IRIS 2021

APPLYING THE HLEG’S ASSESSMENT LIST FOR TRUSTWORTHY AI (ALTAI) IN THE AUTOMOTIVE CONTEXT

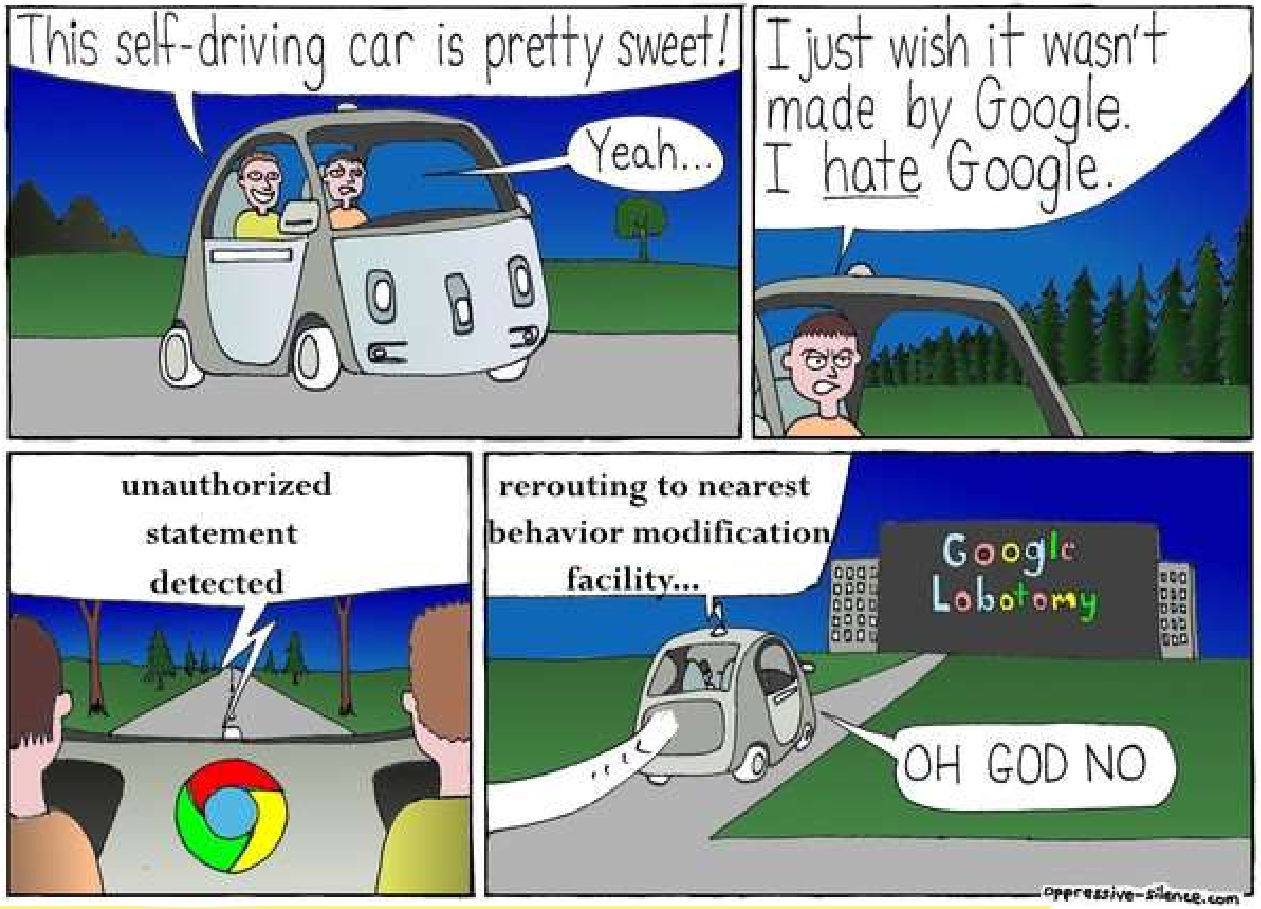
David Andrews & Hristina Veljanova

Das Land Steiermark
Wissenschaft und Forschung, Gesundheit und Pflegemanagement

Outlook

1. Assessment List for Trustworthy AI (ALTAI)
2. VERDI Criteria Catalogue for trustworthy highly-automated vehicles and driver assistance systems (SAE L3)
3. Cross-checking VERDI & ALTAI
4. Wrap up

Imagine a world...


This self-driving car is pretty sweet!
Yeah...
I just wish it wasn't made by Google. I hate Google.
unauthorized statement detected
rerouting to nearest behavior modification facility...
Google Lobotomy
OH GOD NO
oppressive-silence.com

[Oppressive Silence 2016]

~~Imagine~~ Realise a world...

Self-Driving Mercedes Will Be Programmed To Sacrifice Pedestrians To Save The Driver

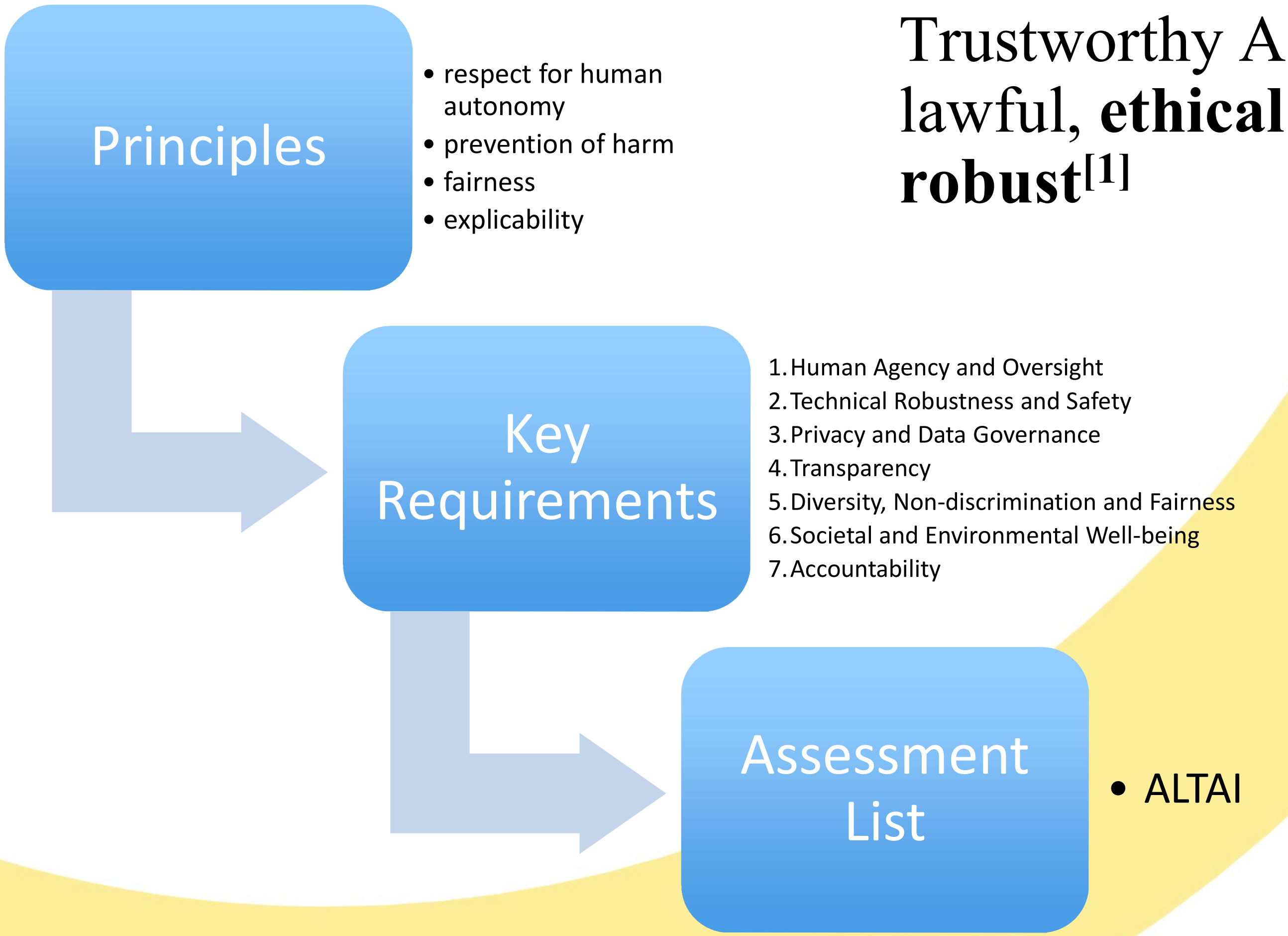
Mercedes gets around the moral issues of self-driving cars by deciding that–of course–drivers are more important than anyone else.

