INDI Conference, 18 June 2021

How to design trustworthy AVs?
Bridging the gap between SSH and technology design

Hristina Veljanova

We work for **tomorrow**

UNI GRAZ

Das Land Steiermark
Wissenschaft und Forschung, Gesundheit und Pflegemanagement

ZUKUNFTSFONDS STEIERMARK

UNI GRAZ

virtual vehicle

VERDI ♪

"Trust in digitisation using the example of systems for (partially) automated driving and driver assistance (SAE L3)"

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

SAE L3 – Conditional Driving Automation

The *sustained* and *ODD*-specific performance by an *ADS* of the entire *DDT* under routine/normal operation with the expectation that the *DDT fallback-ready user* is receptive to *ADS*-issued requests to intervene, as well as to *DDT* performance-relevant *system failures* in other *vehicle* systems, and will respond appropriately.

SAE International (2021): Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, J3016, online: https://www.sae.org/standards/content/j3016_202104/ (27.05.2021)

SAE L3 – Conditional Driving Automation

The *sustained* and *ODD*-specific **performance by an *ADS* of the entire *DDT*** under routine/normal operation with the expectation that the *DDT fallback-ready user* is receptive to *ADS*-issued requests to intervene, as well as to *DDT* performance-relevant *system failures* in other *vehicle* systems, and will respond appropriately.

SAE International (2021): Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, J3016, online: https://www.sae.org/standards/content/j3016_202104/ (27.05.2021)

