes ensuring that the system does not perpetuate harmful biases or stereotypes.
- The system should treat all users equally, and decisions made by the system should be fair and impartial.

Fairness:

- AI systems should be fair, meaning they should treat all individuals and groups equally and impartially.
- The system should not favor one group over another, and its decisions should be based on relevant and fair criteria.
- The system should be transparent about its decision-making process, and users should be able to understand and challenge its decisions if they believe they are unfair.

Societal and environmental well-being

AI systems should benefit all human beings, including future generations.

What are the methods for Societal and environmental well-being?

- **Benefit All Human Beings**: AI systems should benefit all human beings, including future generations. This means that the benefits of AI should be distributed widely, rather than being concentrated in the hands of a few.
- **Sustainability**: AI systems must be sustainable and environmentally friendly. This includes considering the environmental impact of AI systems throughout their lifecycle, from development to deployment to decommission.
- **Social Impact**: The social and societal impact of AI systems should be carefully considered. This includes understanding how AI systems affect social dynamics and structures, and taking steps to mitigate any negative impacts.
- **Involvement of Relevant Stakeholders**: Relevant stakeholders should be involved throughout the entire lifecycle of AI systems. This can help ensure that the AI system is designed and used in a way that aligns with societal values and norms.

What are the characteristics of the methods for Societal and environmental well-being?

Benefit All Human Beings:

- AI systems should be designed and used in a way that benefits all human beings, including future generations.
- The benefits of AI should be distributed widely, rather than being concentrated in the hands of a few.

- This includes considering the needs and interests of all stakeholders, and ensuring that the AI system does not disproportionately benefit or harm any particular group.

Sustainability:

- AI systems should be sustainable and environmentally friendly.
- This includes considering the environmental impact of AI systems throughout their lifecycle, from development to deployment to decommission.
- Efforts should be made to minimize the environmental footprint of AI systems, such as by optimizing their energy efficiency and using renewable energy sources where possible.

Social Impact:

- The social and societal impact of AI systems should be carefully considered.
- This includes understanding how AI systems affect social dynamics and structures, and taking steps to mitigate any negative impacts.
- AI systems should be used in a way that promotes social good and contributes to societal well-being.

Involvement of Relevant Stakeholders:

- Relevant stakeholders should be involved throughout the entire lifecycle of AI systems.
- This can help ensure that the AI system is designed and used in a way that aligns with societal values and norms.
- Stakeholder involvement can also help identify the potential social and environmental impacts of the AI system, and develop strategies to address them.

Accountability

Mechanisms must be in place to ensure responsibility and accountability for AI systems and their outcomes.

What are the methods for Accountability?

- **Responsibility and Accountability for AI Systems and Their Outcomes**: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. This includes the responsibilities of AI developers, users, and regulators.
- **Auditability**: Auditability, which enables the assessment of algorithms, data, and design processes, plays a key role in accountability. It helps in understanding and explaining the decisions made by the AI system.
- **Risk Management and Redress for Adverse Impact**: There should be measures in place for risk management and redress for adverse impact caused by AI systems.
- **Clear Rules and Regulations**: Developing clear rules and regulations about how AI systems can be used can also help in making AI systems more accountable.

What are the characteristics of the methods for Accountability?

Responsibility and Accountability for AI Systems and Their Outcomes:

- AI developers, users, and regulators should all have clearly defined responsibilities.
- Mechanisms should be in place to hold these parties accountable for the AI system's outcomes.
- This includes ensuring that the AI system operates as intended, and addressing any negative impacts that it may have.

Auditability:

- AI systems should be auditable, meaning that their algorithms, data, and design processes can be assessed.
- This helps in understanding and explaining the decisions made by the AI system.

- Auditability also promotes transparency and trust in the AI system.

Risk Management and Redress for Adverse Impact:

- Measures should be in place for managing the risks associated with AI systems.
- This includes identifying potential risks, implementing strategies to mitigate these risks, and monitoring the AI system's operation to detect and address any adverse impacts.
- There should also be mechanisms for redress, meaning that individuals or groups who are negatively impacted by the AI system should have a means of seeking remedy.

Clear Rules and Regulations:

- Clear rules and regulations should be developed about how AI systems can be used.
- These rules and regulations should be designed to ensure that AI systems are used responsibly and ethically.
- They should also provide guidance for AI developers, users, and regulators, helping them understand their responsibilities and the expectations for the AI system's operation.

What is feasible in the time frame of this project?

I would say everything is possible but the Societal and environmental well-being is possibly less important and should be left to an AI group. Everything else is possible to implement within some days. Although not everything should be treated equal so look into the next chapter about what to prioritize.

What are we prioritizing?

This depends on the use case, for which the developer should find their own way. But all seven of these 7 key principles should be used. In what I mean, you at least do something about the Accountability and not that you need to do Auditability for every bot. But Auditability is still important and should be implemented or discussed in terms of how it will be done.

Some cases where one should be prioritized over the other could be:

- Human Agency and Oversight over Transparency: In a healthcare setting, an AI system is used to predict patient health outcomes. While transparency is important, it might be more crucial to ensure that the AI does not operate independently and that healthcare professionals can override its decisions. This is to ensure that the final decision always incorporates human judgement, especially in life-critical situations.
- Fairness over Transparency: An AI system used for loan approval should prioritize fairness to ensure all applicants are treated equally, regardless of their race, gender, or background. While it's important for the system to be transparent, ensuring it doesn't discriminate is arguably more crucial.
- Societal and Environmental Well-being over Transparency: An AI system designed to optimize energy usage in a city should prioritize societal and environmental well-being. While it's beneficial for the system's decisions to be explainable, the primary focus should be on reducing energy consumption and carbon emissions.
- Privacy and Data Governance over Robustness and Safety: In a personal digital assistant application, privacy and data governance might be prioritized over robustness and safety. This is because it's crucial to protect user data and maintain privacy, even if it means the AI might not always function optimally.

Resources used

https://www.europarl.europa.eu/cmsdata/196377/AI HLEG_Ethics Guidelines for Trustworthy AI.pdf
https://futurium.ec.europa.eu/en/european-ai-alliance/best-practices/practical-organizational-framework-ai-accountability