



**ARTIFICIAL  
INTELLIGENCE**

**MEETS SAFETY AND  
HEALTH AT WORK**

# How standardisation brings AI trustworthiness into practice

Dr Sebastian Hallensleben

Head of Digitalisation & AI at VDE e.V.

Chair CEN-CENELEC JTC 21

Co-Chair OECD ONE.AI WG Risk Assessment

UNESCO Expert Group on AI Ethics



# VDE and AI Trustworthiness

Chairmanship of  
CEN-CENELEC JTC21



Steering group for German  
AI Standardization Roadmap

Convenorship of IEC  
SEG 10 AI Ethics



AI Trust Standard & Label (April  
2022)

Observer in Council of Europe  
Committee on AI



Lead & Conception  
„AI Ethics: From Principles to Practice“  
(April 2020)

OECD ONE.AI Co-Chair  
Classification and risk assessment



German Enquete on AI: Advice to MPs

Federal Ministry of Labour and Social affairs:  
Framework for AI Training Data Quality

Advice to EU-Commission  
DG GROW / DG JUST / DG CNECT

BertelsmannStiftung

ERBERHARD KAREL  
UNIVERSITÄT  
TÜBINGEN INTERNATIONAL CENTER FOR  
ETHICS IN THE SCIENCES AND  
HUMANITIES (IZEW)

ITAS Institut für  
Technikfolgenabschätzung  
und Systemanalyse

H L R I S  
High-Performance Computing Center | Stuttgart

TECHNISCHE UNIVERSITÄT  
KAISERSLAUTERN

iRights.Lab  
Think tanks for the  
digital world





Broader perspective: from AI Ethics to AI Quality

## AI Quality Summit

02. November 2022 near Frankfurt Airport

Registration: [www.ai-q.de](http://www.ai-q.de)

**AIQ** | AI QUALITY &  
TESTING HUB

EUROSHNET 20/10/2022



Hessische Staatskanzlei  
Hessische Ministerin für  
Digitale Strategie und Entwicklung



digitales.hessen

**VDE**



# The big challenge

**Operationalise** AI Ethics with an approach ...

- ... that is viable for industry, regulators and consumers / citizens

- ... and that makes ethics measurable and enforceable



# Why standardisation is the right approach

**Standardisation =**

1. Building consensus among all relevant stakeholders
2. Formulating this consensus  
in a concrete, specific, practically useful way



# How to handle AI Ethics through standardisation

consensus unlikely

## Explicit ethical rules

(e.g. „Child more important than old person“, „100 severely injured better than 1 dead“)

viable, flexible and strong

## Standardised description of ethical aspects of systems

(e.g. „Privacy A, Transparency D, Fairness B“)



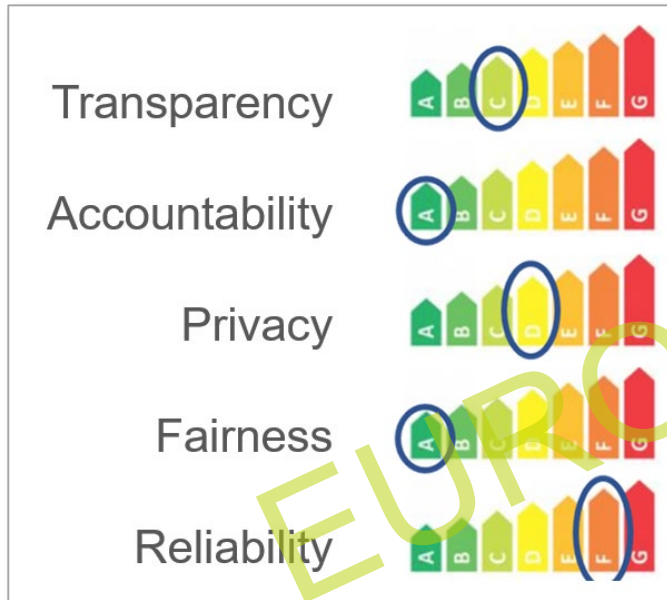
viable but limited on its own

## Only processes and structures for decisions about ethics

(e.g. ethics boards in companies)



Approach: A standardised „label“ / „short datasheet“ that can be attached to AI products



- ✓ provides **positive differentiation** in the marketplace
- ✓ ensures **fair competition**
- ✓ promotes consistency with **organisational and societal values**
- ✓ facilitates **compliance** with regulation
- ✓ supports policymakers in **minimising red tape**