KEEP SUMMER SAFE

[Schwifty Memes 2019], see also Sorrel (2016)

Assessment List for Trustworthy AI (ALTAI)

AI HLEG & Trustworthy AI



Principles

- respect for human autonomy
- prevention of harm
- fairness
- explicability

Key Requirements

1. Human Agency and Oversight
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, Non-discrimination and Fairness
6. Societal and Environmental Well-being
7. Accountability

Assessment List

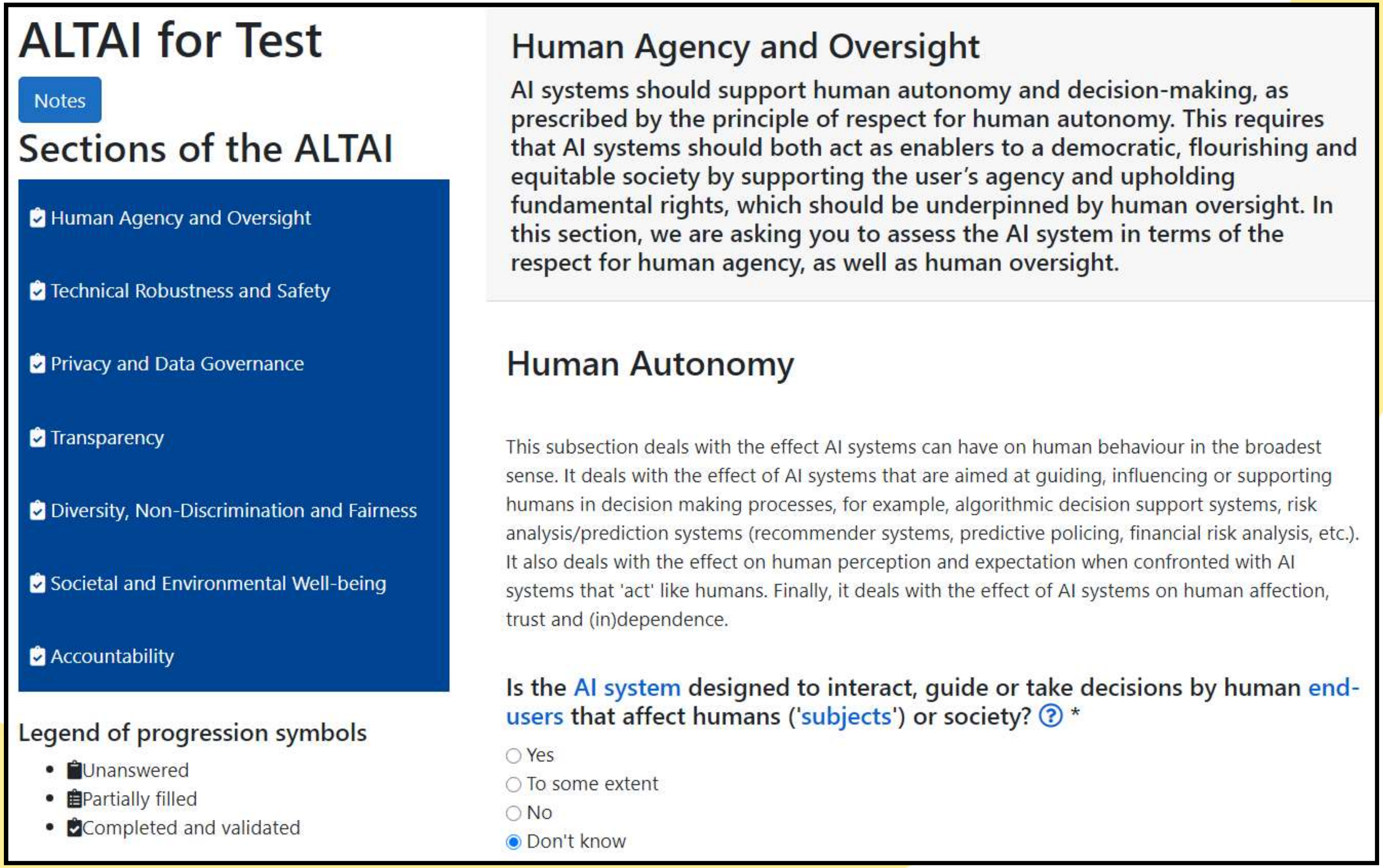
- ALTAI

Trustworthy AI should be lawful, **ethical and robust**[1]

[1] AI HLEG (2019): “Ethics Guidelines for Trustworthy AI”

ALTAI Web Tool

Screenshots ALTAI



ALTAI for Test

Notes

Sections of the ALTAI

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, Non-Discrimination and Fairness
- Societal and Environmental Well-being
- Accountability

Legend of progression symbols

- Unanswered
- Partially filled
- ✓ Completed and validated

Human Agency and Oversight

AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user’s agency and upholding fundamental rights, which should be underpinned by human oversight. In this section, we are asking you to assess the AI system in terms of the respect for human agency, as well as human oversight.

Human Autonomy

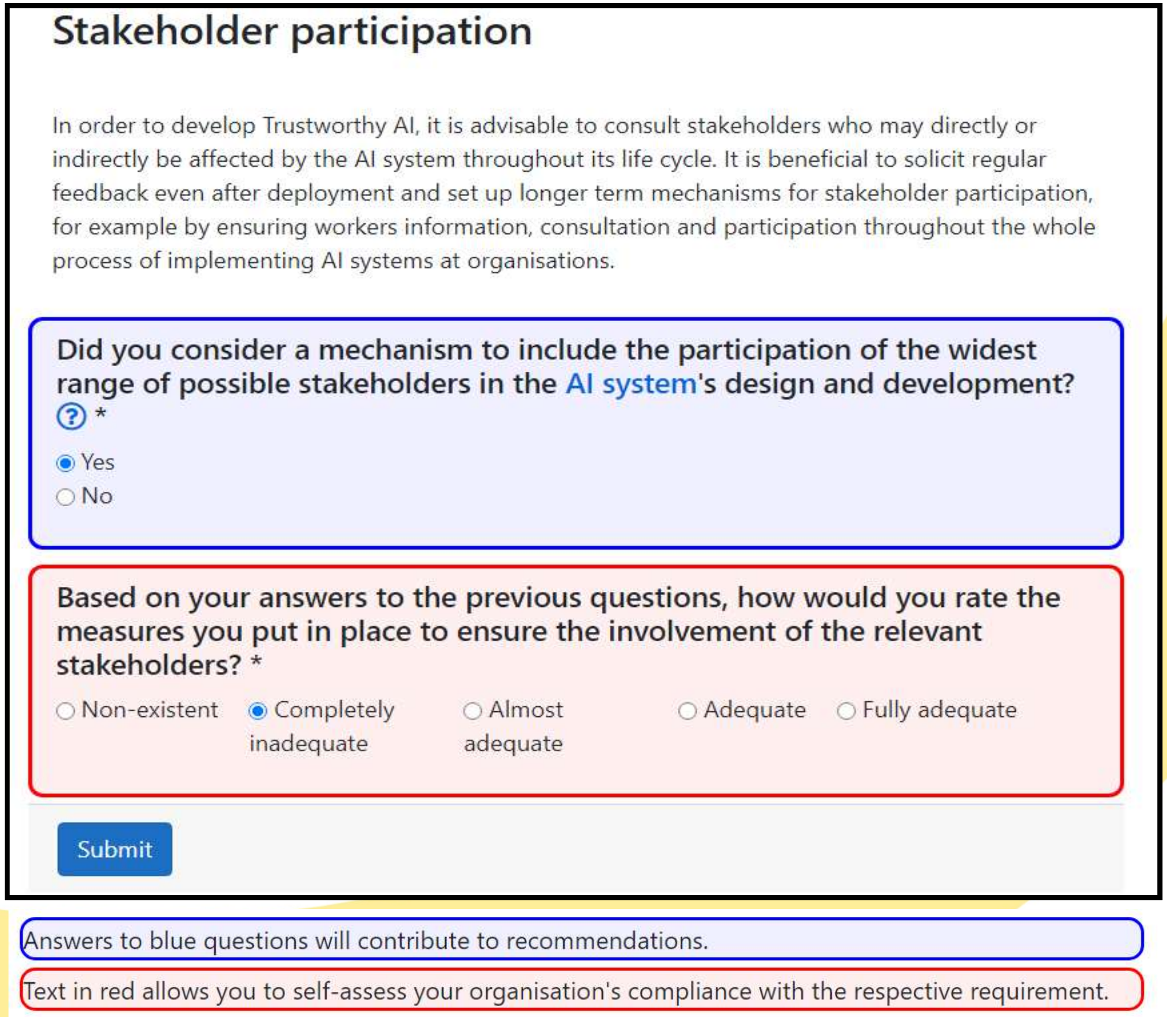
This subsection deals with the effect AI systems can have on human behaviour in the broadest sense. It deals with the effect of AI systems that are aimed at guiding, influencing or supporting humans in decision making processes, for example, algorithmic decision support systems, risk analysis/prediction systems (recommender systems, predictive policing, financial risk analysis, etc.). It also deals with the effect on human perception and expectation when confronted with AI systems that 'act' like humans. Finally, it deals with the effect of AI systems on human affection, trust and (in)dependence.

