Digital Trust Criteria

Version 3

Valid as of June 2024

SWISS DIGITAL INITIATIVE

Our Digital Trust Criteria

This Criteria Catalogue forms the basis for third-party audits and the award of the Digital Trust Label. Auditing Companies performing Digital Trust Label audits must use this document together with our Interpretation Guidelines and our Code of Practice.

This document is regularly reevaluated and updated by our Digital Trust Expert Group to ensure that it reflects the current state of technological as well as regulatory developments with a focus on protecting and empowering the end-user of a digital service. All changes to the last version are documented in the Change Log at the end of this document.

In this version, additional guidance and criteria have been added to reflect the emergence of Generative Artificial Intelligence (GenAI). We rely on the AI system definition provided by the OECD to shape our understanding of the term AI system[1]

> *“An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.”*

And we rely on the definition of GenAI from the Swiss Competence Network for Artificial Intelligence (CNAI)[2]

> *“‘Generative AI’ is a broad term that refers to AI systems that are trained on large amounts of data from the physical and virtual world in order to generate data themselves (e.g. texts, imagery, sound recordings, videos, simulations, and codes). They are often multimodal, e.g. with input and/or output in one or several modalities (e.g. text, image or video). Different model architectures, including diffusion models and transformer-based models, can be used for generative tasks.”*

As a particular technology within AI, GenAI exacerbates questions around the origin and reliability of systems' outputs, anthropomorphism and risks of over reliance. We strive for a Criteria Catalogue that is applicable to as many digital services as possible, but given that generative AI is putting a lot of pressure on digital service providers to integrate those AI capabilities we drafted these additional criteria with GenAI in mind, although many of them also apply to other AI systems.

As the technological and regulatory environment is evolving rapidly will also create a separate document for organizations specifically thinking about the use of GenAI. This guidance document is independent of audits for the Digital Trust Label and will provide additional guidance and support.

Security

Data Protection

Reliability

Fair User Interaction

Criteria reviewed and amended by the Digital Trust Expert Group

Bays, Xavier	*Head of Consulting & Associate, Swiss-SDI*
Blattner, Marcel (Co-President)	*Principal Data Scientist and Team Leader, ETH Swiss Data Science Center*
Böhler, Nikki	*Co-Director, Intersections*
Fetai, Ilir	*Head Competence Center Machine Perception, SBB & Lecturer FFHS*
Fischli, Roberta	*Political scientist, PhD Candidate, University of St. Gallen*
Groth, Maximilian	*Co-Founder & CEO, Decentriq*
Gubser, Rahel	*Researcher Digital Health & Medical Data Science*
Jotterand, Alexandre	*CIPP/E, CIPM, attorney-at-law at id est avocats*
Kende, Michael	*Chair of the Board of the Datasphere Initiative and a Senior Advisor at Analysys Mason*
Koller, Rodolphe (Co-President)	*Editor in Chief, ICTjournal*
Kuonen, Diego	*CEO, Statoo Consulting & Professor Data Science, University of Geneva*
Ochoa, Martin	*Senior Researcher and Lecturer, ETHZ*
Ruiz Barragan, Santiago	*Legal advisor in Data protection and new Technologies*
Scherr, Mitchell	*CEO, Assured Cyber Protection*
Shamsrizi, Manouchehr	*Co-Founder, gamelab.berlin Humboldt-Universität's Cluster of Excellence and Co-Founder RetroBrain R&D*
Steiger, Martin	*Attorney-at-law and Partner, Steiger Legal, Co-Founder and CEO, Datenschutzpartner*
Toplic, Leila	*Chief Communications and Trust Officer, Carbonfuture*
Tuulia, Timonen	*Head of PSC Service Excellence, Posti Group*
van Ooijen Falce, Charlotte	*Senior Policy Analyst, Digital Government and Public Sector Data at CvanO Research & Advice*

Security

The criteria in this dimension cover:

- **Secure communication, data transmission & storage**
- **Secure user authentication**
- **Secure service set-up, maintenance and update**
- **Monitoring & reporting of vulnerabilities & breaches**

1. The service shall apply best practice cryptography to data in transit, ensuring that the cryptography is reviewed and evaluated, delivers the required functions for all transmitted data and is appropriate to the properties of the technology, risk, and usage. All data in transit over open communication lines such as the internet must be encrypted.