 Compatibility with the AI Act  
(Art. 40, 41)  
Consideration of existing  
standards and GDPR





## European and international standardization



CEN-CENELEC Focus Group for Artificial Intelligence



Roadmap report  
October 2020

- IEC SEG 10 Ethics in autonomous and artificial intelligence applications



Final report July 2021



AI Ethics Impact Group  
[www.ai-ethics-impact.org](http://www.ai-ethics-impact.org)

# Bertelsmann Stiftung

EBERHARD KARLS  
UNIVERSITÄT  
TÜBINGEN



INTERNATIONAL CENTER FOR  
ETHICS IN THE SCIENCES AND  
HUMANITIES (IZEW)



H L R I S

High-Performance Computing Center | Stuttgart



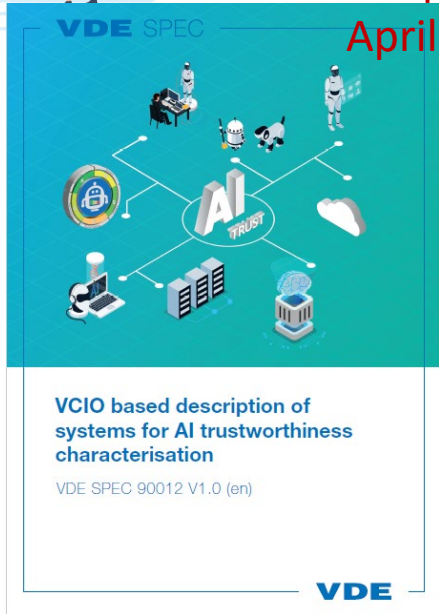
# VDE





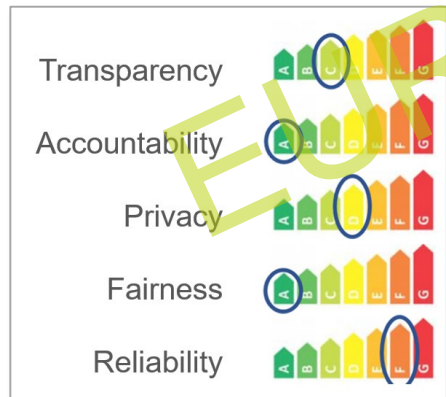
# Comprehensive consortial standard 2021/22

Version 1 published in April 2022



Describes the characteristics of an AI product with regards to:

Transparency – Accountability – Privacy – Fairness – Reliability



Digital Trust Forum





Questions:

1. Which categories do we include?
2. ...
3. ...

EUROSHINET 20/10/2022



privacy protection	17
accountability	17
fairness, non-discrimination, justice	17
transparency, openness	15
safety, cybersecurity	15
common good, sustainability, well-being	15
human oversight, control, auditing	12
explainability, interpretability	10
solidarity, inclusion, social cohesion	10
science-policy link	10
legislative framework, legal status of AI systems	9
responsible/intensified research funding	8
public awareness, education about AI and its risks	8
future of employment	8
dual-use problem, military, AI arms race	7
field-specific deliberations (health, military, mobility etc.)	7
human autonomy	7
diversity in the field of AI	6
certification for AI products	4
cultural differences in the ethically aligned design of AI systems	2
protection of whistleblowers	2
hidden costs (labeling, clickwork, content moderation, energy, resources)	1

- transparency
- justice
- accountability
- privacy
- reliability/safety
- environmental sustainability

EUROSHNET 2011012022



### Questions:

1. Which categories do we include?
2. How can we measure transparency, accountability, etc.?
3. ...



Based on T1.1  
(e.g. from D)

## Transparency



### Negative anchor indicator

"necessary condition"

Prerequisite for T1.2 and T1.3.  
Minimum requirement (e.g. E-G)

### Positive anchor indicator

"sufficient condition"

The fulfilment of one indicator can substitute the fulfilment of one or more other indicators.

## T1. Disclosure of origin of data sets

## T2. Accessibility

### T1.1

Is the origin of the data documented?

### T1.2

Is it for each individual use plausible, which data is being used?

### T1.3

Are the characteristics of the training data set documented and disclosed?  
Are the data sheets to the data sets comprehensive?

### T2.1

Are the modes of interpretability oriented toward the needs of the target groups and developed with them?

### T2.1

Are the modes of interpretability in their target group specific form also intelligible for the target groups?

Yes, comprehensive logging of all training and operating data, version control of data sets etc.

