



Digital Trust Label &
KI-Ethik-Label
Was bringen sie wirklich?

Niniane Päßgen & Lajla Fetic
#Shift2021

SHIFT 2021 – Labels und Deklarationen

Ziele der Breakout Session

Wir wollen:

- beleuchten, wieso die Entwicklung von Labels, Deklarationen und messbaren Regelwerken zu neuen Technologien wichtig und aktuell im Trend sind – und was wir uns davon erhoffen.
- hinterfragen, was Labels und Deklarationen in der Praxis bewirken können und wo die Grenzen sind.

Dafür:

- Zeigen wir mit zwei Beispielen aus der Praxis eine mögliche Umsetzung auf.
- Diskutieren wir mit den Teilnehmenden der SHIFT 2021 aus unterschiedlichen Perspektiven über die Umsetzbarkeit und Wirkung von Konsument:innen-Labels



LAJLA FETIC

Forscht und arbeitet zur
Algorithmenethik, u.a. für die
Bertelsmann Stiftung

Ist Co-Autorin mehrerer
Praxisleitfäden zur Umsetzung
von Ethik-Regeln

Mitglied der
AI Ethics Impact Group





NINIANE PÄFFGEN

Leitet die Stiftung Swiss
Digital Initiative mit Sitz
in Genf

Fördert digitale Ethik und
die Umsetzung ethischer
Prinzipien in die Praxis
und in der digitalen Welt