2. The service shall apply best practice cryptography to data at rest, ensuring that the cryptography is reviewed and evaluated, delivers the required functions for all sensitive and applicable data at rest and is appropriate to the properties of the technology, risk, and usage.

   If AI systems are used, this includes clear policies for how data used for the entire lifecycle of the AI system (e.g. design, data processing, model building, validation, deployment, operation and monitoring) is protected.

3. Privacy-enhancing technologies such as Anonymization and Pseudonymization shall be used according to best practices in order to adequately protect the user's data.

4. All passwords used for the service shall be subject to a state-of-the-art password policy, which includes requirements applicable to the service and ensures that no hard-coded passwords are used, best practice authentication is in place and ensures that brute-force attacks on authentication mechanisms are not feasible.

Security

5. Guidance for secure installation, configuration, and updates shall be in place and updated for each release if necessary. Guidance shall be available in a manner that is easy to access and understand. Any major changes shall lead to a communication to the users in an easy-to-understand format.
6. All software components shall be updatable in a secure manner, and verification of security updates shall be in place.
7. Updates shall be timely. Updates addressing critical security vulnerabilities must be available as soon as possible.
8. Hard-coded critical security parameters in service software source code shall not be used.
9. Any critical security parameters used for integrity and authenticity checks of software updates and for protection of communication with associated service software shall be unique per service and shall implement security measures to protect the integrity and confidentiality of critical security parameters.
10. The service provider shall follow secure management processes for critical security parameters that relate to the service.
11. The service provider shall continually monitor, identify, and rectify security vulnerabilities and/or breaches including for the entire lifecycle of AI systems used (e.g. model or prompt hacking), and shall provide a public point of contact as part of a vulnerability disclosure policy so that security researchers and others are able to report issues.

Security

12. Critical security vulnerabilities shall be communicated to relevant authorities within 72 hours if not corrected, and the impacted users shall be timely and adequately informed.

13. Personal data breaches shall be communicated to relevant authorities and impacted data subjects within 72 hours.

14. The service provider shall implement and monitor legal and technical restrictions on AI systems.

Data Protection

The criteria in this dimension cover:

- **User consent**
- **Data retention & processing**

15. The user shall be informed about the purpose of the processing and the legal basis for processing of their personal data in clear and plain language. Where there is more than one purpose/legal basis, they need to be listed separately in a way that the user is able to easily distinguish between one purpose/legal basis and another.
    If AI systems are used, this includes informing the user about the use of personal data throughout the entire lifecycle of the AI system (e.g. design, data processing, model building, validation, deployment, operation and monitoring) and providing the option to opt-in.

16. Where user consent is sought for the processing of personal data, such consent shall be expressly collected from the user for each of the purposes and legal basis listed by the service provider and obtained separately from the terms and conditions of use of the services.

17. Where initial user consent is sought for the processing of personal data, the user shall be provided with the option of either opting in, or opting out, expressed through a valid and affirmative action. If checkboxes are used, they shall not pre-ticked. The user shall also be given the possibility of requesting additional information.

18. The user shall be provided with a separate, easy, and accessible way of withdrawing consent.

Data Protection

19. The user shall be informed of the definite time period for which the personal data will be stored. If that is not possible, the user shall be informed of the criteria and reasons used to determine the indefinite period, and a regular timeframe for which a review will be undertaken.