Yes, the use of data and the individual application are intelligible

Yes and the data sheets are comprehensive

Yes

Yes, the modes of interpretability have been tested with target groups for intelligibility

Yes, logging and version control through an intermediary (e.g. data supplier)

Yes, it is intelligible on an abstract, not case specific level, which data is being used

Yes, but the data sheet contains few or missing information

Yes, but without participation of the target groups

Yes, target groups can complain or ask when they do not understand a mode of interpretability

No logging. Data used is not controlled or documented in any way

No, but a summary on the data usage is available

No

Yes, but only toward one target group

No

No

### Score indicators

Build on anchor indicators.  
Scoring of the score indicators are added and averaged to determine the level of the label

Yes, but only one mode of interpretability is developed without regard to target groups' needs



### Questions:

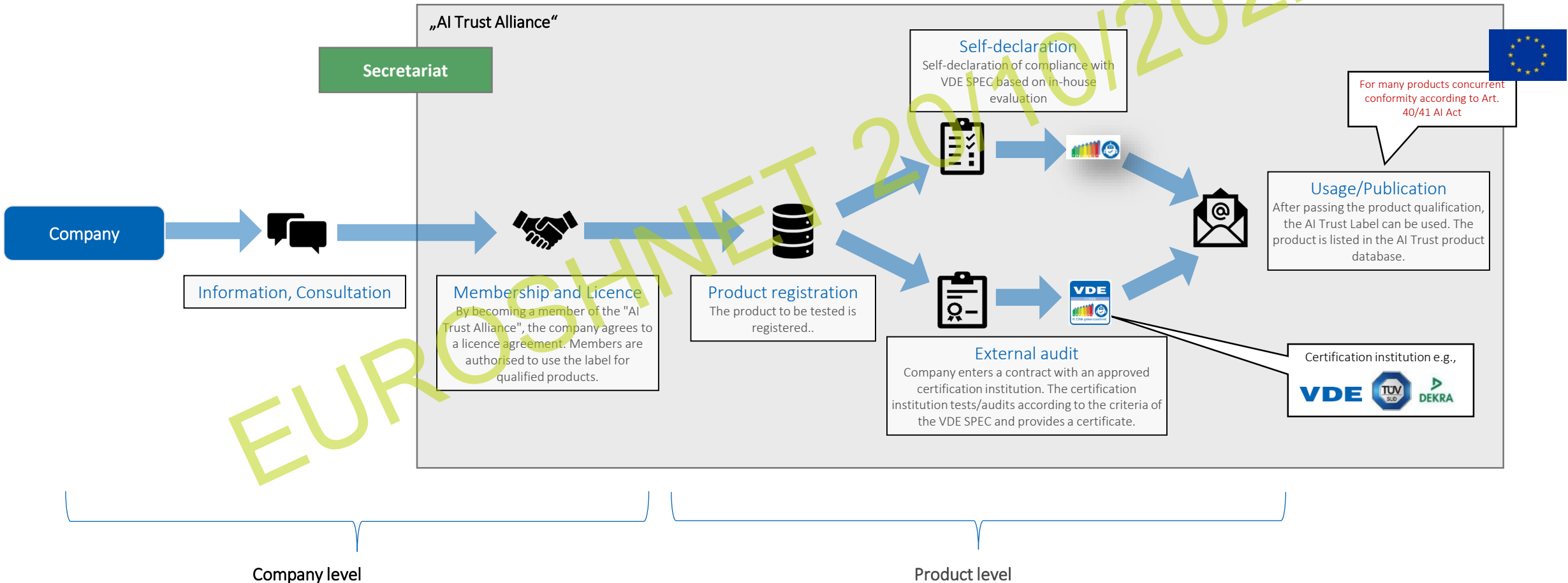
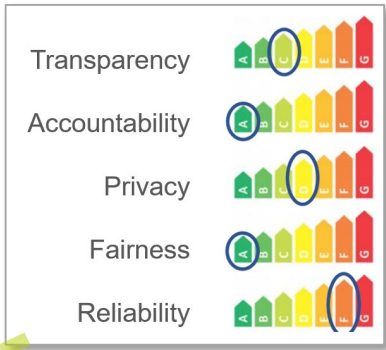
1. Which categories do we include?
2. How can we measure transparency, accountability, etc.?
3. What levels are acceptable in a given application?







# AI Trust Standard & Label from a company perspective



EUROPEAN DIGITAL CHARTER 2021/01/2022

# Towards a European approach



**VDE**

Transparency	
Accountability	
Privacy	
Fairness	
Reliability	

Combining complementary work metrics – tools – governance

