Operationalising Trust: The AI Trust Standard & Label with the VCIO framework

Dr Sebastian Hallensleben

Head of Digitalisation & AI at VDE e.V.

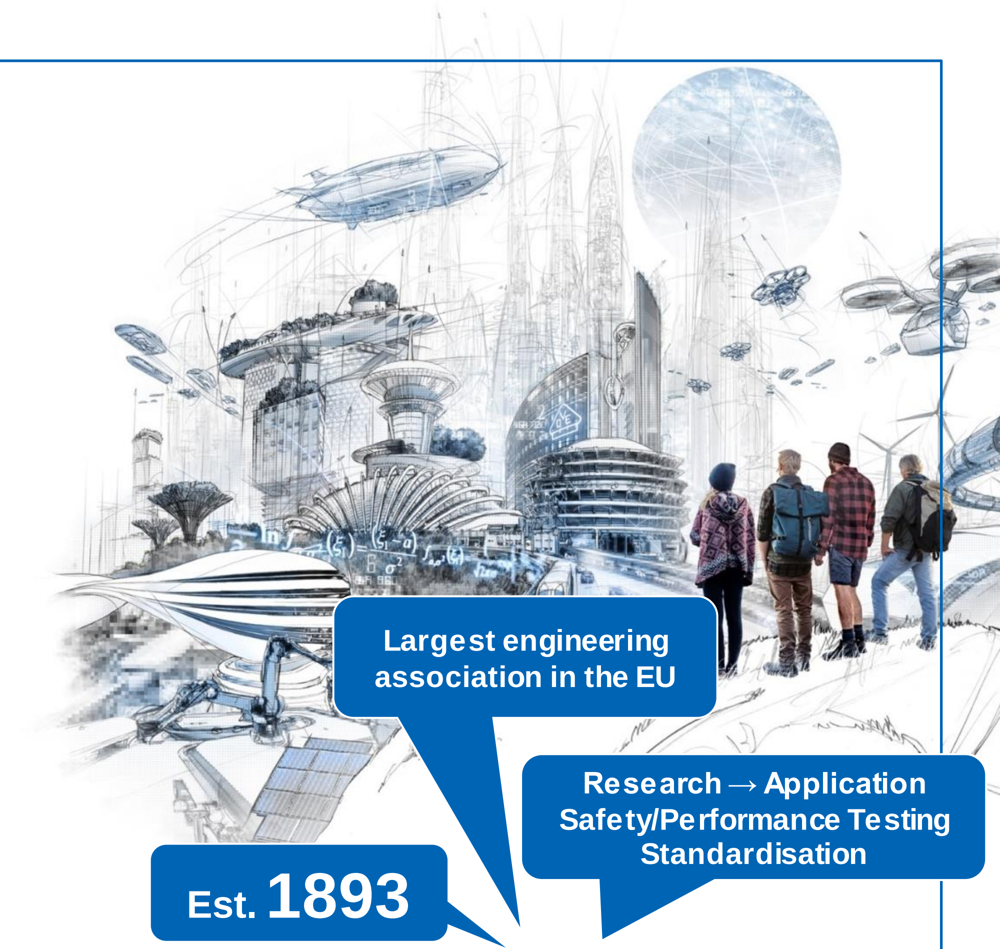
Chair CEN-CENELEC JTC 21

Co-Chair OECD ONE.AI WG Risk & Accountability

UNESCO Expert Group on AI Ethics

HLRS Summer School

2023-07-26



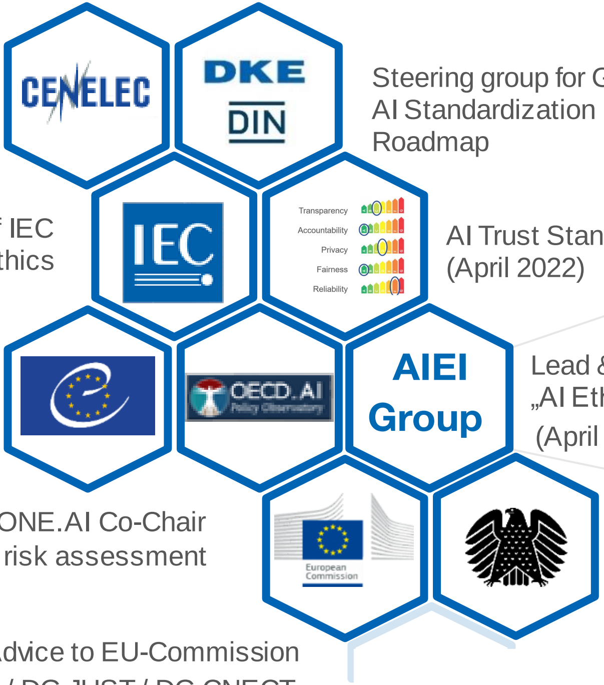
Largest engineering association in the EU

