TRUSTWORTHY AUTONOMOUS VEHICLES: LET’S LEARN FROM LIVE

Axel Hahn, Ingo Stierand, André Bolles

DLR

Why is it **widely accepted** that people get their driver’s license after a **very limited** set of driving lessons – for example, 9 hours in Germany?

AI-generated by GPT-4o

At the same time, at least 50 % of the people say they **don’t trust** autonomous driving.

AI-generated by GPT-4o

What’s the difference?

Well ….
Trust in **natural** intelligence
vs.
trust in **artificial** intelligence

But, what’s really the difference?
Where does trust really come from?

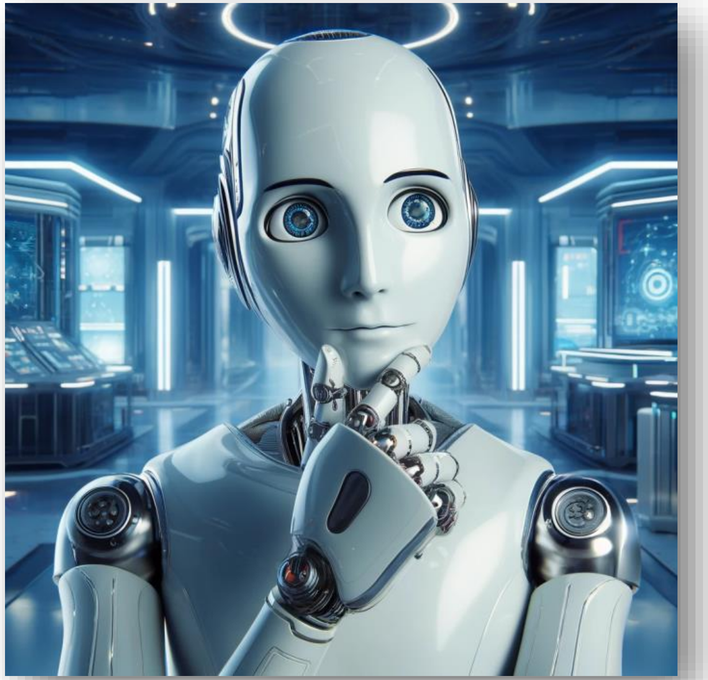
Paper

Motivation

- Autonomous vehicles represent a completely new class of systems
- Substantial changes in our daily lives
- Need for deeper understanding of trust and trustworthiness
  - in these systems
  - in systems engineering to create these systems

AI-generated by GPT-4

Motivation



AI-generated by GPT-4o

- **Uncertainties** will continue to exist within AI-based systems even after full certification.
- **Negative framings and mistrust** overshadow the perception of usefulness.
- The lack of trust is not a matter of **persuasion**, but of **communication**.
- Thus, the role of trust has to be considered **increasingly important** as a “human factor” in systems engineering.

So, let’s have a deeper look into trust. 👀

Deep dive into trust

What the literature says

- Trust **develops** over time, **changes** with experiences and indeed **shapes** our experiences in turn.

- It has **emotional** and **cognitive** facets and fulfils important **normative** and **legal** functions and has **corresponding implications**.

- It has very different **meanings** in different **disciplines** and **research fields**.

