EUROSHNET

ARTIFICIAL INTELLIGENCE

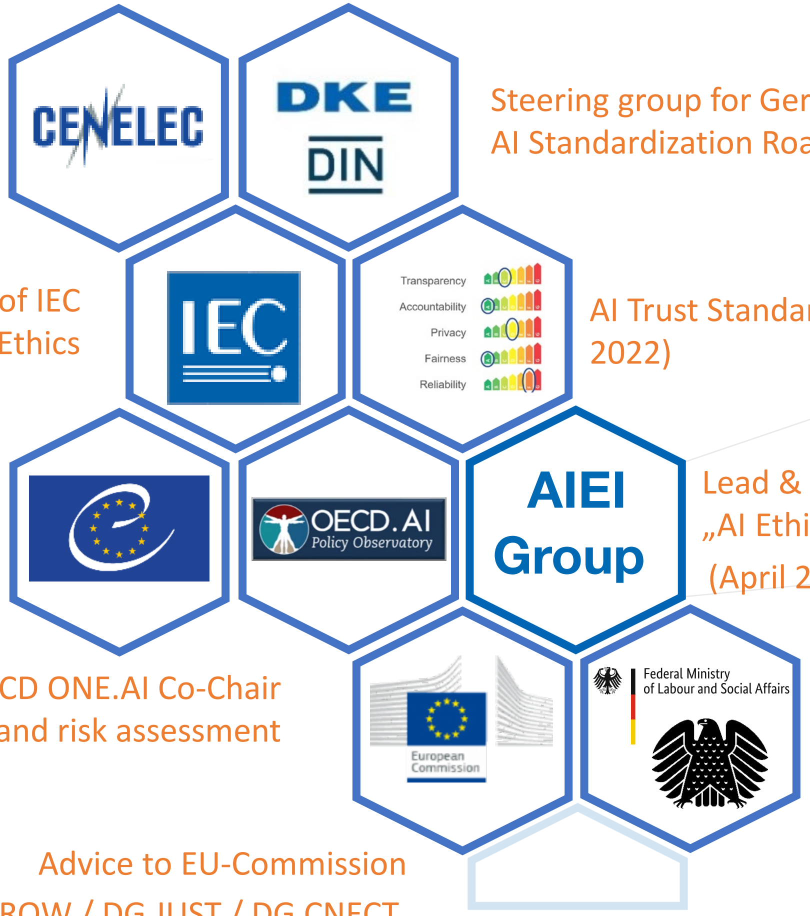
MEETS SAFETY AND HEALTH AT WORK

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)

Bertelsmann Stiftung
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

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

Broader perspective: from AI Ethics to AI Quality

AI Quality Summit

02. November 2022 near Frankfurt Airport

Registration: www.ai-q.de

AIQ | AI QUALITY & TESTING HUB

HESSEN

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 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“)