Research → Application
Safety/Performance Testing
Standardisation

Est. 1893

VDE

VDE and AI Ethics



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

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

BertelsmannStiftung

Eberhard Karls Universität Tübingen

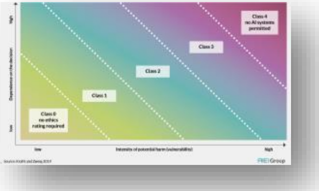
International Center for Ethics in the Sciences and Humanities (IZEW)

ITAS Institut für Technikfolgenabschätzung und Systemanalyse

HLRS High-Performance Computing Center | Stuttgart

Technische Universität Kaiserslautern

iRights.Lab Think tank for the digital world



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 enforcable**

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

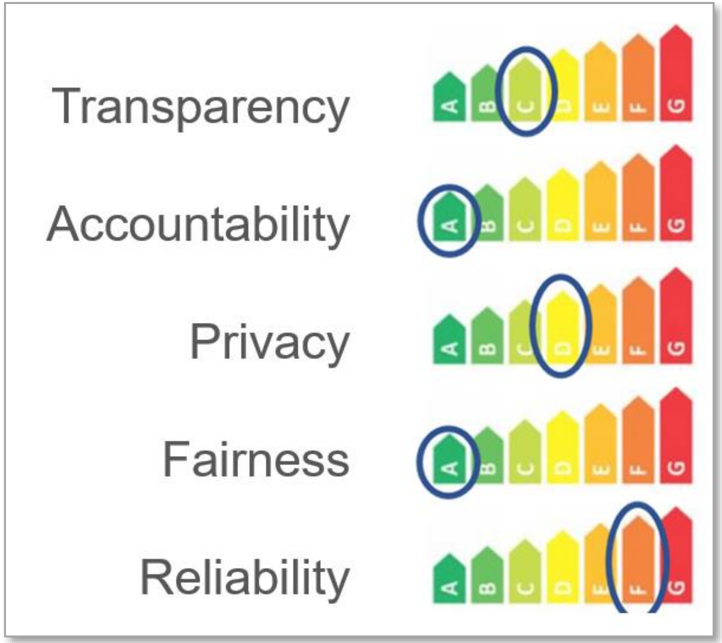
(e.g. „Privacy A Transparency D, Fairness B“)

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



Standard:
describes the metric for quantifying characteristics

Label:
communicates the adherence to the standard in a concise way

European and international standardization

cen

CENELEC

CEN-CENELEC Focus Group for Artificial Intelligence

Roadmap report October 2020

- IEC SEG 10 Ethics in autonomous and artificial intelligence applications

IEC

Final report July 2021

AI Ethics Impact Group www.ai-ethics-impact.org

BertelsmannStiftung

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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

ITAS Institut für Technikfolgenabschätzung und Systemanalyse

HLRS High-Performance Computing Center | Stuttgart

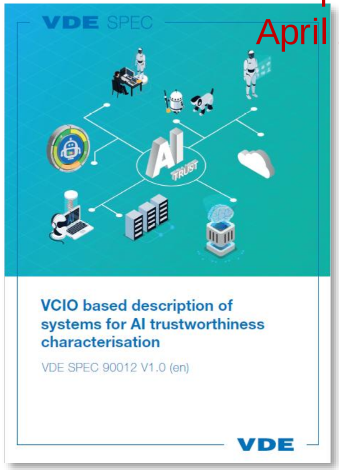
TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

iRights.Lab Think tank for the digital world

VDE

Comprehensive consortial standard 2021/22

Version 1 published in April 2022



VDE SPEC

VCIO based description of systems for AI trustworthiness characterisation

VDE SPEC 90012 V1.0 (en)

VDE

Describes the characteristics of an AI product with regards to:

Transparency – Accountability – Privacy – Fairness – Reliability

BOSCH

SIEMENS

TECHNISCHE UNIVERSITÄT DARMSTADT

SAP

TÜV SÜD

VDE

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
INTERNATIONALES ZENTRUM FÜR ETHIK IN DEN WISSENSCHAFTEN (IZEW)