Arbeit am Digital Trust
Label

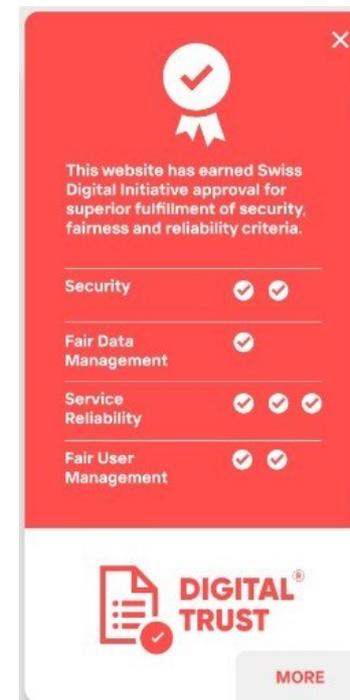




Digital Trust Label

Im Überblick

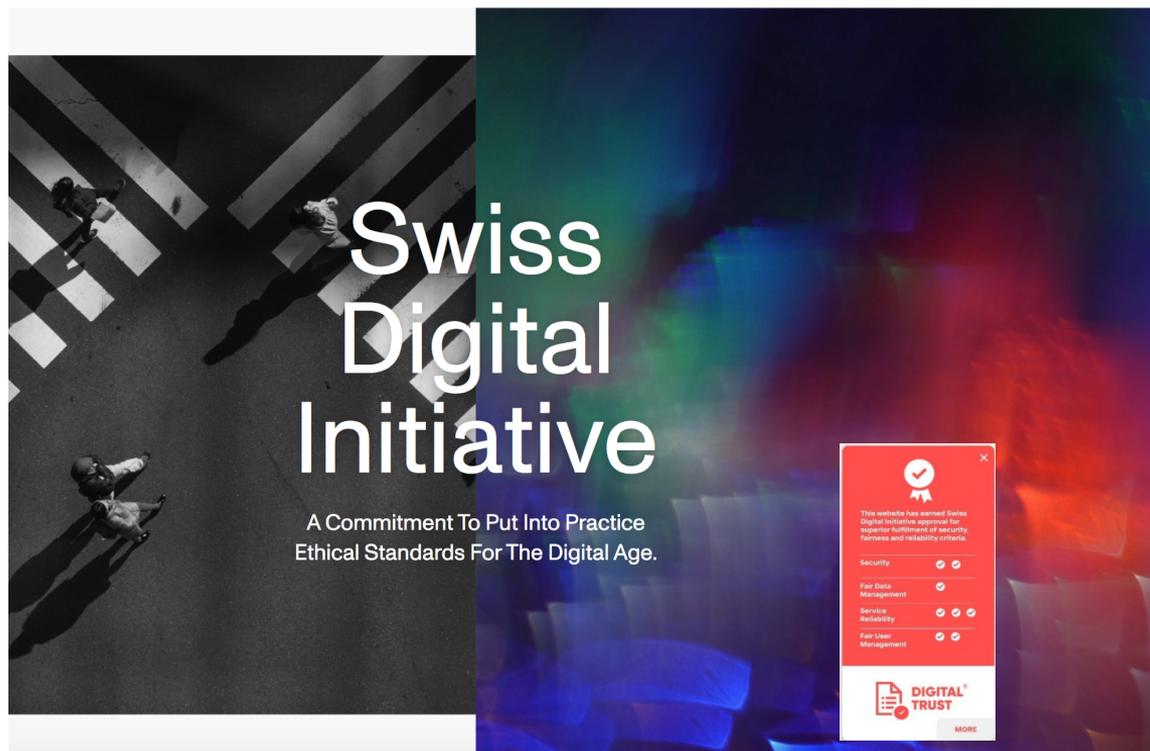
- Wir wollen das erste weltweit führende Digital Trust Label anbieten, welches die Vertrauenswürdigkeit eines digitalen Dienstes (Websites, Apps) in klarer, visuell verständlicher und nicht-technischer Sprache kennzeichnet.
- Das Label:
 - gibt Nutzerinnen und Nutzern digitaler Dienste **mehr Informationen und Transparenz**,
 - zeigt, dass ein Unternehmen **die Verantwortung gegenüber den Nutzerinnen und Nutzern** seiner digitalen Dienste ernst nimmt,
 - umfasst 4 Kategorien: **Sicherheit, Fair Data Management, Verlässlichkeit des Services, Fair User Management**,
 - besteht aus technischen Kriterien, die von einer **unabhängigen dritten Partei verifiziert und auditiert werden**.



BEISPIELILLUSTRATION

Digital Trust Label

Beispiel User Journey (noch nicht final) – Ebene I



Digital Trust Label

Beispiel User Journey (noch nicht final) – Ebene II

Understanding trustworthiness

The table offers a clear overview of which Swiss Digital Initiative defined standards a web-based service fulfills. This gives you the **transparency and detail needed to form an opinion on the trustworthiness of this service.**

Security

Your data and accounts are concealed in encrypted form to protect them from unauthorized access.

- ✓ Secure Communication, data transmission and storage
- ✓ Secure user authentication

Fair Data Management

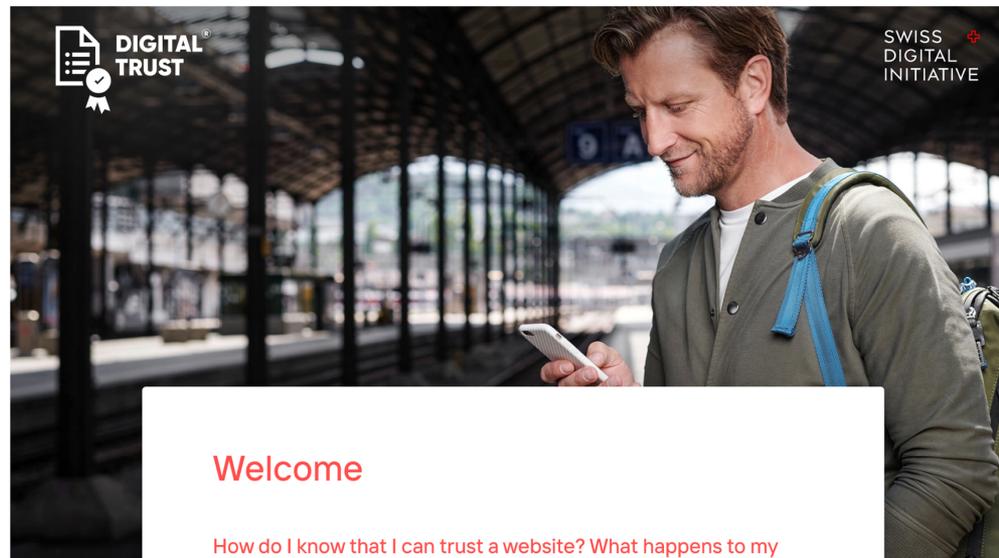
Your data is managed responsibly in regard to privacy policies, consent, data collection, and usage.