20. In cases in which the service provider anonymises personal data, upon a request by the user, such service provider shall provide a detailed explanation of how personal data is being anonymised, and the safety measures used to prevent de-anonymisation. The service provider shall also update the user on the anonymisation status of any personal data held by the provider at the time of the request.

21. Once the data retention period lapses, the service provider shall either anonymise or delete the personal data. In case of indefinite data retention periods where regular reviews are to be undertaken (criteria 19), the user shall be informed of the outcomes of the review within 30 days.

22. The service provider shall ensure that the user can access their data. Any requests for access need to be acceded to within 30 days. Together with a copy of the personal data, a user is to be provided with names of third parties with whom such personal data has been shared, together with the legal basis under which such data is being held.

23. The service provider shall make the best efforts to ensure that the AI system, once trained, does not inadvertently reveal confidential or private information or patterns that could be traced back to individual data points or personal data or replicate real personal data when generating output ("leaking").

Reliability

The criteria in this dimension cover:

- **Reliable service updates**
- **Resilience to service outage**
- **Functional reliability**
- **Accountability**

24. The software version of the service shall be easy to access and understand. If AI systems are used this includes as much information about the models as can be shared publicly.

25. There shall also be policies in place to review and update the systems to ensure they meet standards of fairness and non-discrimination.

26. The service provider shall publish, in a way that is easy to access and understand for the user, the defined support period and the need for that support period.

27. Disaster recovery, business continuity and data backup and restore policies and procedures shall be in place and regularly tested to ensure ongoing availability of the service and associated data.

28. The service shall provide its users with an extensive, easy-to-access, easy-to-understand description of its functionalities, and shall operate in strict accordance with this description.

29. If relevant, the service shall provision for a secure, precise and efficient billing and payment system which employs two-factor authentication and adheres to local and regional norms.

Reliability

30. If relevant, the service shall provision for a delivery system which fulfills state-of-the-art conditions of the associated specific activity domain.

31. The service shall provide its users with an easy-to-access, easy-to-understand, and easy-to-print service and service provider identification.

32. The service shall document its compliance with all applicable laws and regulation and assign a contact representative for easy-to-access and easy-to-understand information about legislation that the service is subject to.

33. User inquiries and complaints shall be treated in a timely fashion, and relevant alternative dispute resolution mechanisms must be in place to facilitate these processes.

34. The service provider shall clearly communicate the limitations of the AI systems, ensuring that users are aware of contexts where the AI might not be reliable.

35. There shall be a proactive & reactive review (e.g. from unexpected results or user feedback) and update policy in place to improve the AI systems reliability and accuracy.

36. The service provider shall implement a technical standard to disclose information about who created or changed a piece of digital content (e.g. texts, images, sound recordings, codes, etc.), what was changed and how it was changed.

Fair User Interaction

The criteria in this dimension cover:

- **Non-discriminating access**
- **Fair user interfaces**
- **Fair use of AI-based algorithms**

37. The system shall provide a non-discriminating access to all its potential users.

38. Service interfaces shall be designed so as not to arouse over reliance deceive, nor to manipulate the users, and, in particular, shall exclude clearly manipulative techniques ("dark patterns") such as Interface Interference (Preselection, Obstruction), Aesthetic Manipulation (Toying with emotions, False Hierarchy), Anthropomorphism, Disguised ads (Trick questions, Sneaking), Forced Actions (Social Pyramid, Gamification, Privacy Zuckering). In addition, the use and objectives of mildly manipulative techniques (incl. nudging) shall be clearly announced to the users and proportionate to the objectives of the service.

39. The service shall not be designed to exclusively cause user addiction and shall provide the users with an easy-to-access, easy-to-understand information about potential addiction risks during its set up.

40. Service providers whose services are illegal to users under the age of 18 shall take proportional steps to verify users' age and prevent under-18s from accessing those services.