Digital Trust Forum

BASF
We create chemistry

Ferdinand-Steinbeis-Institut

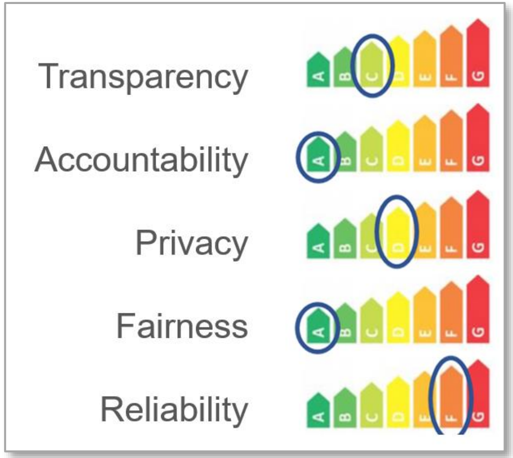
iRights.Lab

KIT
Karlsruher Institut für Technologie

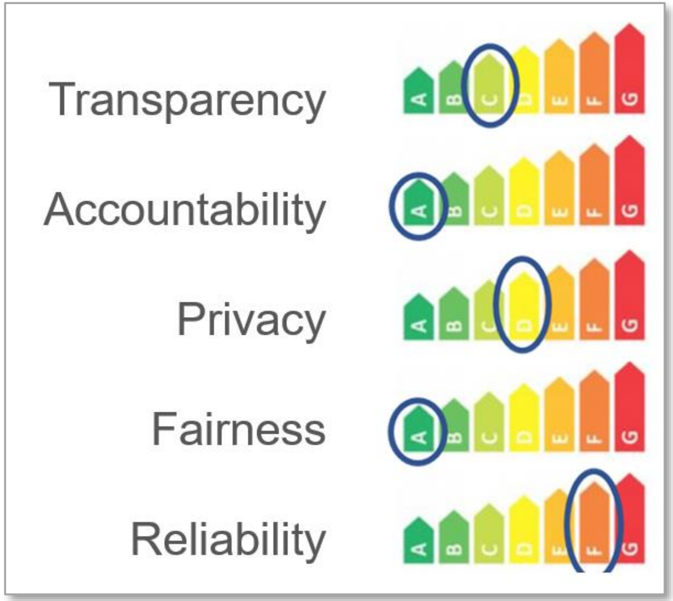
VDE

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

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



- ✓ 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**

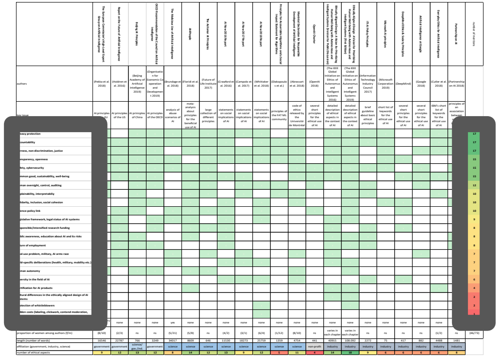


Questions:

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

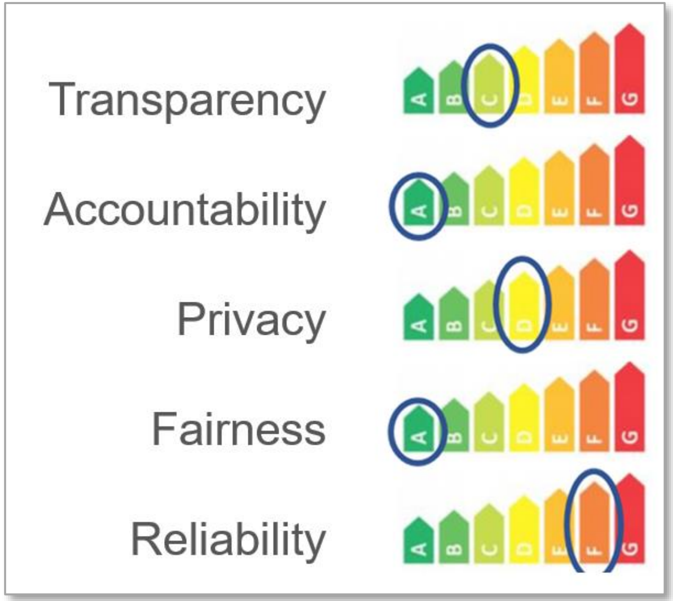
Meta analysis of position papers on AI ethics principles