www.digitaltrust.com



Digital Trust Label

Beispiel User Journey (noch nicht final) – Ebene III



Welcome

How do I know that I can trust a website? What happens to my personal data? How protected is my privacy? As digitalization becomes more prevalent in our society, these questions are increasingly relevant. These are the concerns that gave rise to the Swiss Digital Initiative. Its aim is to motivate companies to adopt ethical web-based standards and practices

[> BACK TO OVERVIEW](#)

Digital Trust Label

Warum ein Label – Was erhoffen wir uns davon?

- **Mehr Informationen und Transparenz** für Nutzerinnen und Nutzer digitaler Dienste.
- Nutzerinnen und Nutzer können die **Vertrauenswürdigkeit digitaler Dienste besser einschätzen**.
- Ein Label **reduziert technische Komplexität und sensibilisiert in einer nutzerfreundlichen Sprache**: Hat die Nutzerin beispielsweise mit einem Algorithmus zu tun?
- Es steht für einen **Benchmark und signalisiert, dass Unternehmen die Vertrauenswürdigkeit ihrer digitalen Dienste** ernst nehmen.
- Die Service Provider nehmen durch die Anwendung der Labelkriterien **ihre Verantwortung gegenüber den Nutzerinnen und Nutzern digitaler Services** wahr.
- **Werte messbar** und von der abstrakten Theorie in die Praxis überführen.

Digital Trust Label - Erfolgskriterien

Erkenntnisse aus dem bisherigen Prozess

1. **Globale Nutzerstudie:** Hohe (globale) Nachfrage nach einer Lösung zu Vertrauen im Internet besteht.
2. **Inklusiver und transparenter Multi-Stakeholder** Entwicklungs-Prozess sind zentral
3. Das Label wird auch nach dem Launch **ständig weiterentwickelt** werden müssen
4. Das Label muss eine Kombination aus **normativen** (Garantie eines gewissen Standards) und **deskriptiven Elementen** (Informationen) sein.
5. Digitales Vertrauen kann **nicht rein technisch** gelöst werden: Zusammenspiel verschiedener Faktoren (Kontext, offline Erfahrungen etc.)
6. Die **Einhaltung und Erfüllung der Kriterien** müssen durch eine unabhängige Organisation **geprüft** werden.