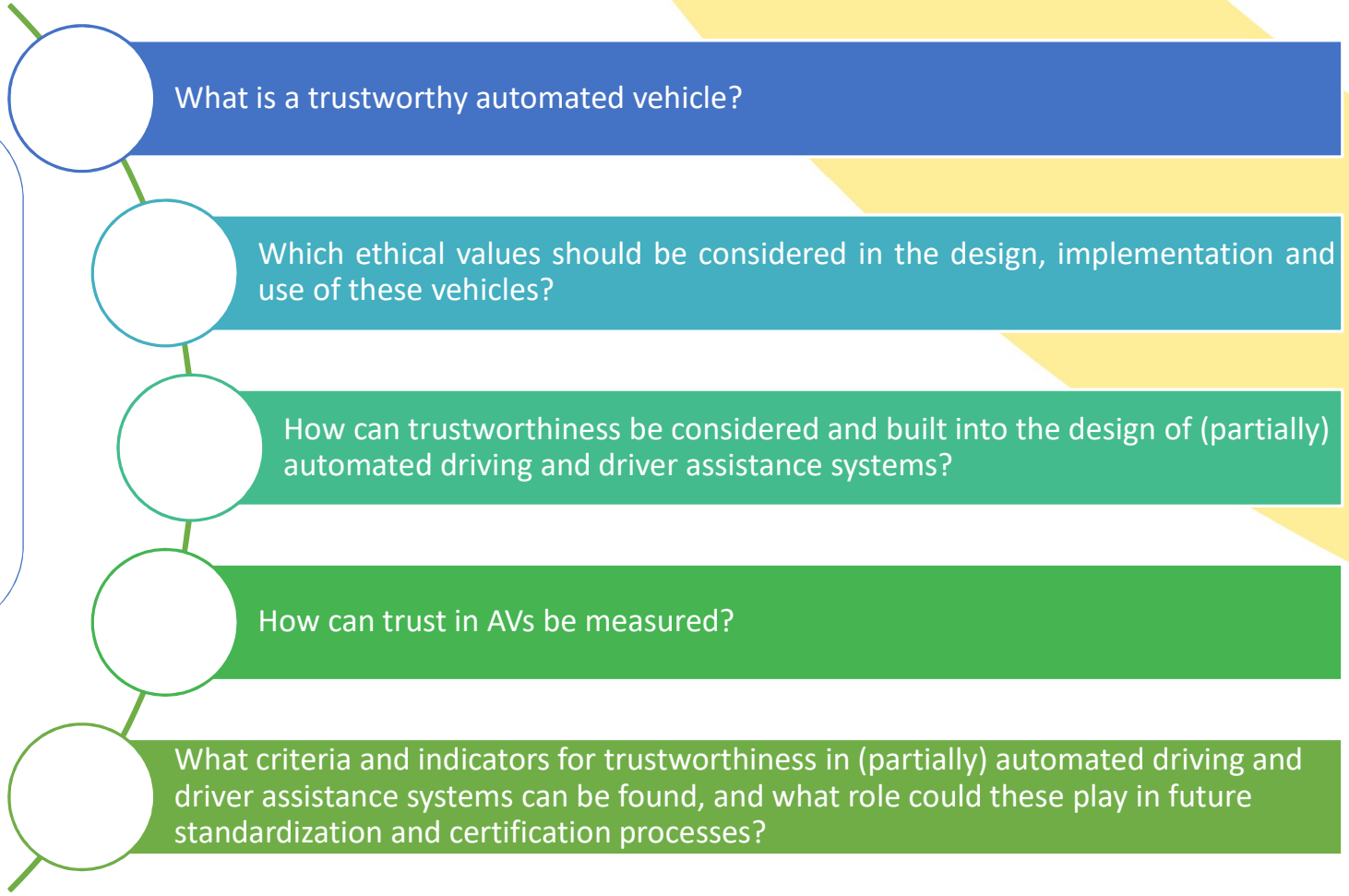
SAE L3 – Conditional Driving Automation

The *sustained* and *ODD*-specific **performance by an *ADS* of the entire *DDT*** under routine/normal operation with the expectation that the *DDT* **fallback-ready user is receptive** to *ADS*-issued requests to intervene, as well as to *DDT* performance-relevant *system failures* in other *vehicle* systems, and will respond appropriately.

SAE International (2021): Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, J3016, online: https://www.sae.org/standards/content/j3016_202104/ (27.05.2021)

Research questions

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



Today…

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

Approach

Pillars

Structure

How to use it?

Approach

Values by design

- Inlcude values throughout the whole lifecycle of the product (design, implementation and use)

Beyond compliance

- Not a mere translation of existing legal frameworks
- What more could be done?

Manufacturers as the main target group

- Other relevant stakeholders: vehicles users, policy makers, NGOs etc.
- VERDI Recommendations

Few observations

- Essential vs.non-essential functions
- Driver vs Fallback-ready user
- Autonomous vs automated vehicles

Essential vs.non-essential functions

- Performance of the driving task
- Vehicle operation

Essential functions

- Additional functions and services
- In-car entertainment

Non-Essential functions

VERDI ♪

☞ **Privacy and data protection**

Driver vs Fallback-ready user

SAE L3 – introduction of a new role – the fallback-ready user

To what extent is a smooth driver-system interaction and transfer of control at all realistic and safe?

If the driver remains legally responsible in the role of FRU, then the question arises, to what extent is the term “fallback-ready user” truly necessary?

Autonomous vs automated vehicles

Increased tendency to ascribe traditional human-related characteristics to technology

What do we mean when we refer to technology as autonomous?

☞ Restrict autonomy to humans?