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, interpretabiliy	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, contend moderation, energy, resources)	1



	The European Commission's High-Level Expert Group on Artificial Intelligence	Report on the Future of Artificial Intelligence	Beijing AI Principles	OECD Recommendation of the Council on Artificial Intelligence	The Malicious Use of Artificial Intelligence	AI4People	The Asilomar AI Principles	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Montréal Declaration for Responsible Development of Artificial Intelligence	OpenAI Charter	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version for Public Discussion)	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition)	ITI AI Policy Principles	Microsoft AI principles	DeepMind Ethics & Society Principles	Artificial Intelligence at Google	Everyday Ethics for Artificial Intelligence	Partnership on AI	number of mentions
authors	(Pekka et al. 2018)	(Holdren et al. 2016)	(Beijing Academy of Artificial Intelligence 2019)	(Organisation for Economic Co-operation and Development 2019)	(Brundage et al. 2018)	(Floridi et al. 2018)	(Future of Life Institute 2017)	(Crawford et al. 2016)	(Campolo et al. 2017)	(Whittaker et al. 2018)	(Diakopoulos et al.)	(Abrassart et al. 2018)	(OpenAI 2018)	(The IEEE Global Initiative on Ethics of Autonomus and Intelligent Systems 2016)	(The IEEE Global Initiative on Ethics of Autonomus and Intelligent Systems 2019)	(Information Technology Industry Council 2017)	(Microsoft Corporation 2019)	(DeepMind)	(Google 2018)	(Cutler et al. 2018)	(Partnership on AI 2018)	
key issue	AI principles of the EU	AI principles of the US	AI principles of China	AI principles of the OECD	analysis of abuse scenarios of AI	meta-analysis about principles for the beneficial use of AI	large collection of different principles	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	principles of the FAT ML community	code of ethics released by the Université de Montréal	several short principles for the ethical use of AI	detailed description of ethical aspects in the context of AI	detailed description of ethical aspects in the context of AI	brief guideline about basic ethical principles	short list of keywords for the ethical use of AI	several short principles for the ethical use of AI	several short principles for the ethical use of AI	IBM's short list of keywords for the ethical use of AI	principles of an association between [illegible]	
[illegible]	[illegible]	none	none	none	yes	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	[illegible]	
proportion of women among authors (f/m)	(8/10)	(2/3)	ns	ns	(5/21)	(5/8)	ns	(4/2)	(3/1)	(6/4)	(1/12)	(8/10)	ns	varies in each chapter	varies in each chapter	ns	ns	ns	ns	(1/2)	ns	(46/73)
length (number of words)	16546	22787	766	3249	34017	8609	646	11530	18273	25759	1359	4754	441	40915	108.092	2272	75	417	882	4488	1481	
affiliation (government, industry, science)	government	government	science/ gov./ind.	government	science	science	science	science	science	science	science	science	non-profit	industry	industry	industry	industry	industry	industry	industry	industry	
number of ethical aspects	9	12	13	12	8	14	12	13	9	12	5	11	4	14	18	9	6	6	6	6	8	

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



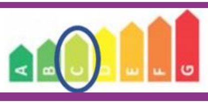
Questions:

1. **Which categories do we include?**

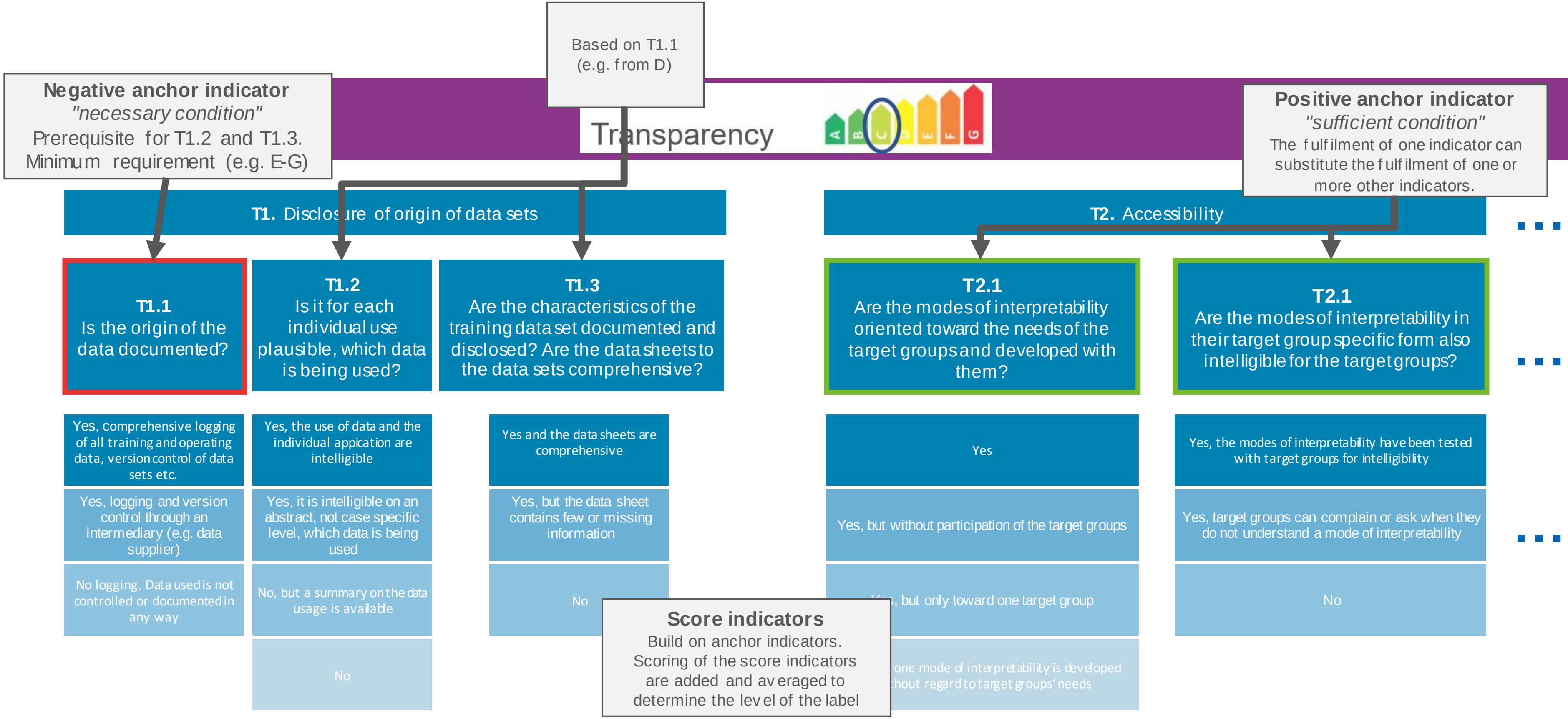
2. **How can we measure transparency, accountability, etc.?**

3. **…**

Transparency



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 appication 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		No, only one mode of interpretability is developed without regard to target groups' needs	



Negative anchor indicator
"necessary condition"
Prerequisite for T1.2 and T1.3.
Minimum requirement (e.g. E-G)

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

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 appication 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		one mode of interpretability is developed thout regard to target groups’ needs	

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

VDE

TRANSPARENCY

T1
Documentation of data sets

T1.1 - Is the data's origin documented?

T1.2 - Are the characteristics of data sets analyzed and documented?

T2
Documentation about the AI systems operation

T2.1 - Are the characteristics of the AI system(s) documented?

T2.2 - Are the characteristics of the AI application documented?

T3
Intelligibility

T3.1 - Have the most intelligible AI models/ systems been selected that can fulfil the application purpose?

T3.2 - What degree of explainability including a regarding documentation is provided?

T3.3 - Are the explanations of the AI system/application outcome designed in a way that adequately informs the affected persons?

T3.4 - Is the interface of the AI system/application designed in a way that adequately informs the user groups about the outcomes and mechanisms?

T4
Accessibility (outside of relevant authorities)

T4.1 - Who has access to the AI system and the AI application?

T4.2 - Who has access to the datasets?

T4.3 - Who has access to the documentation regarding the AI system/application and its data?

T4.4 - Who can see which data attributes (including pre-processing) were used as an input for the AI system/application to generate its output?

PRIVACY

P1
Process for processing of data

P1.1- Does the organization comply to the GDPR?

P2
Protection of personal data (AI related)

P2.1- Which grade of anonymity has the used data?

P2.2 - Is it assured that no personal data can be extracted from the AI System?

P2.3 - What measures have been taken, to prevent attacks on the AI system and application with the aim to inferred data/information?

P3
Consent-Process, information and influence for users and affected

P3.1 - Is the privacy impact assessment presented in the consent process?

P3.2 - Is the privacy impact assessment accessible for affected Persons?