Technical trustworthiness

Reliability

Availability

Ethics

Safety

Privacy

Collection of well-established technical properties

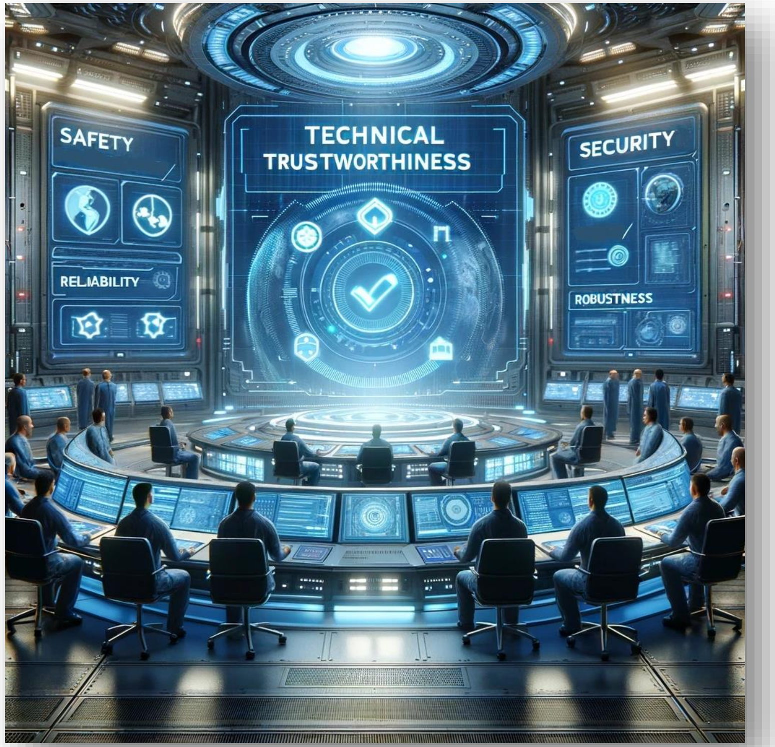
Usability

Robustness

Explainability

Security

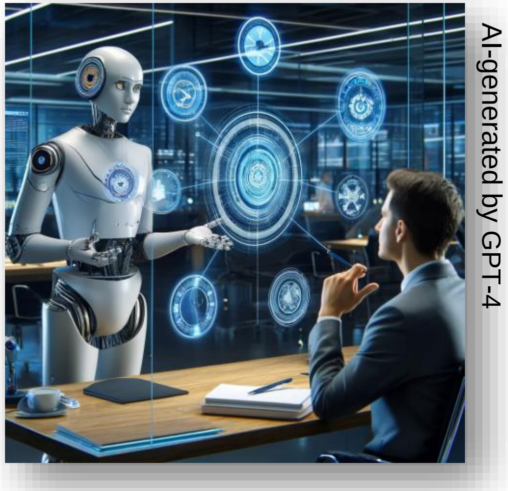
Maintainability



AI-generated by GPT-4

Technical trustworthiness

- Some aspects are not well understood
  - **Explainability**: What is a **good** explanation?
  - Ethical/moral concepts and terms like “**unacceptable**” or “**unreasonable**”
- What is their **relation** to trust?
- How can these aspects be **integrated** into systems engineering? **Should we even do that**?



AI-generated by GPT-4

AI-generated by GPT-4

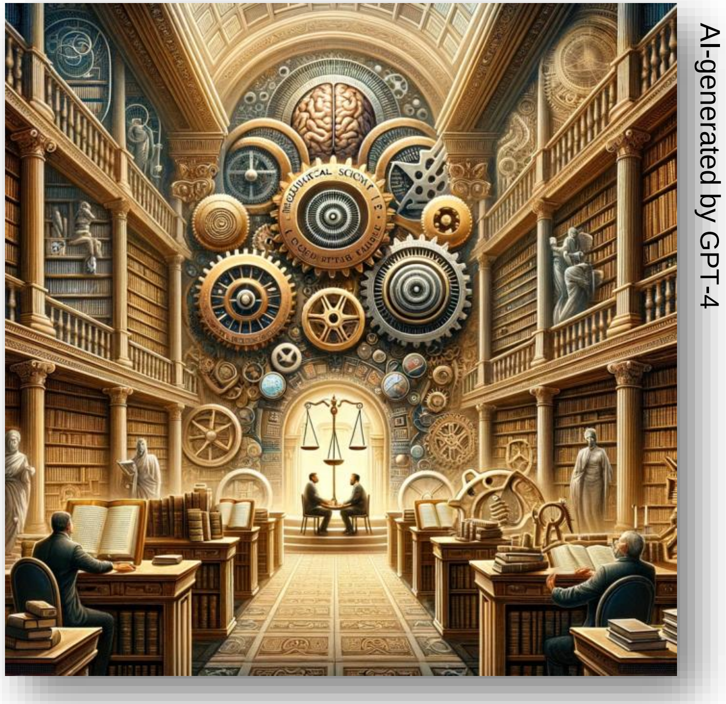
Non-technical perspective

- **Social sciences** and **humanities** also deal with aspects of AI
- Own fields of research and key concepts considered in **interdisciplinary** approaches
- Examples of such fields:
  - “robopsychology”
  - “sozionics”
  - “deep mediatization”

Non-technical perspective

Humanities

- Historically, primarily examined in the context of faith and fidelity.
- Modern views include trust in government, contracts, and contractors.
- Trust is a relation: **A trusts B with respect to C**
- Risk assessment perspective on trust
  - Trust as hope that trustee will prove to be trustworthy
- Motives-based perspective on trust
  - Trustee is motivated to be considered trustworthy