Is the AI system designed to interact, guide or take decisions by human end-users that affect humans ('subjects') or society? ? *

○ Yes
○ To some extent
○ No
● Don't know

[Own screenshot]

Screenshots ALTAI



Stakeholder participation

In order to develop Trustworthy AI, it is advisable to consult stakeholders who may directly or indirectly be affected by the AI system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations.

Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system's design and development? ? *

● Yes
○ No

Based on your answers to the previous questions, how would you rate the measures you put in place to ensure the involvement of the relevant stakeholders? *

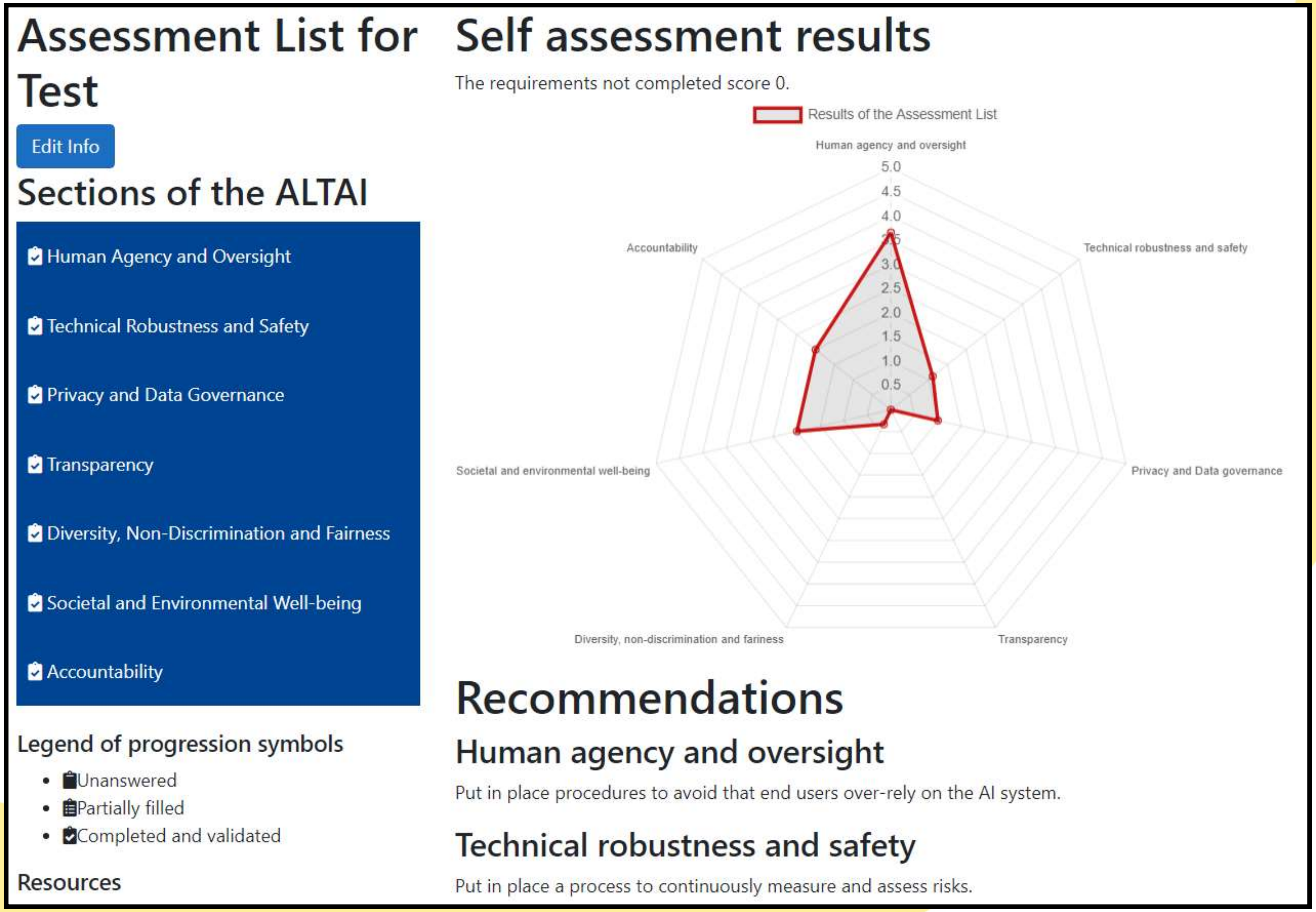
○ Non-existent ● Completely inadequate ○ Almost adequate ○ Adequate ○ Fully adequate

Answers to blue questions will contribute to recommendations.

Text in red allows you to self-assess your organisation's compliance with the respective requirement.

[Own screenshot]

Screenshots ALTAI



Assessment List for Test

Edit Info

Sections of the ALTAI

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, Non-Discrimination and Fairness
- Societal and Environmental Well-being
- Accountability

Legend of progression symbols

- Unanswered
- Partially filled
- Completed and validated

Resources

Self assessment results

The requirements not completed score 0.

Results of the Assessment List
Human agency and oversight
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Technical robustness and safety
Privacy and Data governance
Transparency
Diversity, non-discrimination and fariness
Societal and environmental well-being
Accountability

Recommendations

Human agency and oversight

Put in place procedures to avoid that end users over-rely on the AI system.