P3.3 - Can affected persons review and rectify data concerning them?

P3.4 - Design of the consent-process

ACCOUNTABILITY

A1
Processes in life cycle to ensure accountability

A1.1 - How detailed is the process of data collection and management logged/recorded and how easily can relevant stakeholders access it?

A1.2 - Are the development and training process logged/recorded?

A1.3 - Is the traceability of the system-composition (including soft- and hardware-composition and components) guaranteed?

A1.4 - Are systems with a learning component monitored in their interaction with their environment throughout the runtime?

A2
Corporate/institutional liability (retrospective)

A2.1 - Is there a defined channel for giving feedback and obtain information about system characteristics?

A2.2 - How "easy" is the access to the feedback channels?

A3
Responsible Human Oversight

Human in Command (Control):

A3.1 - Is the user expertise needed to judge the results of the AI system to avoid overconfidence defined?"

A3.2 - Which effort is needed to understand and interact with the AI system? (depending on the application context)

Human in the Loop:

A3.3 - Which measures are taken to ensure that the AI system does not affect human autonomy by interfering with the operator’s decision-making process in an unintended way?"

A3.4 - Is the human takeover of the system designed so that the user understands the current state of the system and can therefore take over quickly?“

Human on the Loop:

A3.6 - Does the system makes the decision parameters transparent and allows post-hoc changes?"

FAIRNESS

F1
Assuring fairness during development

F1.1 - Are all entities impacted and/or influenced by the system considered?

F1.2 - Are target groups defined?

F1.3 - Are there marginalised entities within the target group and does risk arise for them being marginalised?

F1.4 - Is there a commitment to a fairness definition that considers F1.2 and F1.3.?

F1.5 - Are metrics to track/evaluate fairness with respect to F1.2 and F1.4 in place?

F1.7 - Has the data been analysed for potential harmful, unintended biases with regard to F1.4 and F1.5?

F1.8 - Have trade-offs between fairness and other objectives been identified, assessed and justified?

F2
Working and supply chain conditions

F2.1 - Are the working conditions of external persons involved in the labelling process evaluated?"

F2.2 - Is the supply chain monitored to evaluate working conditions and to prevent human rights violation and child labour?

F3
Ecological sustain development

F3.1 - Are data centres or servers,which are used for developing, supplied with renewable energy?

F3.2 - Is a report available detailing of energy consumption during training of the AI system?

F3.3 - How is the disposal of electronic waste processed?

RELIABILITY

R1
Robustness & reliability qua design

R1.1 - Is the operational design domain of the AI system/application clearly defined and documented?

R1.2 - Was ensured, that the quality and quantity of the data fit to the intended purpose and Operational Design Domain?

R1.3 - Was the quality of the development of the AI systems ensured?

R1.4 - Is the system robust against varying environments (i.e. distribution shift) and outliers?

R1.5 - Are all possible risks assessed and the harms the system could have classified (e.g. life and health, violation of rights etc.)?

R1.6 - Are measures in place to ensure the integrity, robustness and overall security of the AI system/application against potential attacks over its life cycle?

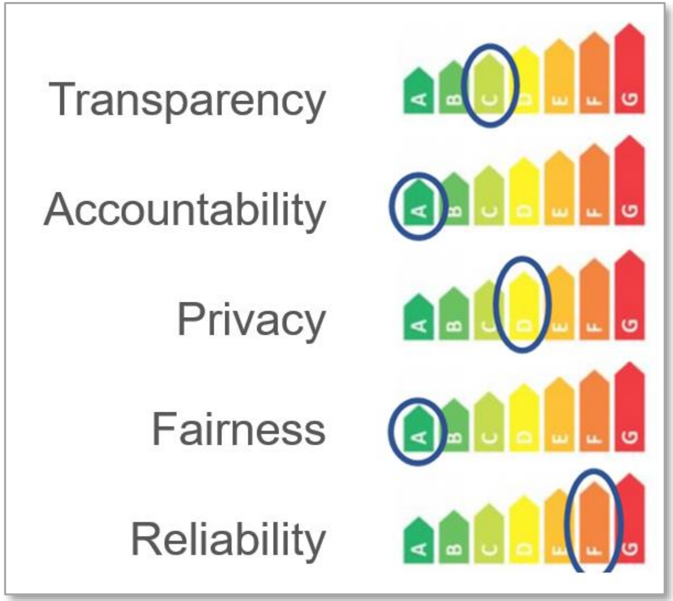
R1.7 - Did you inform end-users of the duration of security coverage and updates? What length is the expected timeframe within which you provide security updates for the AI system?

