Ethics Label for Digital Systems to Promote Transparency and User Awareness

Marco Autili, University of L’Aquila, **Riccardo Corsi,** University of Pisa/GSSI,
Martina De Sanctis, Gran Sasso Science Institute, **Paola Inverardi**, Gran Sasso Science Institute,
Patrizio Pelliccione, Gran Sasso Science Institute

NB: This interview will be audio recorded for later analysis.

The study goal and background

Why an ethics label?

Assessing ethics as a quality for modern digital systems in a user-centered perspective.

The goal of our work is to move beyond mere compliance with standards, proposing an ethics label that helps users understand how software systems impact human, societal, and environmental values—both throughout their development and during real-world use.

How do we build the ethics label?

(i) reviewing existing guidelines, framework, recommendations, laws and regulations on digital systems to identify human, societal, and environmental drivers of innovation;

(ii) identifying and analyzing existing standards, e.g., the SQuaRE family

(iii) realizing an ethical label user-centered.

Identified limitations

(i) Need of new qualities

(ii) Flourishing of standards

(iii) Focus mostly on product-oriented aspects

(iv) Limitation of CE marking / legal compliance

(v) Need of transparency and digital-ethical literacy

Considered sources - Guidelines, Frameworks, Laws, Regulations, Standards

Guidelines, frameworks, regulations

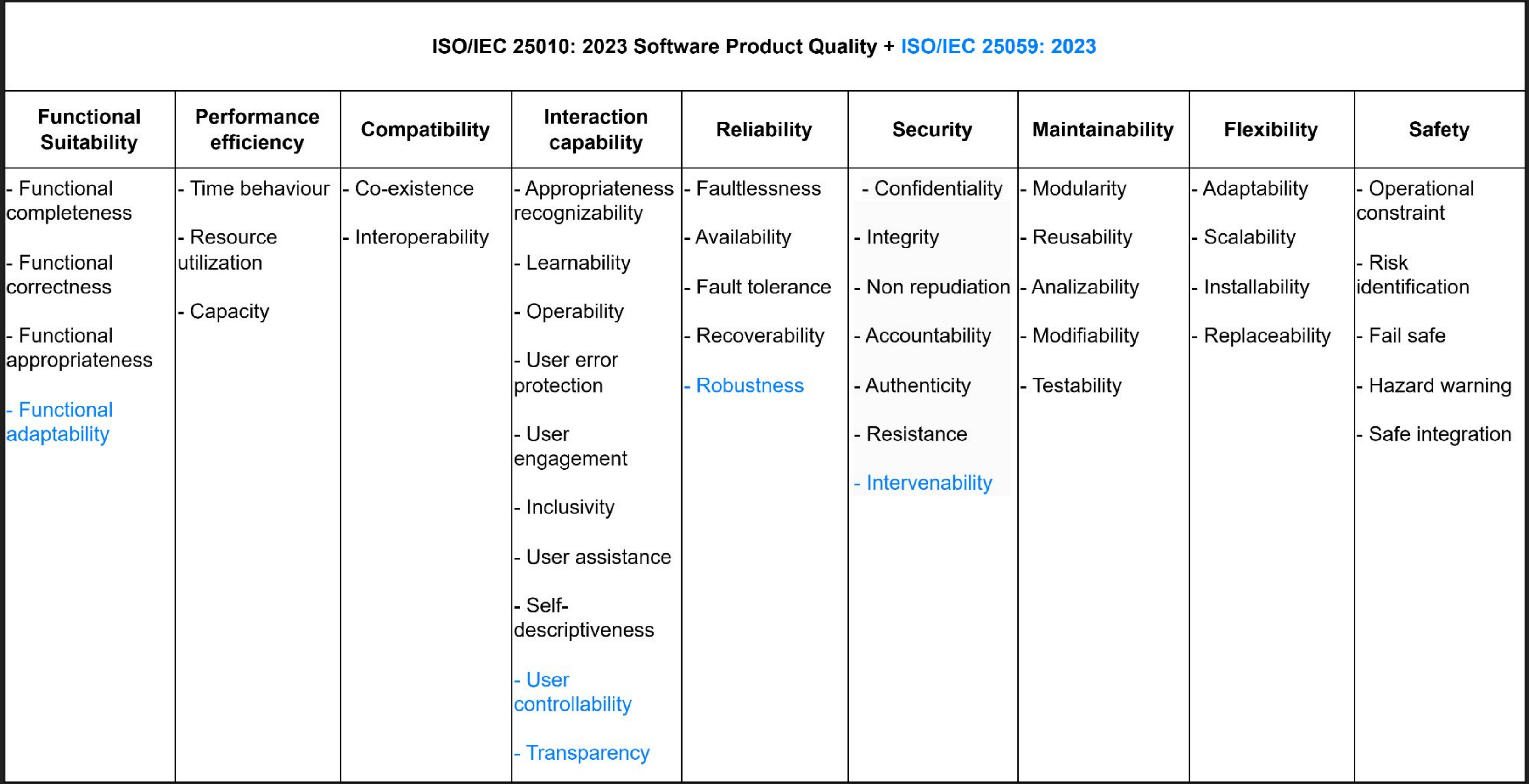
- High-Level Expert Group on AI: Ethics guidelines for trustworthy AI (2019/2024)
- Beijing AI Principles (2019)
- OECD.AI Policy Observatory: OECD AI Principles overview (2019/2024)
- UNESCO: Recommendation on the ethics of artificial intelligence (2022)
- G7 Hiroshima Process: International Guiding Principles for Advanced AI Systems (2023)
- United States Government: Blueprint for an AI Bill of Rights Making Automated Systems Work for the American People (2022)
- USA Algorithmic Accountability Act (2022)
- Council of Europe Framework Convention on AI and Human Rights (2024)
- China’s Deep Synthesis Provisions (2023)
- China’s Interim Measures on Generative AI (2023)
- China’s Governance principles for the New Generation Artificial Intelligence (2019)
- China’s Ethical Norms for the NGAI (2021)
- Government Data Quality Frameworks: UE; Canada