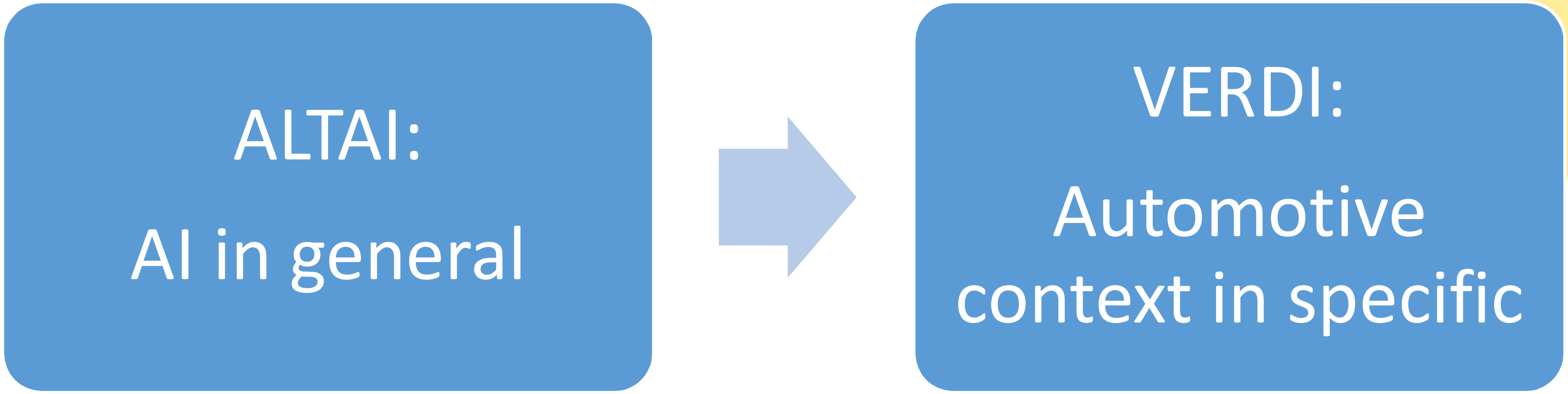
Technical robustness and safety

Put in place a process to continuously measure and assess risks.

[Own screenshot]

ALTAI → VERDI

ALTAI: “is intended for **flexible use**: organisations can […] add elements to it as they see fit, taking into consideration the **sector** they operate in” [2]



ALTAI:

AI in general

VERDI:

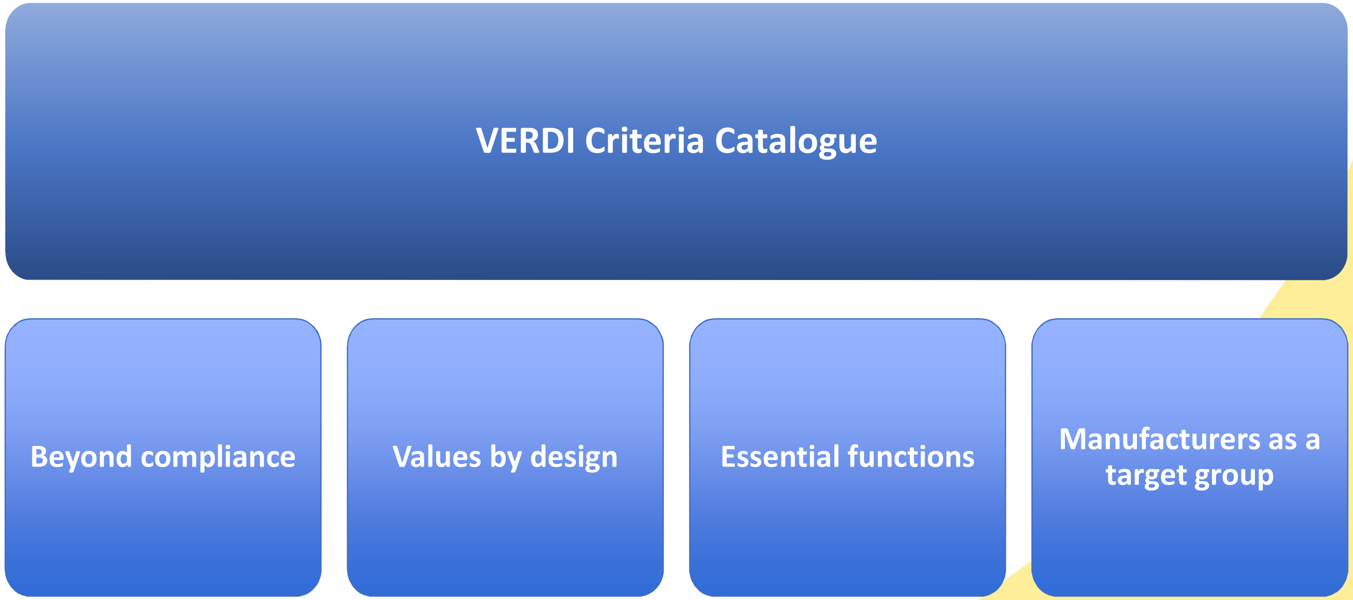
Automotive context in specific

Trustworthiness

[2] AI HLEG (2020): “The Assesment List For Trustworthy Artificial Intelligence (ALTAI)”

VERDI Criteria Catalogue
for trustworthy partially automated vehicles and driver assistance systems
(SAE L3)

Approach



VERDI Criteria Catalogue

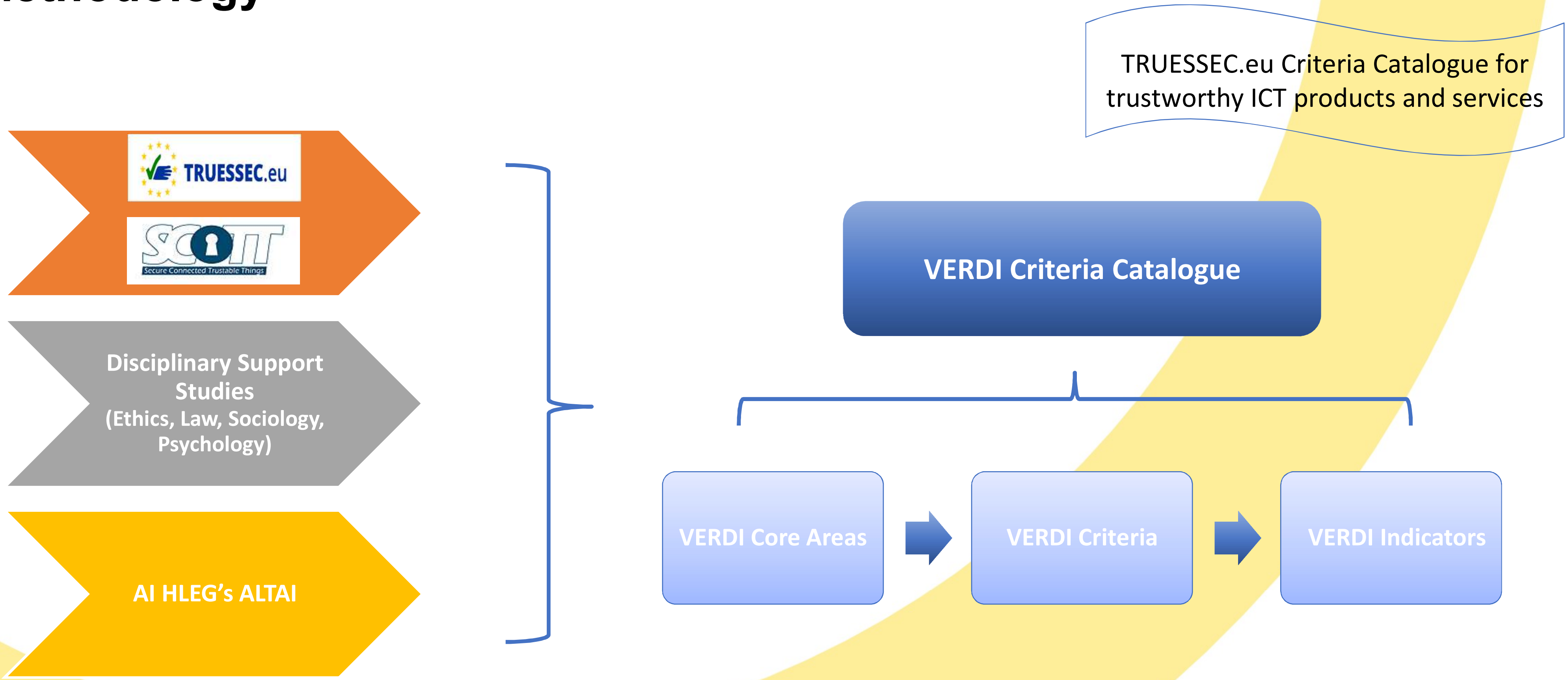
Beyond compliance

Values by design

Essential functions

Manufacturers as a target group

Methodology



Disciplinary Support Studies
(Ethics, Law, Sociology, Psychology)