R1.8 - Are technical documentations documented, including standards, that need to be applied by the AI system/application?

R2
Robustness & reliability in operation

R2.1 - Is the applied AI lifecycle management robust to changes in the operational domain?

R2.2 - Is a failure mitigation strategy for the AI-based system in place?



Questions:

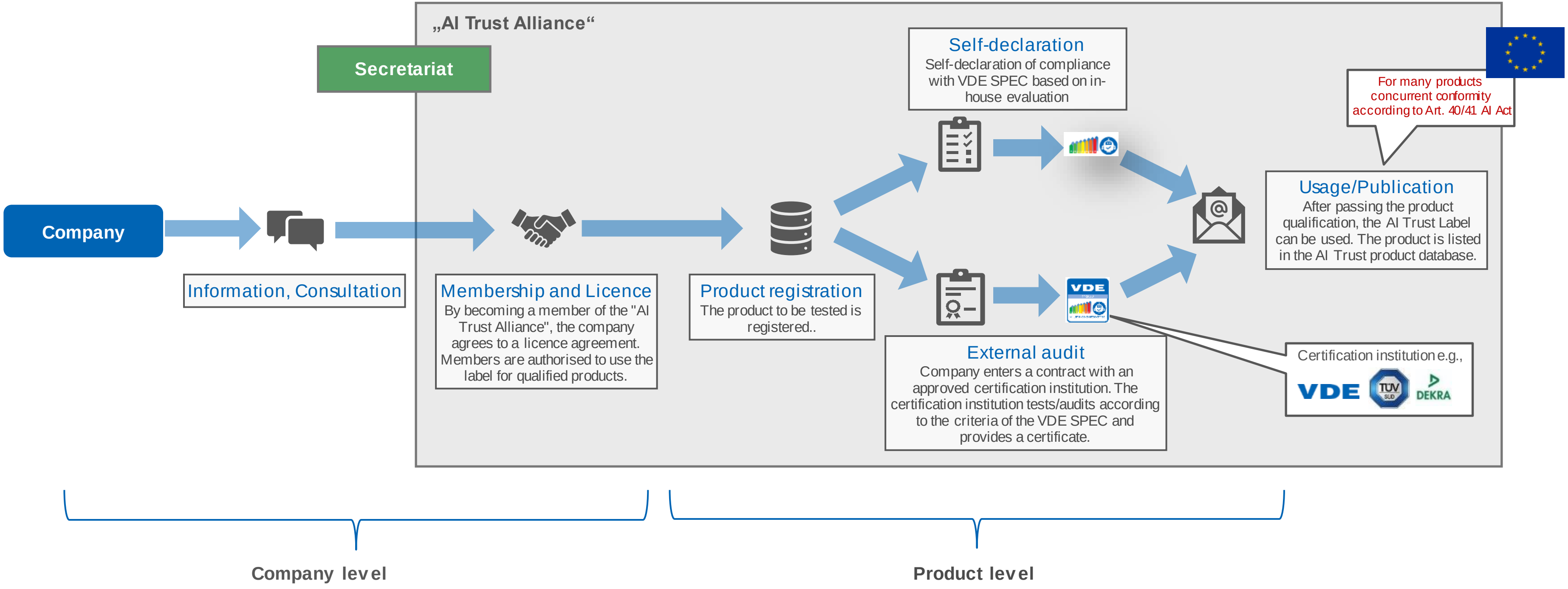
1. Which categories do we include?

2. How can we measure transparency, accountability, etc.?

3. What levels are acceptable in a given application?

… this is a political decision, taken differently in every jurisdiction

AI Trust Standard & Label from a company perspective



„AI Trust Alliance“

Secretariat

Company

Information, Consultation

Membership and Licence
By becoming a member of the "AI Trust Alliance", the company agrees to a licence agreement. Members are authorised to use the label for qualified products.

Product registration
The product to be tested is registered..

Self-declaration
Self-declaration of compliance with VDE SPEC based on in-house evaluation

External audit
Company enters a contract with an approved certification institution. The certification institution tests/audits according to the criteria of the VDE SPEC and provides a certificate.

Certification institution e.g.,
VDE TÜV SÜD DEKRA

Usage/Publication
After passing the product qualification, the AI Trust Label can be used. The product is listed in the AI Trust product database.