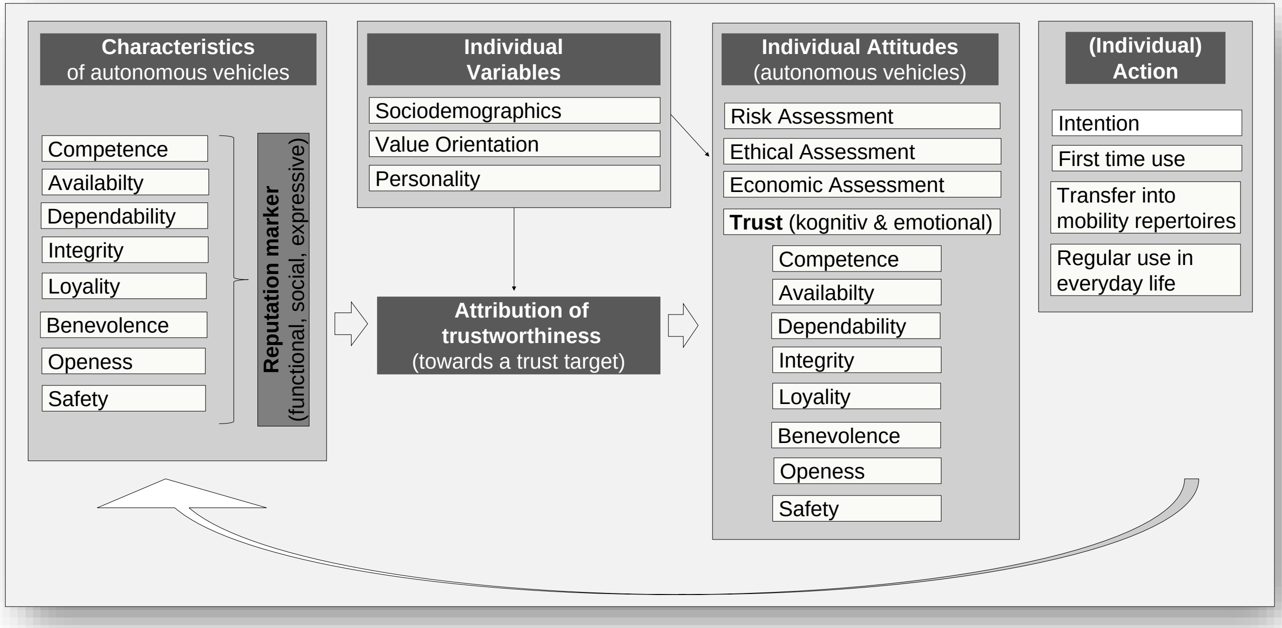
AI-generated by GPT-4

Non-technical perspective

Social sciences

- Focus on relationship between trustor and trustee

- ***“The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor***” (Mayer et al.)



Characteristics
of autonomous vehicles

Competence
Availabilty
Dependability
Integrity
Loyality
Benevolence
Openess
Safety

Reputation marker
(functional, social, expressive)

Individual Variables

Sociodemographics
Value Orientation
Personality

Attribution of trustworthiness
(towards a trust target)

Individual Attitudes
(autonomous vehicles)

Risk Assessment
Ethical Assessment
Economic Assessment
Trust (kognitiv & emotional)
Competence
Availabilty
Dependability
Integrity
Loyality
Benevolence
Openess
Safety

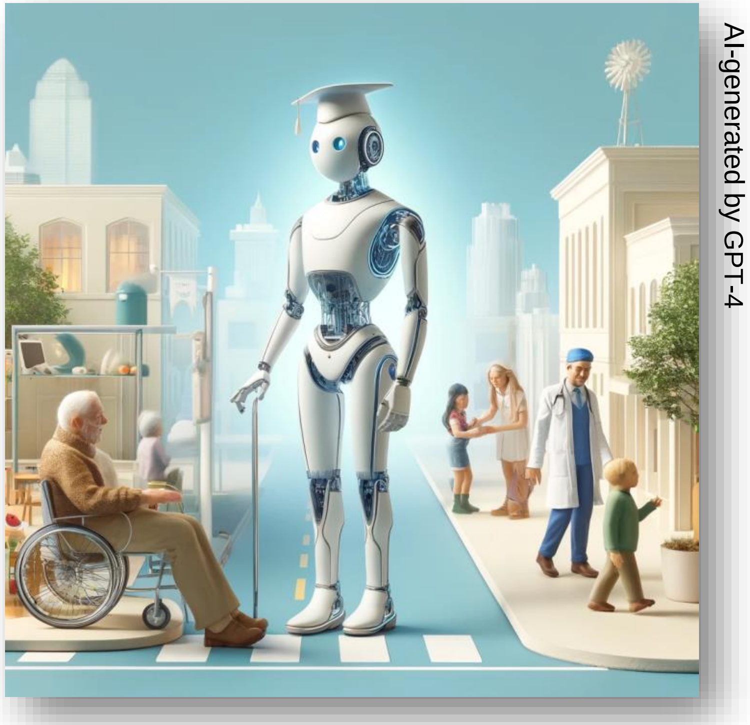
(Individual) Action

Intention
First time use
Transfer into mobility repertoires
Regular use in everyday life