AI HLEG's ALTAI

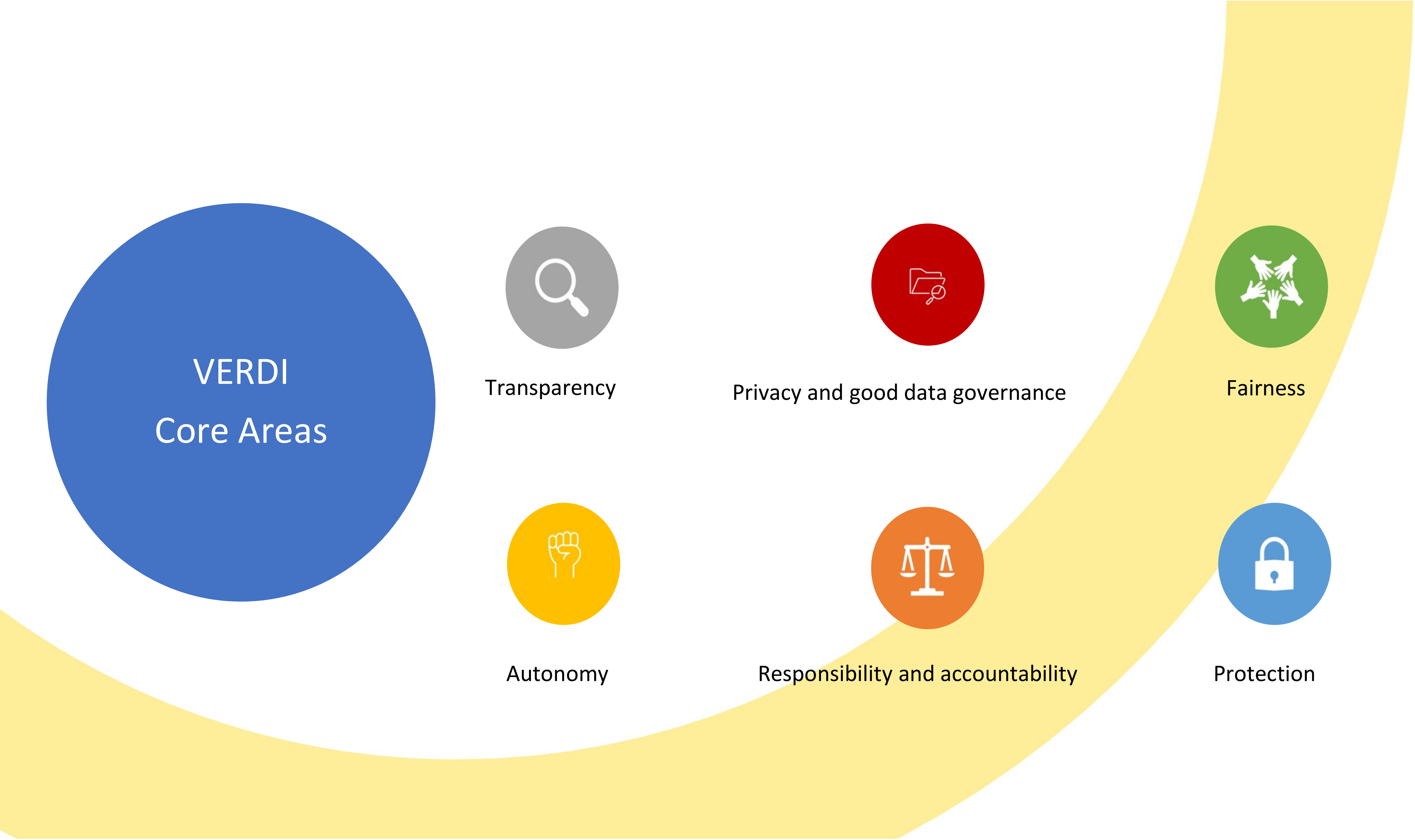
TRUESSEC.eu Criteria Catalogue for trustworthy ICT products and services

VERDI Criteria Catalogue

VERDI Core Areas

VERDI Criteria

VERDI Indicators



VERDI

Core Areas

Transparency

Privacy and good data governance

Fairness

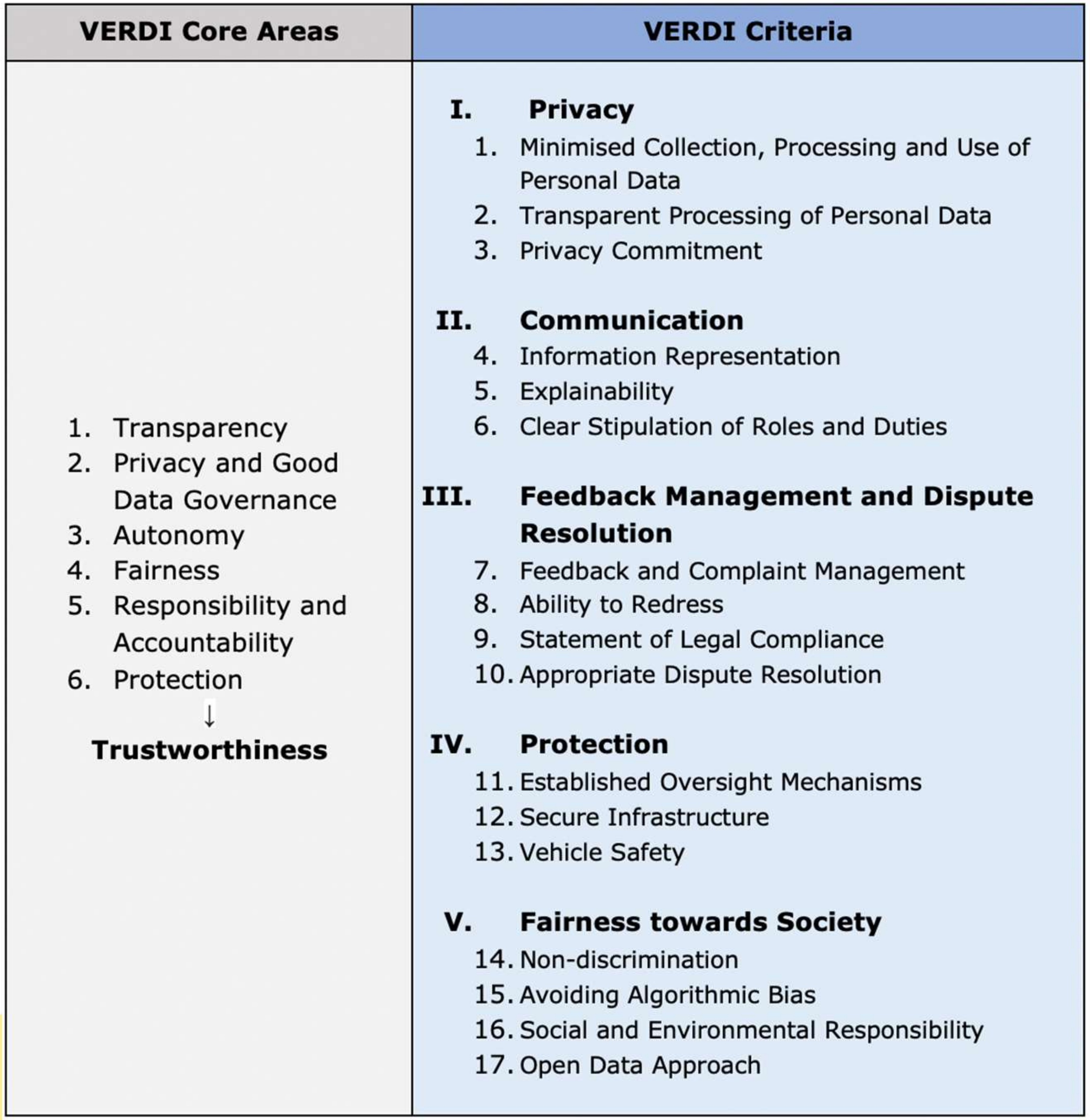
Autonomy

Responsibility and accountability

Protection

Transparency	The Core Area ‘Transparency’ encompasses provider’s **information duties** towards the user regarding the **system’s functionality and limitations** as well as the **data** that is processed by the system. Additionally, it also focuses on **information representation**.
Privacy and good data governance	This Core Area entails two aspects: (1) any **personal data** processed as part of the interaction with the system should be **protected**, and (2) the user should have the possibility to **contro**l that data.
Autonomy	Autonomy refers to the ADS providing the user with the possibility to **choose and make decisions** regarding the (non-)use of certain automation aspects and services as well as acknowledging other parties’ **rights and freedoms**.
Fairness	Fairness stands for **preventing cases of discrimination** due to algorithmic biases and societal factors (e.g. socio-economic status) and considering effects and contributions towards **social in- and exclusion**.
Responsibility and accountability	**Respect** and **clear information** about the stipulation of **roles and liabilities**. It furthermore addresses the legitimate and reasonable **expectations** of the user and society in relation to the system’s functionality and reliability.
Protection	This core area refers to the protection of users, other road users and the surrounding from any **harms and risks** that might be caused by the ADS, including physical harm (**safety**) and protection of software errors and data (**security**).