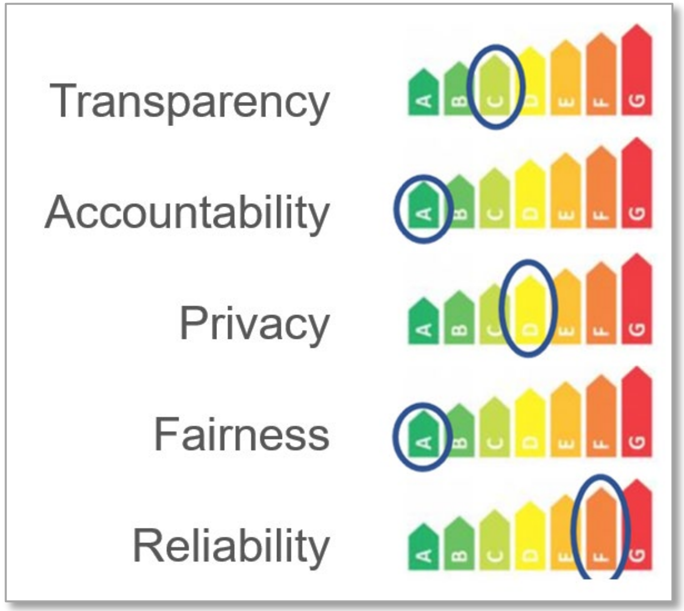
viable but limited on its own

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

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**

VDE

www.linkedin.com/in/sebastianhallensleben sebastian.hallensleben@vde.com

European and international standardization

cen

CENELEC

CEN-CENELEC Focus Group for Artificial Intelligence

Roadmap report October 2020

IEC

- IEC SEG 10 Ethics in autonomous and artificial intelligence applications

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

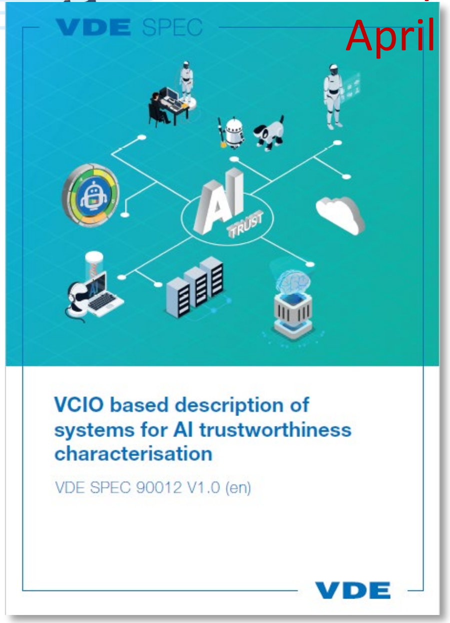
TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

iRights.Lab
Think tank for the digital world

VDE

Comprehensive consortial standard 2021/22

Version 1 published in April 2022

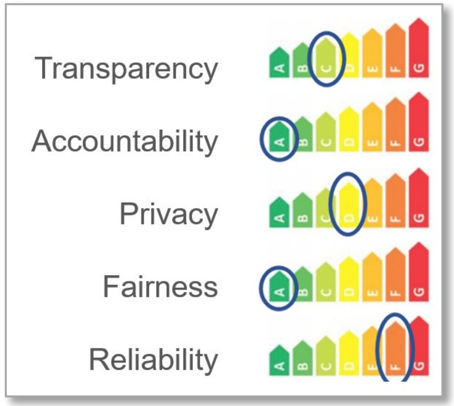


VDE SPEC

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)