Fair User Interaction

41. There shall be clear and easily understandable information to the user when interacting with AI systems, especially with automated decision-making algorithms or content-generation algorithms. The service provider shall also indicate which user-related data is processed by the AI-system and its relationship to the objectives of the service, in addition to informing why an AI is used for the service. Any risks and limitations inherent to the AI systems must be clearly and concisely described to the user.

42. If AI-based algorithms and, especially, automated decision-making algorithms or content-generation algorithms, are used, the service shall provision for specific mechanisms to assess their robustness, resilience, and accuracy, as well as the risks associated with their exploitation, and shall provide the user with the possibility to request that a representative of the service provider, reviews and validates the outputs produced by the algorithm.

43. If an AI system creates personalized content for users, the service provider shall undertake and document best efforts so that such content doesn't reinforce harmful biases, stereotypes, or misinformation. There shall be policies in place to ensure the fairness and impartiality of outputs.

Change Log DTL Criteria Catalog

Old criteria Version 2	**Updated criteria Version 3 (changes in bold)** *Valid since June 2024*
2: The service shall apply best practice cryptography to data at rest, ensuring that the cryptography is reviewed and evaluated, delivers the required functions for all sensitive and applicable data at rest and is appropriate to the properties of the technology, risk, and usage.	2: The service shall apply best practice cryptography to data at rest, ensuring that the cryptography is reviewed and evaluated, delivers the required functions for all sensitive and applicable data at rest and is appropriate to the properties of the technology, risk, and usage. **If AI systems are used, this includes clear policies for how data used for the entire lifecycle of the AI system (e.g. design, data processing, model building, validation, deployment, operation and monitoring) is protected.**
11: The service provider shall continually monitor, identify, and rectify security vulnerabilities and/or breaches, and shall provide a public point of contact as part of a vulnerability disclosure policy so that security researchers and others are able to report issues.	11: The service provider shall continually monitor, identify, and rectify security vulnerabilities and/or breaches **including for the entire lifecycle of AI systems used (e.g. model or prompt hacking)**, and shall provide a public point of contact as part of a vulnerability disclosure policy so that security researchers and others are able to report issues.
new	**14: The service provider shall implement and monitor legal and technical restrictions on AI systems.**
13: The user shall be informed about the purpose of the processing and / or the legal basis for processing of their personal data in clear and plain language. Where there is more than one purpose and /or legal basis, they need to be listed separately in a way that the user is able to easily distinguish between one purpose and / or legal basis and another.	**15**: The user shall be informed about the purpose of the processing and / or the legal basis for processing of their personal data in clear and plain language. Where there is more than one purpose and /or legal basis, they need to be listed separately in a way that the user is able to easily distinguish between one purpose and / or legal basis and another. **If AI systems are used, this includes informing the user about the use of personal data throughout the entire lifecycle of the AI system (e.g. design, data processing, model building, validation, deployment, operation and monitoring) and providing the option to opt-in.**
19: Once the data retention period lapses, the service provider shall either anonymise or delete the personal data. In case of indefinite data retention periods where regular reviews are to be undertaken (criteria 17), the user shall be informed of the outcomes of the review within 30 days.	**21**: Once the data retention period lapses, the service provider shall either anonymise or delete the personal data. In case of indefinite data retention periods where regular reviews are to be undertaken (criteria **19**), the user shall be informed of the outcomes of the review within 30 days.

Change Log DTL Criteria Catalog

Old criteria Version 2	**Updated criteria Version 3 (changes in bold)** *Valid since June 2024*
new	**23: The service provider shall make the best efforts to ensure that the AI system, once trained, does not inadvertently reveal confidential or private information or patterns that could be traced back to individual data points or personal data or replicate real personal data when generating output ("leaking").**
21: The software version of the service shall be easy to access and understand.	**24:** The software version of the service shall be easy to access and understand. **If AI systems are used this includes as much information about the models as can be shared publicly.**
new	**25: There shall also be policies in place to review and update the systems to ensure they meet standards of fairness and non-discrimination.**
new	**34: The service provider shall clearly communicate the limitations of the AI systems, ensuring that users are aware of contexts where the AI might not be reliable.**
new	**35: There shall be a proactive & reactive review (e.g. from unexpected results or user feedback) and update policy in place to improve the AI systems reliability and accuracy.**
new	**36: The service provider shall implement a technical standard to disclose information about who created or changed a piece of digital content (e.g. image, video, audio recording, document), what was changed and how it was changed.**