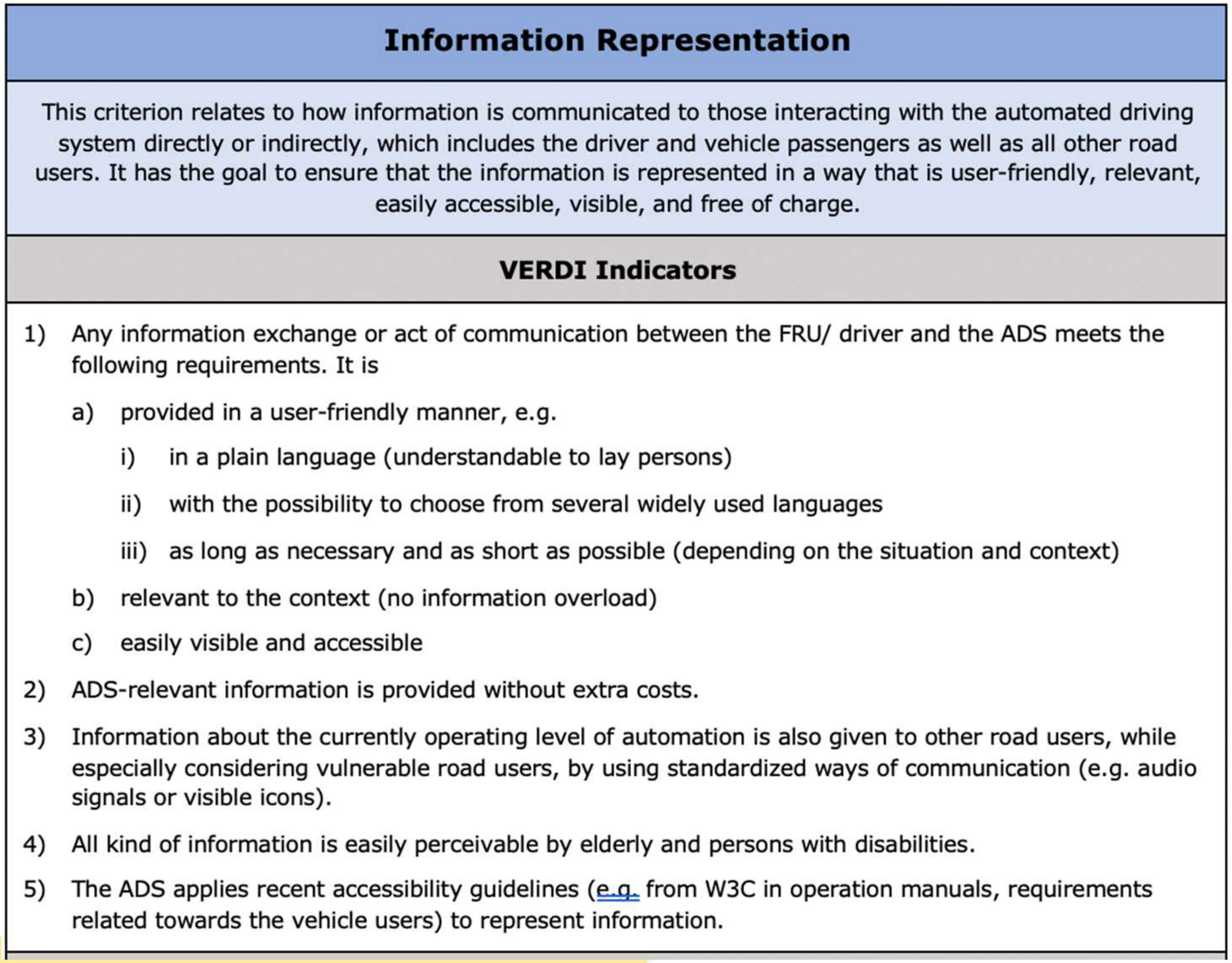
VERDI Criteria



VERDI Core Areas	VERDI Criteria
1. Transparency 2. Privacy and Good Data Governance 3. Autonomy 4. Fairness 5. Responsibility and Accountability 6. Protection ↓ **Trustworthiness**	**I. Privacy** 1. Minimised Collection, Processing and Use of Personal Data 2. Transparent Processing of Personal Data 3. Privacy Commitment **II. Communication** 4. Information Representation 5. Explainability 6. Clear Stipulation of Roles and Duties **III. Feedback Management and Dispute Resolution** 7. Feedback and Complaint Management 8. Ability to Redress 9. Statement of Legal Compliance 10. Appropriate Dispute Resolution **IV. Protection** 11. Established Oversight Mechanisms 12. Secure Infrastructure 13. Vehicle Safety **V. Fairness towards Society** 14. Non-discrimination 15. Avoiding Algorithmic Bias 16. Social and Environmental Responsibility 17. Open Data Approach

VERDI

Criteria & Indicators



Information Representation

This criterion relates to how information is communicated to those interacting with the automated driving system directly or indirectly, which includes the driver and vehicle passengers as well as all other road users. It has the goal to ensure that the information is represented in a way that is user-friendly, relevant, easily accessible, visible, and free of charge.

VERDI Indicators

1) Any information exchange or act of communication between the FRU/ driver and the ADS meets the following requirements. It is
   a) provided in a user-friendly manner, e.g.
      i) in a plain language (understandable to lay persons)
      ii) with the possibility to choose from several widely used languages
      iii) as long as necessary and as short as possible (depending on the situation and context)
   b) relevant to the context (no information overload)
   c) easily visible and accessible
2) ADS-relevant information is provided without extra costs.
3) Information about the currently operating level of automation is also given to other road users, while especially considering vulnerable road users, by using standardized ways of communication (e.g. audio signals or visible icons).
4) All kind of information is easily perceivable by elderly and persons with disabilities.
5) The ADS applies recent accessibility guidelines (e.g. from W3C in operation manuals, requirements related towards the vehicle users) to represent information.