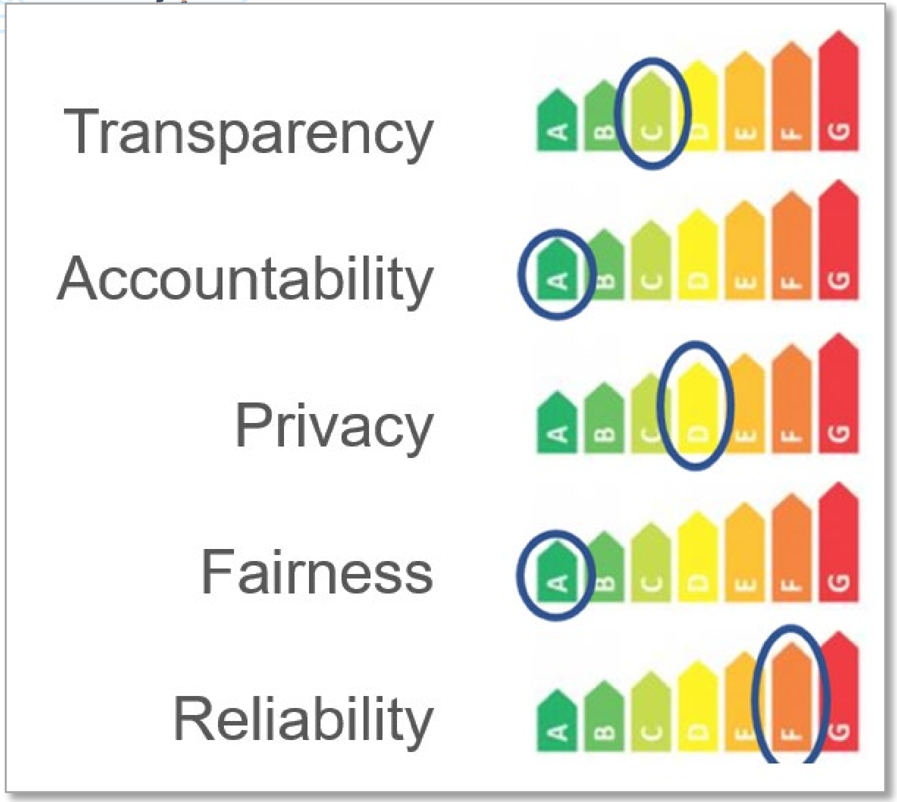
Digital Trust Forum

BASF
We create chemistry

iRights.Lab

KIT
Karlsruher Institut für Technologie

Ferdinand-Steinbeis-Institut



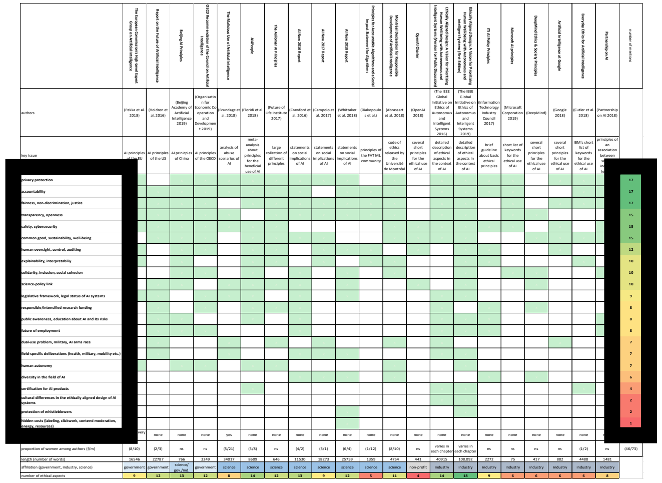
Questions:

1. Which categories do we include?

2. …

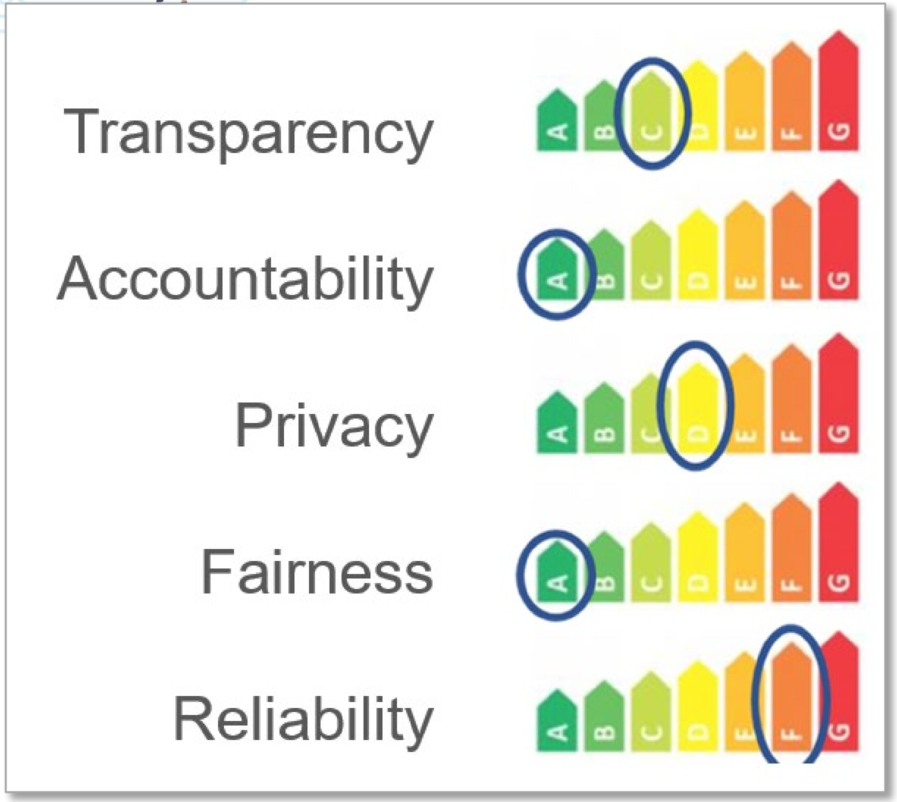
3. …

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

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

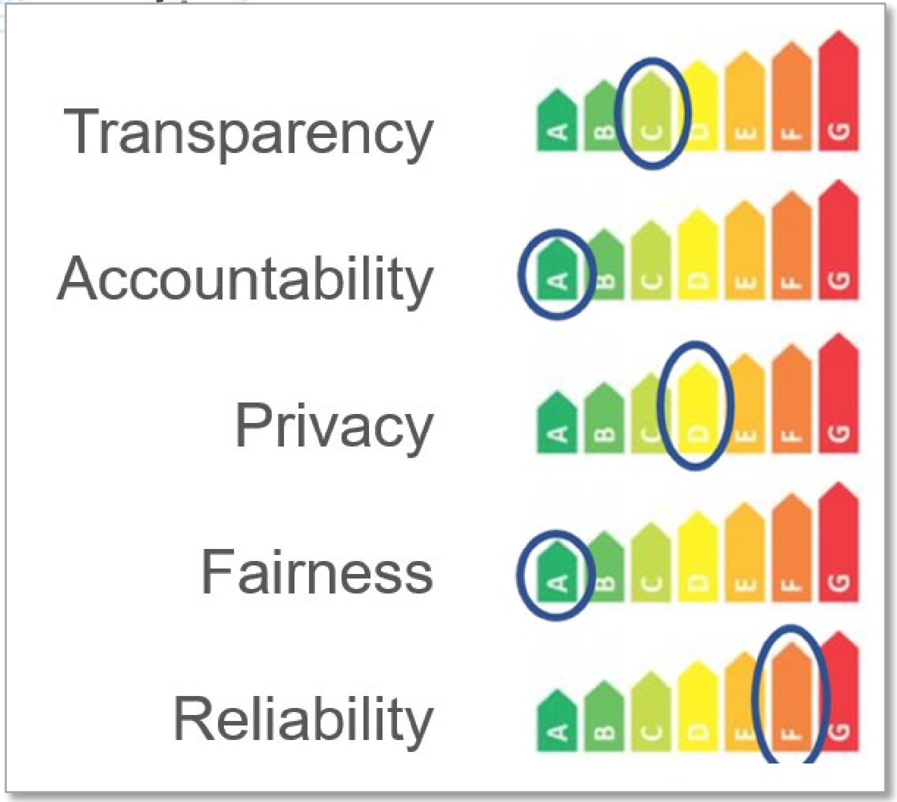
Based on T1.1
(e.g. from D)

Transparency

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	s, but only toward one target group	No
	No		one mode of interpretability is developed thout regard to target groups' needs	

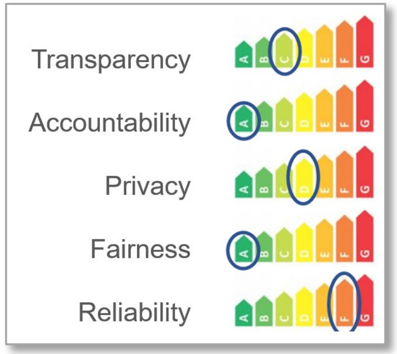
Score indicators
Build on anchor indicators.
Scoring of the score indicators are added and averaged to determine the level of the label