The pillars of the VERDI Criteria Catalogue

→ 2 H2020 projects

TRUESSEC.eu Criteria Catalogue for trustworthy ICT products and services

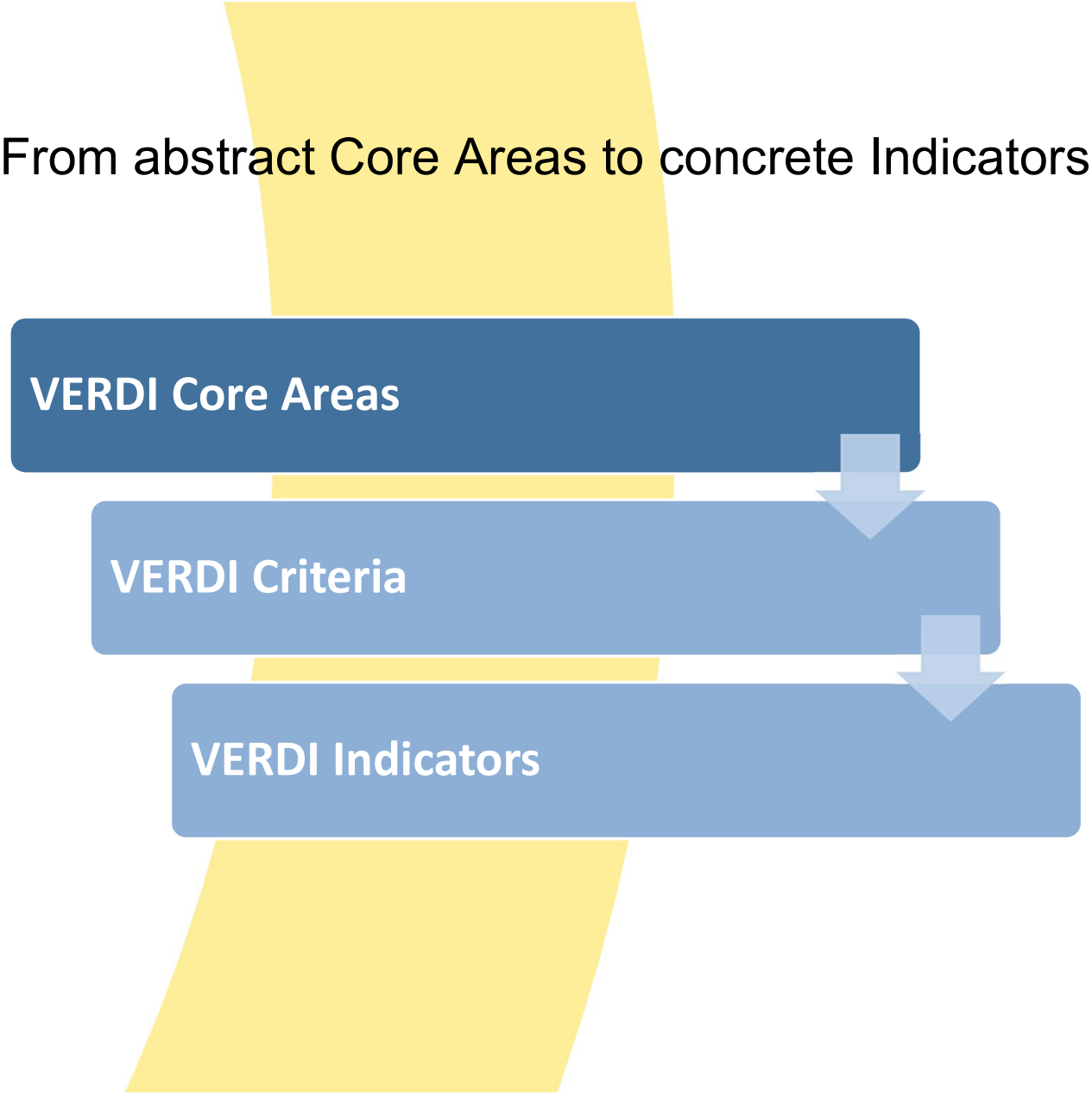
→ 4 Disciplinary Support Studies
(Ethics, Law, Sociology, Psychology)

→ AI HLEG's Ethics Guidelines and ALTAI

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

The structure of the VERDI Criteria Catalogue

From abstract Core Areas to concrete Indicators

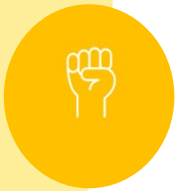


VERDI Core Areas

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**).