Digital Trust Label

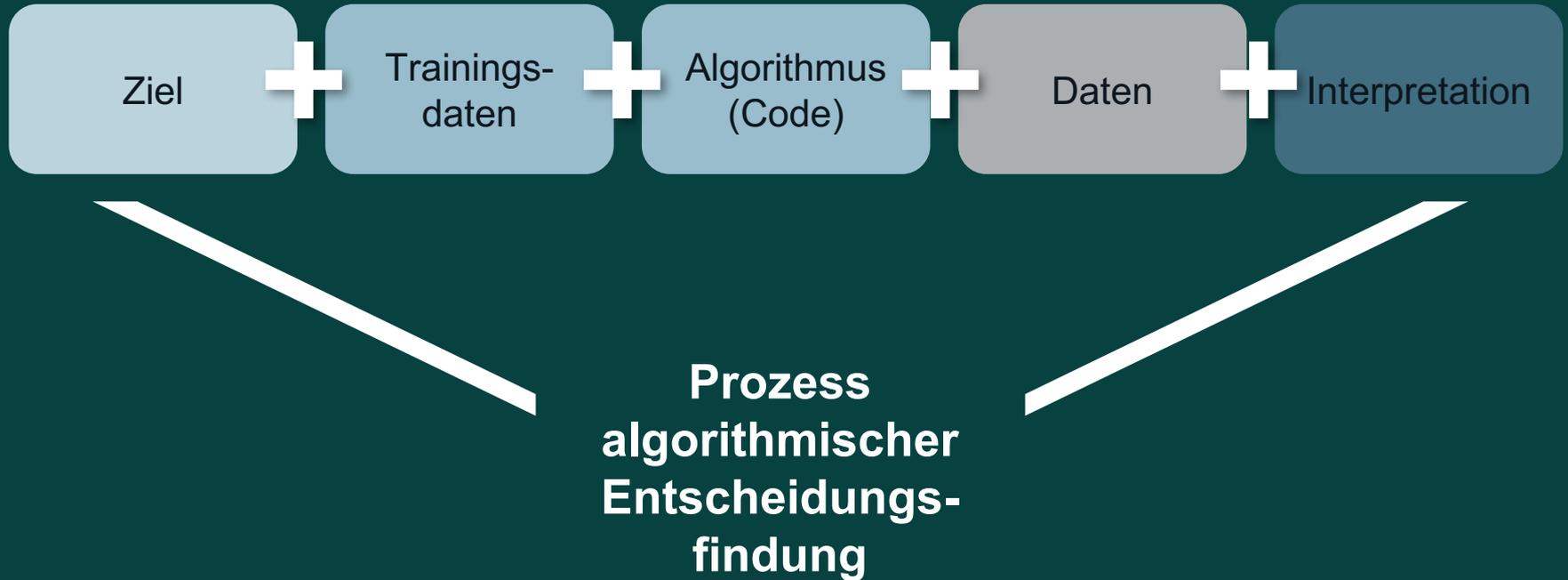
50 + Initiativen weltweit mit ähnlichem Ziel

Trustmark for the Internet (by NGI.eu)	Regional	Concept (stage 1)	Label (end user)	Next Generation Internet (NGI)	EU
A Trustworthy Tech Mark	International	Abandoned	Label (end user)	DotEveryone	Association
AI Certification	National	Prototype (stage 3)	Certification	Fraunhofer Institute, Germany's Federal Office f...	Academia Government
D-Seal, Seal for Data Ethics	National	Running (stage 4)	Label (end user)	Danish Minister for Industry, Business and Financial ...	Government
Data Ethics Framework	Regional	Realisation (stage 2)	Label (end user)	Bertelsmann Stiftung + AI Ethics Impact Group	Foundation
IHAN (Human-Driven Data Economy)	Regional	Realisation (stage 2)	Ecosystem	Sitra	Foundation Government
Independent Audit of AI Systems	International	Prototype (stage 3)	Label (end user)	For Humanity	Foundation
The Digital Standard	International	Running (stage 4)	Evaluation	Consumer Reports, Disconnect, Ranking Digital...	Association
The Ethics Certification Program for Autonomous ...	International	Prototype (stage 3)	Label (end user)	IEEE	International Organizati...

„KI ist sehr gut darin, die Welt zu beschreiben, so wie sie heute ist, mit all ihren Vorurteilen. Aber KI weiß nicht, wie die Welt sein sollte.“

Joanne Chen, Partner bei Foundation Capital

Algorithmische Entscheidungsprozesse sind von Mensch-Technik-Interaktionen geprägt.