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



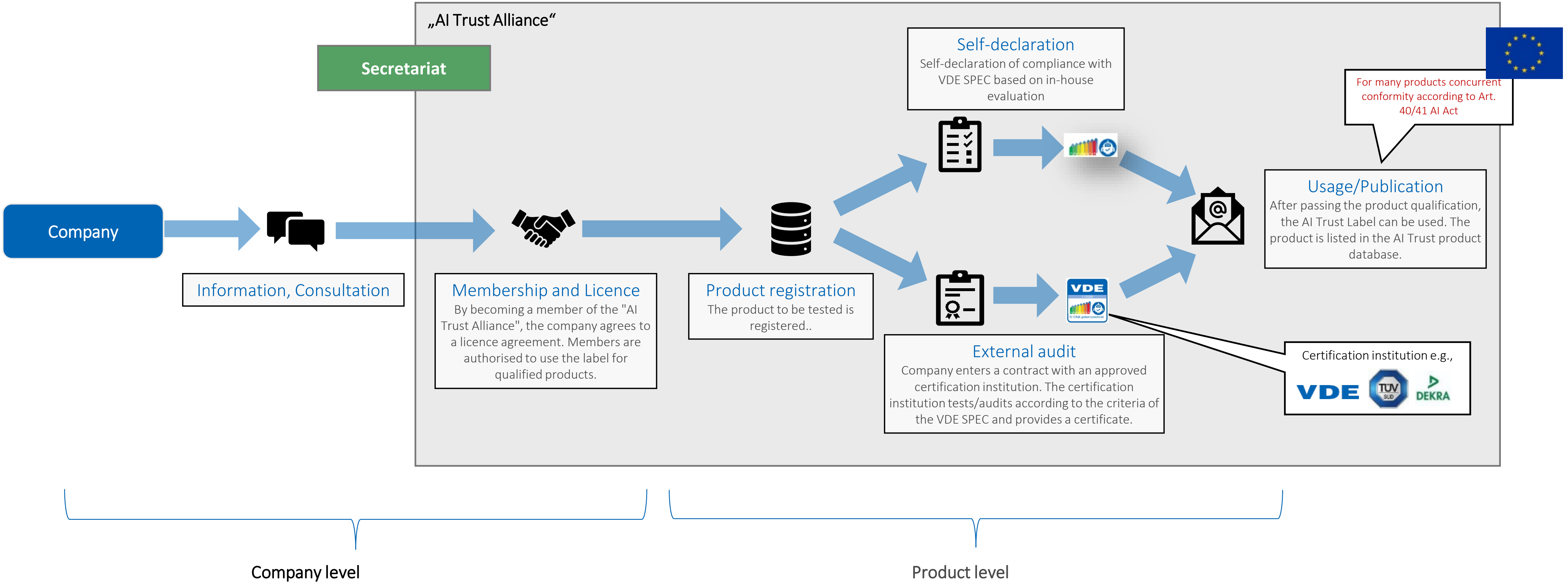
Transparency

Accountability

Privacy

Fairness

Reliability



„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.

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

Usage/Publication

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

Certification institution e.g.,

Company level

Product level

Towards a European approach

VDE SPEC

VCIO based description of systems for AI trustworthir characterisation

VDE SPEC 90012 V1.0 (en)

AI
TRUST
Transparency
Accountability
Privacy
Fairness
Reliability
A
B
C
D
E
F
G

VDE

CEN

JTC21

CENELEC

Combining complementary work
metrics – tools – governance

Confiance ai

Reminder: AI Quality Summit | 02. November 2022 Frankfurt Airport | Registration: www.ai-q.de