Cross-checking VERDI & HLEG’s ALTAI

Cross-checking VERDI & HLEG's ALTAI

- Structure
- Terminology
- Tool

INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

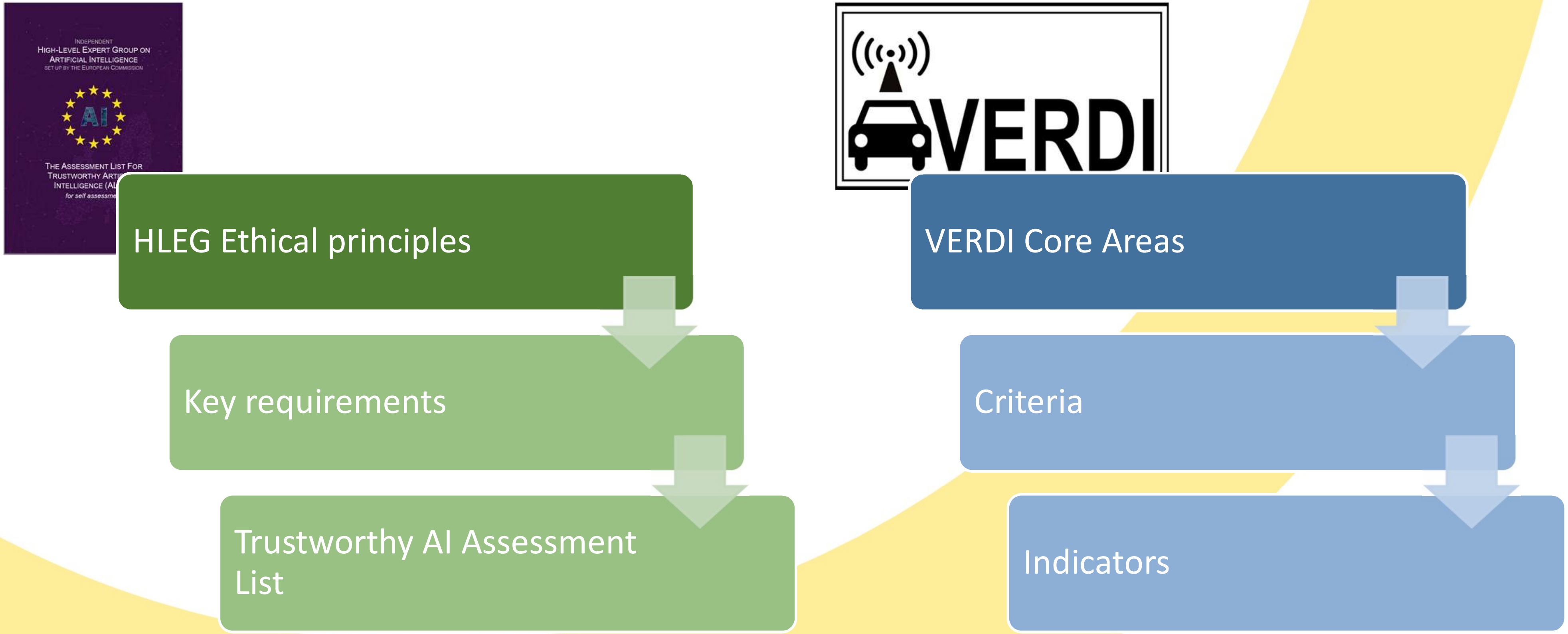
AI

THE ASSESSMENT LIST FOR
TRUSTWORTHY ARTIFICIAL
INTELLIGENCE (ALTAI)
for self assessment

[AI HLEG 2020]