Change Log DTL Criteria Catalog

Old criteria Version 2	**Updated criteria Version 3 (changes in bold)** *Valid since June 2024*
31: Service interfaces shall be designed so as not to deceive, nor to manipulate the users, and, in particular, shall exclude clearly manipulative techniques ("dark patterns") such as Interface Interference (Preselection, Obstruction), Aesthetic Manipulation (Toying with emotions, False Hierarchy), Disguised ads (Trick questions, Sneaking), Forced Actions (Social Pyramid, Gamification, Privacy Zuckering). In addition, the use of mildly manipulative techniques shall be clearly announced to the users and proportionate to the objectives of the service.	**38:** Service interfaces shall be designed so as not **to arouse over reliance,** deceive, nor to manipulate the users, and, in particular, shall exclude clearly manipulative techniques ("dark patterns") such as Interface Interference (Preselection, Obstruction), Aesthetic Manipulation (Toying with emotions, False Hierarchy), **Anthropomorphism**, Disguised ads (Trick questions, Sneaking), Forced Actions (Social Pyramid, Gamification, Privacy Zuckering). In addition, the use **and objectives** of mildly manipulative techniques **(incl. nudging)** shall be clearly announced to the users and proportionate to the objectives of the service.
34: There shall be clear and easily understandable information to the user when interacting with AI systems, especially with automated decision-making algorithms. The service provider shall also indicate which user-related data is processed by the AI-system and its relationship to the objectives of the service, in addition to informing why an AI is used for the service. Any risks and limitations inherent to the AI systems must be clearly and concisely described to the user.	**41:** There shall be clear and easily understandable information to the user when interacting with AI systems, especially with automated decision-making algorithms **or content-generation algorithms**. The service provider shall also indicate which user-related data is processed by the AI-system and its relationship to the objectives of the service, in addition to informing why an AI is used for the service. Any risks and limitations inherent to the AI systems must be clearly and concisely described to the user.
35: If AI-based algorithms and, especially, automated decision-making algorithms, are used, the service shall provision for specific mechanisms to assess their robustness, resilience, and accuracy, as well as the risks associated with their exploitation, and shall provide the user with the possibility to request that a representative of the service provider, reviews and validates the outputs produced by the algorithm.	**42:** If AI-based algorithms and, especially, automated decision-making algorithms **or content-generation algorithms**, are used, the service shall provision for specific mechanisms to assess their robustness, resilience, and accuracy, as well as the risks associated with their exploitation, and shall provide the user with the possibility to request that a representative of the service provider, reviews and validates the outputs produced by the algorithm.
new	**43: If an AI system creates personalized content for users, the service provider shall undertake and document best efforts so that such content doesn't reinforce harmful biases, stereotypes, or misinformation. There shall be policies in place to ensure the fairness and impartiality of outputs.**

References

1. https://www.oecd-ilibrary.org/science-and-technology/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en
2. https://cnai.swiss/en/products/terminology/

SWISS
DIGITAL
INITIATIVE

CONTACT

Swiss Digital Initiative
Chemin des Mines 9
1202 Geneva
Switzerland

info@sdi-foundation.org

Disclaimer: This document has been prepared only for purposes of the use of the recipient and only serves for the scope and the terms agreed. This document may not be reproduced or circulated without the Swiss Digital Initiative's prior written consent. We accept no liability (including for negligence) to anyone else in connection with this document, and it may not be provided to anyone else.