Die 9 Algo.Rules – Den Prozess im Blick

#1 Kompetenz aufbauen

#5 Kennzeichnung durchführen

#2 Verantwortung definieren

#6 Nachvollziehbarkeit sicherstellen

#3 Ziele und erwartete Wirkung dokumentieren

#7 Beherrschbarkeit absichern

#4 Sicherheit gewährleisten

#8 Wirkung überprüfen

#9 Beschwerden ermöglichen

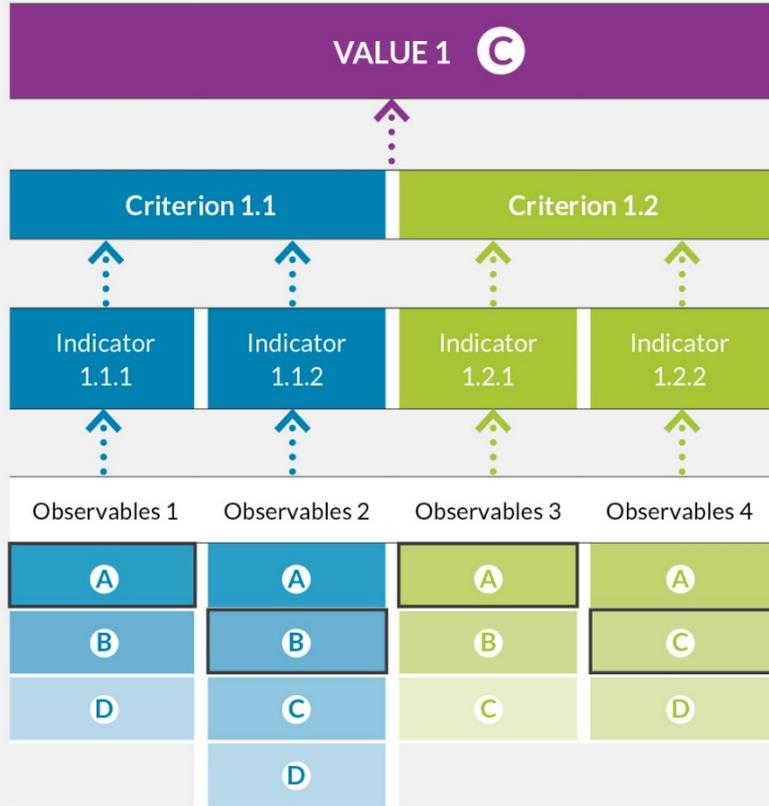
Nach einer Studie der ETH Zürich (Jobin et. al. 2019) finden sich bestimmte ethische Prinzipien immer wieder in den Ethik-Richtlinien.

SPANNEND:

Die Definitionen der Prinzipien fehlen, unterscheiden oder widersprechen sich teilweise.

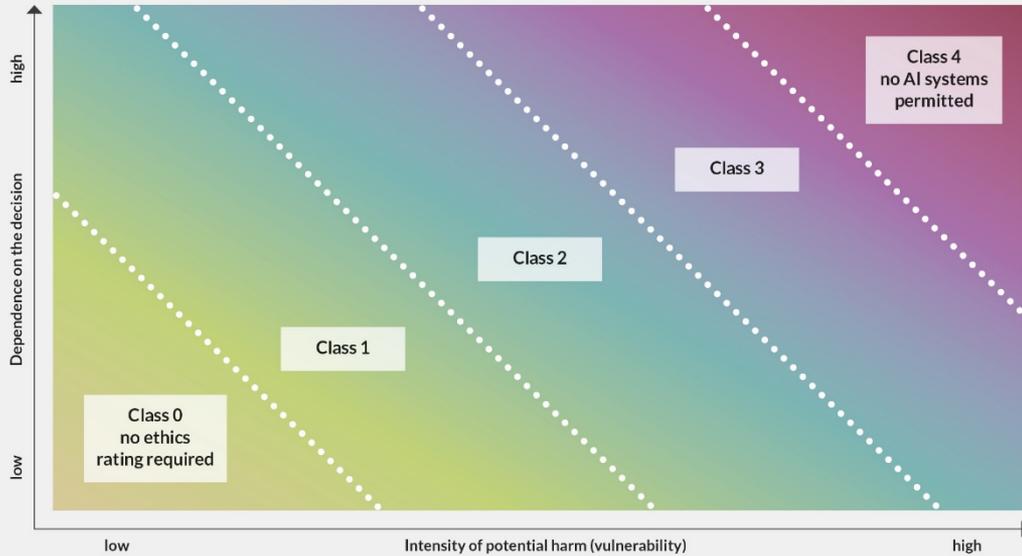
Fairness
Accountability
Transparenz:
Was bedeutet das überhaupt?