VERDI

VERDI & HLEG's ALTAI: Structure



HLEG Ethical principles

Key requirements

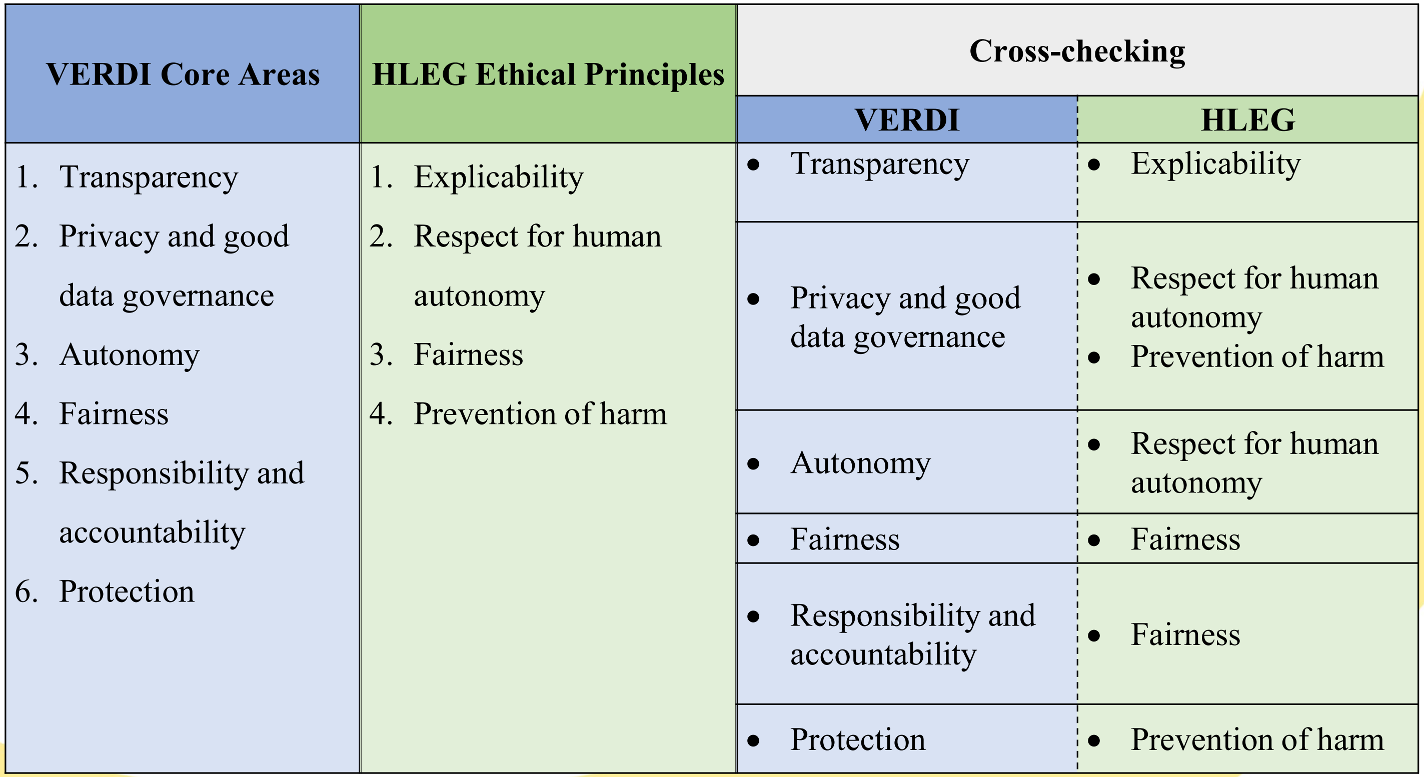
Trustworthy AI Assessment List

VERDI Core Areas

Criteria

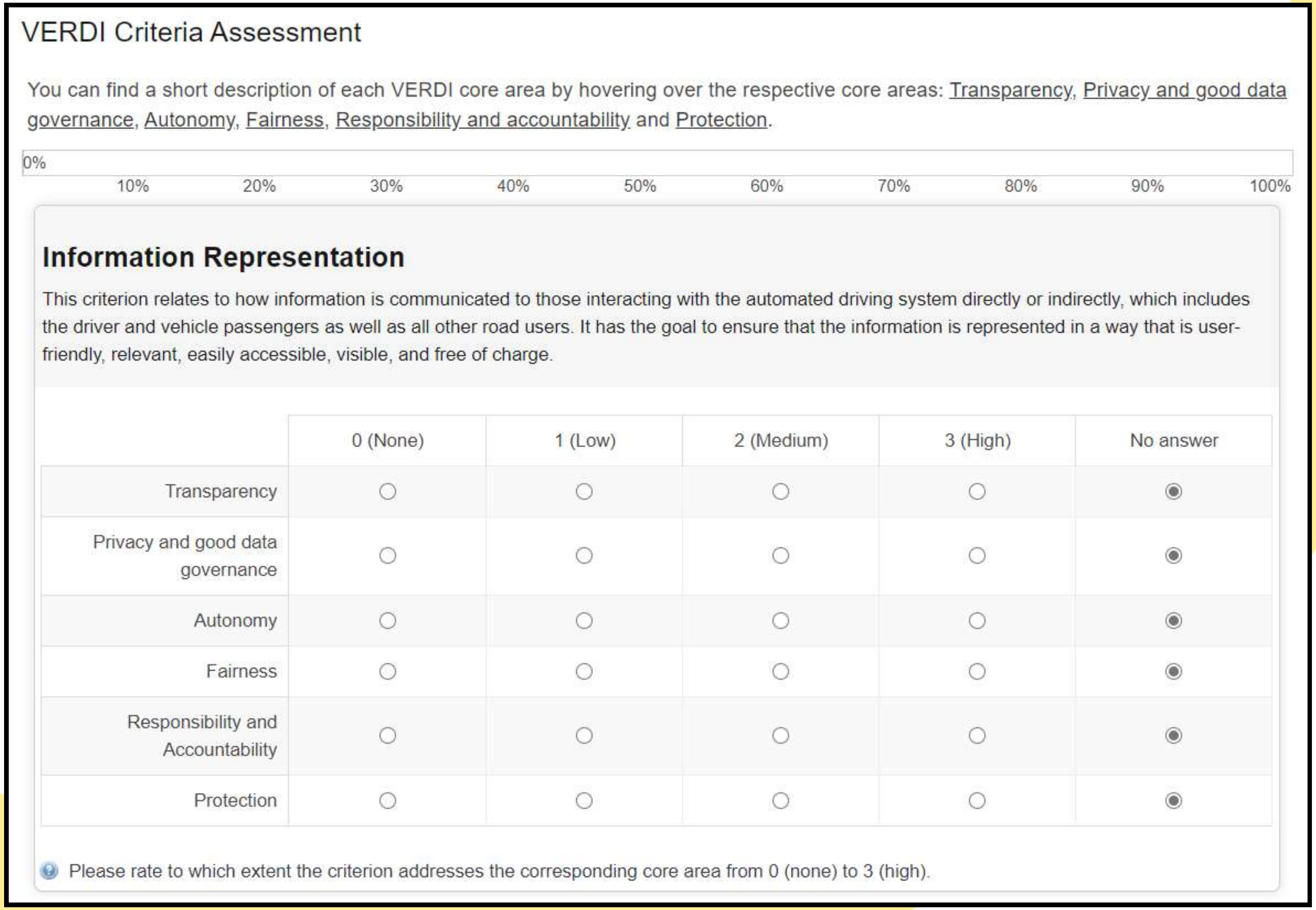
Indicators

VERDI Core Areas vs HLEG Ethical principles



VERDI Core Areas	HLEG Ethical Principles	Cross-checking	
		VERDI	HLEG
1. Transparency 2. Privacy and good data governance 3. Autonomy 4. Fairness 5. Responsibility and accountability 6. Protection	1. Explicability 2. Respect for human autonomy 3. Fairness 4. Prevention of harm	• Transparency	• Explicability
		• Privacy and good data governance	• Respect for human autonomy • Prevention of harm
		• Autonomy	• Respect for human autonomy
		• Fairness	• Fairness
		• Responsibility and accountability	• Fairness
		• Protection	• Prevention of harm

VERDI survey and next steps



VERDI Criteria Assessment

You can find a short description of each VERDI core area by hovering over the respective core areas: Transparency, Privacy and good data governance, Autonomy, Fairness, Responsibility and accountability and Protection.

Information Representation

This criterion relates to how information is communicated to those interacting with the automated driving system directly or indirectly, which includes the driver and vehicle passengers as well as all other road users. It has the goal to ensure that the information is represented in a way that is user-friendly, relevant, easily accessible, visible, and free of charge.

	0 (None)	1 (Low)	2 (Medium)	3 (High)	No answer
Transparency	○	○	○	○	●
Privacy and good data governance	○	○	○	○	●
Autonomy	○	○	○	○	●
Fairness	○	○	○	○	●
Responsibility and Accountability	○	○	○	○	●
Protection	○	○	○	○	●

Please rate to which extent the criterion addresses the corresponding core area from 0 (none) to 3 (high).

[Own screenshot]

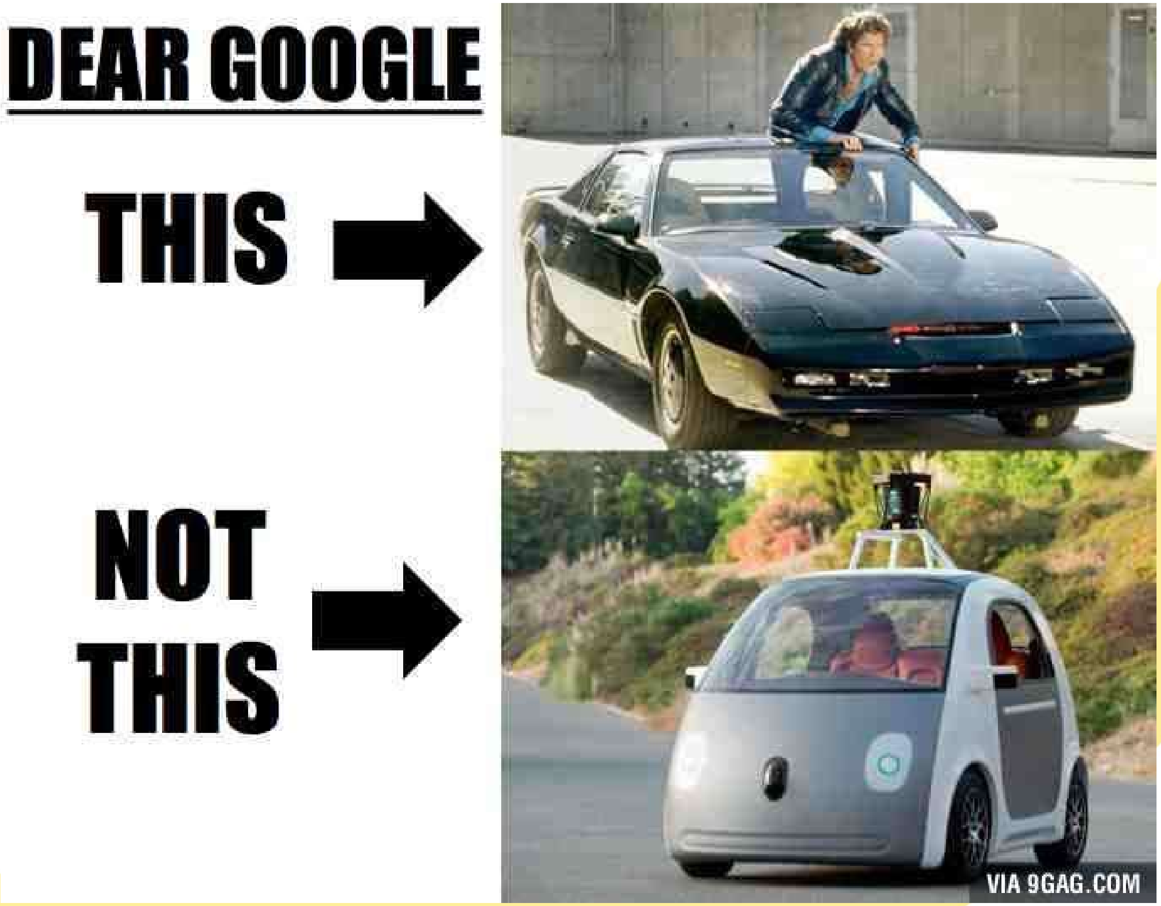
Wrap up

https://verdi.uni-graz.at/en/

Final symposium in Spring 2021

Follow-up project on standardisation and certification

~~Imagine Realise~~ Shape a world...



[9GAG 2014]

References

AI HLEG (2019): Ethics Guidelines for Trustworthy AI. [online] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [24.02.2021]

AI HLEG (2020): The Assesment List For Trustworthy Artificial Intelligence (ALTAI). [online] https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment [24.02.2021].

Sorrel, Charlie (2016): Self-Driving Mercedes Will Be Programmed To Sacrifice Pedestrians To Save The Driver. In: Fastcompany.com [online] https://www.fastcompany.com/3064539/self-driving-mercedes-will-be-programmed-to-sacrifice-pedestrians-to-save-the-driver [24.02.2021]

Images

9GAG (2014): Google's new self driving car is unacceptable…[online] https://9gag.com/gag/aYb5ZeN [24.02.2021]

Oppressive Silence (2016): self driving car. [online] https://imgur.com/gallery/I0ivc64 [24.02.2021]

Schwifty Memes (2019): [online] https://schwifty-memes.tumblr.com/post/189712495869 [24.02.2021]

Thank you for your attention!

david.andrews@uni-graz.at
hristina.veljanova@uni-graz.at

We work for tomorrow

UNI GRAZ