Key ingredients for trust

Identified in humanities and social sciences

- **Abilities**: skills, competencies and characteristics of the system
  - Open question: How to implement this in systems engineering
- **Benevolence**: “good will” of trustee or believe in trustee that he will do good.
  - Objectifiable characteristics/metrics for “good will” are needed for systems engineering.
- **Integrity**: acting according to norms, standards and principles
  - Systems engineering: technical background on standards;
  - Social Background on standards needed

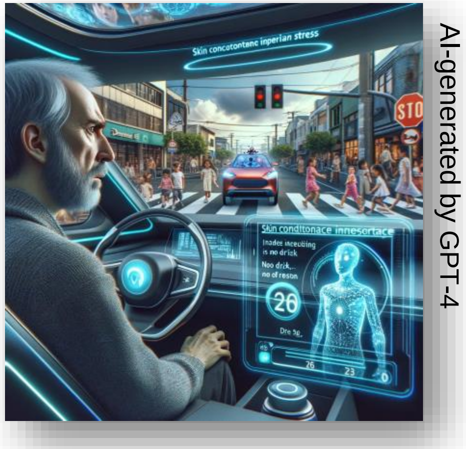


AI-generated by GPT-4

Trust and explainability

Key ingredients are no guarantee for trust,…

- **Example 1:** A person is sitting in an autonomous car. She or he sees a busy area with many children and elderly people on the sidewalks. The car measures increased stress of the passenger. The car reports back that the vehicle has not identified any risk and has been observing the speed limit.
  However, the skin conductance measurement still indicates stress.


Skin concotontanc inperlan stress
STO
Skin conditonace innesortace
26

AI-generated by GPT-4

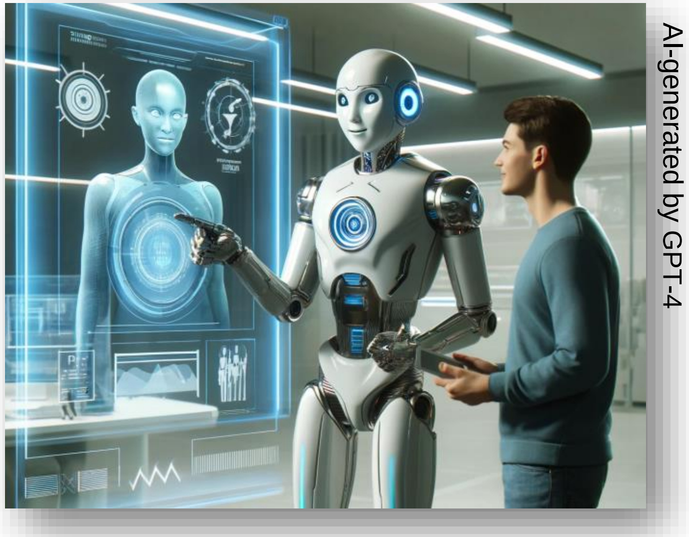
- **Example 2:** An autonomous bus is driving down a country road. The bus passes a bad accident with several injured people. The passengers of the bus are informed that the bus has sent an automatic emergency call.
  However, some passengers are nevertheless very worried about how to deal with the situation and what to do next.

AI-generated by GPT-4

Trust and explainability

… neither does explainability.

- Explanations given in both examples
- However, involved humans are not convinced
- Challenges to be solved by interdisciplinary research
  - **When** do we need **explanations**?
  - **Which** explanation is **sufficient**?
  - **When** is **human** intervention needed for **ethical** and **social** reasons?
  - **What** are metrics to **decide** above and **how** can they be sensed?



AI-generated by GPT-4

Summary

- The role of trust has to be considered increasingly important as “human factor” in systems engineering.

- Key ingredients of trust: abilities, benevolence, integrity and explainability

- Technical and non-technical understanding necessary for implementation
  - Non-technical understanding for defining social and ethical norms
  - Interdisciplinary research to identify corresponding indicators
  - Definition of metrics and sensing mechanisms needed

- More autonomy of systems needs more interdisciplinary research