System rating and operationalisation of a value using minimum requirements and aggregation



Der Weg von abstrakten Werten zu messbaren Indikatoren ist kompliziert.

Risk matrix with 5 classes of application areas with risk potential ranging from 'no ethics rating required' in class 0 to the prohibition of AI systems in class 4



Source: Krafft and Zweig 2019

AI&I Group

Die Anforderungen
sind nicht für
jeden
Anwendungs-
kontext gleich.

Applying the VCIO approach to transparency as a value

Value	TRANSPARENCY						TRANSPARENCY						Value
Criteria	Disclosure of origin of data sets			Disclosure of properties of algorithm/model used			Accessibility						Criteria
Indicators	Is the data's origin documented?	Is it plausible for each purpose, which data is being used?	Are the training data set's characteristics documented and disclosed? Are the corresponding data sheets comprehensive?	Has the model in question been tested and used before?	Is it possible to inspect the model so far that potential weaknesses can be discovered?	Taking into account efficiency and accuracy, has the simplest and most intelligible model been used? ¹	Are the modes of interpretability target-group-specific and have been developed with the target groups?	Who has access to information about data sets and the algorithm/model used?	Is the operating principle comprehensible and interpretable?	Are the modes of interpretability in their target-group-specific form intelligible for the target groups?	Are the hyperparameters (parameters of learning methods) accessible?	Has a mediating authority been established to settle and regulate transparency conflicts?	Indicators
Observables	Yes, comprehensive logging of all training and operating data, version control of data sets etc. ²	Yes, the use of data and the individual application are intelligible	Yes and the data sheets are comprehensive	Yes, the model is widely used and tested both in theory and practice ³	Yes, the model can easily be inspected and tested	Yes, the model has been evaluated and the most intelligible model has been used	Yes	Everyone	Yes, the model itself is directly comprehensible	Yes, the modes of interpretability have been tested with target groups for intelligibility	Yes, to everyone	Yes, a competent authority has been established	Observables
	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 (some) data sheets contain few or missing information	Yes, the model is known and tested in either theory or practice	Yes, but the model can only be tested by certain people due to non-disclosure	No, but the model was evaluated regarding interpretability and this evaluation is disclosed to the public	Yes, but without participation of the target groups	All people directly affected	Yes, the modes of interpretability are provided with the model itself	Yes, target groups can complain or ask if they do not understand a mode of interpretability	Yes, but only to information and trust intermediaries (regulators, watchdogs, researchers, courts)	Yes, a competent authority has been established but its powers are limited	
	No, but a summary on data usage is available	No, but a summary on data usage is available	Yes, the model is known to some experts but has not been tested yet	Yes, the model is known to some experts but has not been tested yet	No	No, the model has not been evaluated	Yes, but the modes or interpretability are only specific for one target group	Only information and trust intermediaries (regulators, watchdogs, research, courts)	No, the modes of interpretability need to be adjusted to the individual model and use by experts	No	No	No	
	No logging; data used is not controlled or documented in any way	No	No	No, the model has been developed recently	No	No, the model has not been evaluated	No, the modes of interpretability ⁴ are not target-group-specific	Nobody	No, but the model is theoretically comprehensible	No	No	No	
									No, there are no known modes of interpretability				

¹ This indicator would require further specification regarding the balance between using an efficient and accurate model and using a model which is technically simple and thus naturally easier to comprehend and follow.