For many products concurrent conformity according to Art. 40/41 AI Act

Company level

Product level

Towards a European+ approach

VDE

Confiance.ai

... ...

Combining complementary work
metrics – tools – governance

Cooperation Germany/France
announced October 2022,

further consolidation ongoing
⇒ **AI Trust Alliance**

For **measuring** product characteristics - STANDARDS -
For **communicating** product characteristics - LABEL(S) -
For **proving** that standards are followed and labels are justified - CERTIFICATION / AUDITING -
For **implementing** the label and **achieving** good ratings - TOOLS / AUTOMATION -

Input

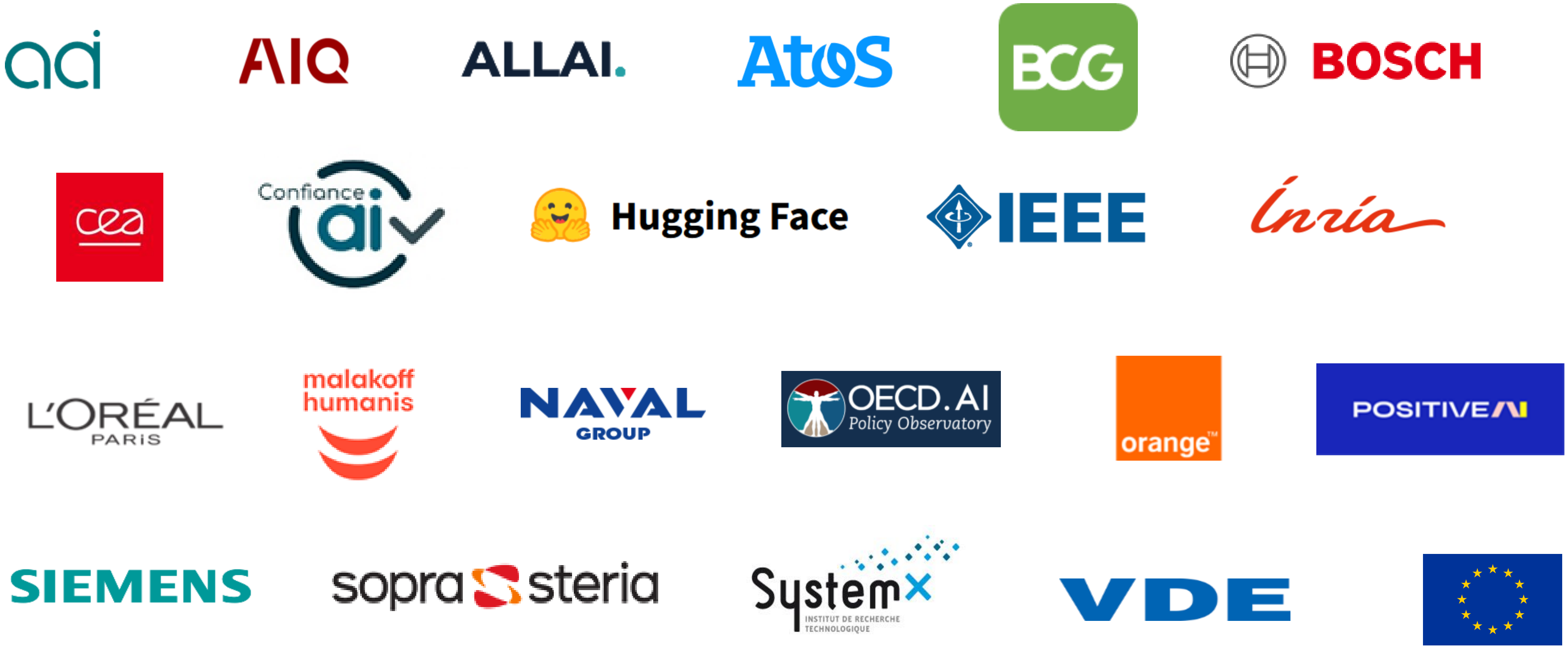
Interoperability

cen

CENELEC

JTC21

AI Trust Alliance (under construction)



adi
AIQ
ALLAI.
AtoS
BCG
BOSCH
cea
Confiance ai
Hugging Face
IEEE
inria
L'ORÉAL PARIS
malakoff humanis
NAVAL GROUP
OECD.AI Policy Observatory
orange™
POSITIVE AI
SIEMENS
sopra steria
SystemX INSTITUT DE RECHERCHE TECHNOLOGIQUE
VDE