Example: Core Area Transparency

Ethics	Law	Sociology	Psychology	VERDI Core Areas SAE L3
Transparency generally stands for openness, accessibility, availability of information. It is considered to be an instrumental value as it enables and supports the realisation of other values such as privacy, accountability, autonomy, safety/security etc. For example, in the context of SAE L3, it relates to providing information to the user about: ○ SAE L3 functionality → safety/security ○ activities with their personal data collected and processed in the context of SAE L3 → privacy, autonomy ○ clear division of roles, duties as well as insight into the decision-making process → accountability etc. It also plays an important role the way the information is delivered to the user (meaningful, relevant, comprehensible etc).	**Transparency** includes legal information duties; these appear especially in consumer law (e.g. informed consent). Moreover, information duties are bindingly defined by the GDPR on EU-level and by the AutomatFahrV on Austrian level. For example, as per the GDPR the data controller (e.g. vehicle manufacturers) is obligated to the respective data subject (e.g. vehicle owner) or as per the AutomatFahrV the test operator is obligated to the BMVIT, the respective governor and road maintainer. Besides, non-binding soft law such as the Austrian Code of Practice includes recommendations and best practices, e.g. the test operator should inform the public, while particularly addressing disabled people and children. Technical standards (e.g. ISO, EN, ÖNORM) are generally not legally binding, unless the parties agree otherwise, or the legal basis bindingly refers to a technical standard either explicitly or generally (i.e. state of the art). In the context automated driving the user has to be informed as long as she is liable. This is the case with SAE L3. When it comes to autonomous driving (SAE L5) information duties will not be (as) necessary.	Covering the (contingent) societal need for controlling a desirable behaviour, process, or structure, transparency can replace the need to trust by tackling uncertainty (synchronous transparency, being able to observe) or necessitates trust (asynchronous transparency, observation is not possible at all time). It is therefore important to define or empirically evaluate what behaviour is socially desired (public opinions), considering possible differences in desirability not just of all stakeholders involved, but between their individual needs, abilities, skills, as well as cultural and social backgrounds. On an institutional level, this means transparency in policy making of compliance, soft-laws, and standardisation, with the public disclosure of the actors and their corresponding interests involved.	System **transparency** relates to the quality with which the system supports the operator in understanding system behaviour, intentions and future goals. The user needs to be able to understand the functionality of the system: what the system is doing, why is it doing that and what it will do next. Transparency guides the development of an adequate mental model of the system functioning and limitations which is central for subsequently forming a calibrated level of trust and an adequate reliance strategy. It needs to be considered what level of transparency is adequate for the specific system. E.g. the driver does not need to have a detailed understanding of the driving automation, but she definitely needs to understand in which contexts it is safe to use the system, in which situations the system might fail and give back control (system limitations) and needs to be able to understand why the system was acting in these specific ways. Furthermore, with vehicles capable of several different levels of automation, it has to be clear for the driver which mode is currently active. Related to that are the responsibilities the driver has within the varying levels of automation (Level 2 vs. Level 3 – what is the driver allowed to do? What means Level 2 and what means Level 3?).	➔ **Transparency** The SAE L3 is provided in line with information duties regarding • the understanding of the system SAE L3 i.e. its functionality (What is it doing/can do?), limitations (What is it not doing/cannot do?) and anticipation (What will it do next?) • dealing with the (personal) data collected and processed for the purposes of L3. In that sense, the understandability of information plays an important role as well, that is, how the information is provided to the user. The above understanding of transparency is predominantly user-centered. Nevertheless, it can be argued that transparency can also be relevant for producers/manufacturers as well as for the society as a whole for the purpose of making system improvements.

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**	1. Minimised Collection, Processing and Use of Personal Data 2. Transparent Processing of Personal Data 3. Privacy Commitment 4. Information Representation 5. Explainability 6. Clear Stipulation of Roles and Duties 7. Feedback and Complaint Management 8. Ability to Redress 9. Statement of Legal Compliance 10. Appropriate Dispute Resolution 11. Established Oversight Mechanisms 12. Secure Infrastructure 13. Vehicle Safety 14. Non-discrimination 15. Avoiding Algorithmic Bias 16. Social and Environmental Responsibility 17. Open Data Approach

VERDI 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 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.

VERDI Indicators

Social and Environmental Responsibility
Mobility has an environmental impact and the potential to lead to unequal treatment. This criterion ensures that negative ecological effects are minimized and that all road users are treated as equal to vehicle users/owners.
VERDI Indicators
1) The manufacturer implements measures to reduce negative environmental impact that go beyond what is required by law and/or other mandatory standards. These include the following: a) The ADS's prime route suggestion is by default the most ecological one. b) The ADS's driving settings are by default environmentally friendly (e.g. slow speed, gasoline saving). c) The ADS's driving style is robust towards other road users, especially regarding vulnerable road users. 2) The ADS's ODD is geographically widely available (e.g. not only in wealthy areas). 3) The ADS is available for a broad range of users in diverse socio-economic situations. 4) The ADS does not favour vehicle users over other road users in critical decisions.

How to use the VERDI Criteria Catalogue?

Manufacturers

Tool for the evaluation of the trustworhiness of partially-automated vehicles

Values to be considered into the design, development and implementation of these vehicles.

Check the adherence to these values at any time.

Value conflicts to be considered

Vehicle users

Evaluate the trustworthiness of their own vehicle and of the vehicle manufacturer

Awareness - raising regarding issues of trustworthiness

Crowd assessment

The VERDI solution:

“The trustworthy European connected car”

- ✓ Protection comes first and before economization
- ✓ Respect user’s privacy and autonomy
- ✓ Focus on transparency
- ✓ Clear division of roles and responsibilities
- ✓ Focus on driving functionality
- ✓ Data cooperation

It is up to all of us…

Thank you for your attention!

hristina.veljanova@uni-graz.at

We work for tomorrow

UNI GRAZ