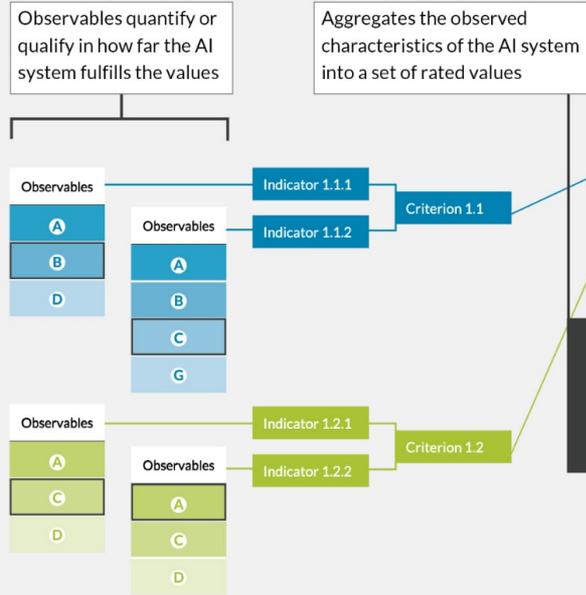
² This observable could include further levels of logging and documentation of data sets.

³ This observable could help to determine the levels needed in other observables: If the model has been widely used and tested, it might not require additional testing.

⁴ "Modes of interpretability" refers to different methods to ensure or increase interpretability (use of simple model, explanations of data and model used, etc.).

Illustration of the composition of the whole system rating using minimum requirements

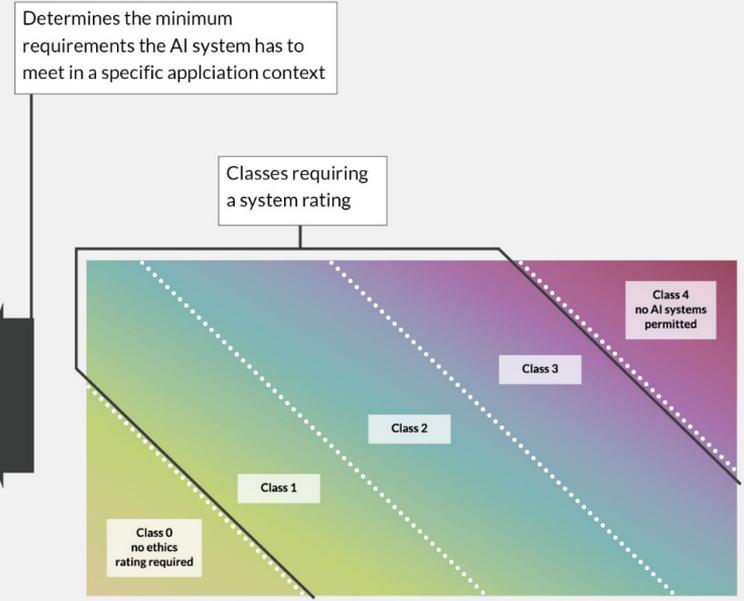
VCIO Approach



AI ETHICS LABEL



Classification of the Application Context



Wir müssen die Delegation von Verantwortung verhindern.

Label ersetzen die gesellschaftliche Debatte über den Einsatz und die Anwendung von Technologie nicht.

Wir brauchen einen Werkzeugkasten für KI-Ethik und eine systemische Perspektive.

**Reicht das
KI-Ethik-Label
aus?**

SWISS 
DIGITAL
INITIATIVE

Niniane Päßgen
Geschäftsführerin
Swiss Digital Initiative

niniane@sdi-foundation.org

c/o Campus Biotech, Chemin des
Mines 9, 1202 Genève

<https://swiss-digital-initiative.org>

Lajla Fetic
Wissenschaftlerin und Beraterin
KI-Ethik und Regulierung

lajla.fetic@email.de

[@lajlafetic](https://www.linkedin.com/in/lajlafetic/)
www.linkedin.com/in/lajlafetic/
www.lajlafetic.de

L
F