SQuaRE Standards

- ISO/IEC 25010:2023 Software Product Quality
- ISO/IEC 25059:2023 Quality model for AI systems
- ISO/IEC 25012:2008 Data quality model
- ISO/IEC 5259:2024 - Quality model for data
- analytics and AI based on ML
- ISO/IEC 25019:2023 Quality-in-use model

Human, Societal, Environmental (HSE) Drivers and ISO Standards SQuaRe

ISO Standards SQuaRe



ISO/IEC 25010: 2023 Software Product Quality + ISO/IEC 25059: 2023

Functional Suitability	Performance efficiency	Compatibility	Interaction capability	Reliability	Security	Maintainability	Flexibility	Safety
- Functional completeness - Functional correctness - Functional appropriateness - Functional adaptability	- Time behaviour - Resource utilization - Capacity	- Co-existence - Interoperability	- Appropriateness recognizability - Learnability - Operability - User error protection - User engagement - Inclusivity - User assistance - Self-descriptiveness - User controllability - Transparency	- Faultlessness - Availability - Fault tolerance - Recoverability - Robustness	- Confidentiality - Integrity - Non repudiation - Accountability - Authenticity - Resistance - Intervenability	- Modularity - Reusability - Analizability - Modifiability - Testability	- Adaptability - Scalability - Installability - Replaceability	- Operational constraint - Risk identification - Fail safe - Hazard warning - Safe integration

ISO Standards SQuaRe

ISO/IEC 25019:2023 - Quality in use model + ISO/IEC 25059:2023			
Beneficialness	**Freedom from risk**	**Acceptability**	**Satisfaction**
- Usability - Accessibility - Suitability	- Freedom from economic risk - Freedom from environmental and societal risk - Freedom from health risk - Freedom from human life risk - Societal and ethical risk mitigation	- Experience - Trustworthiness - Compliance	Transparency

ISO/IEC 25012:2008 + ISO/IEC 5259:2024 - Quality model for data analytics and AI based on ML			
Inherent data quality	**Inherent and system-dependent data quality**	**System-dependent data quality**	**Additional characteristics**
- Accuracy - Completeness - Consistency - Credibility - Currentness	- Accessibility - Compliance - Confidentiality - Efficiency - Precision - Traceability - Understandability	- Availability - Portability - Recoverability	- Auditability - Identifiability - Effectiveness - Balance - Diversity - Relevance - Representativeness - Similarity - Timeliness

HSE Drivers

In our previous work we identified a first version of human, societal, and environmental drivers, organized according to the following categories (which are, in turn, organized in sub-categories):

- Societal and Environmental well-being
- Accountability and Responsibility
- Privacy and Data Governance
- Human Agency and Oversight
- Transparency and Explainability
- Diversity, Fairness, and non-discrimination

HSE Drivers

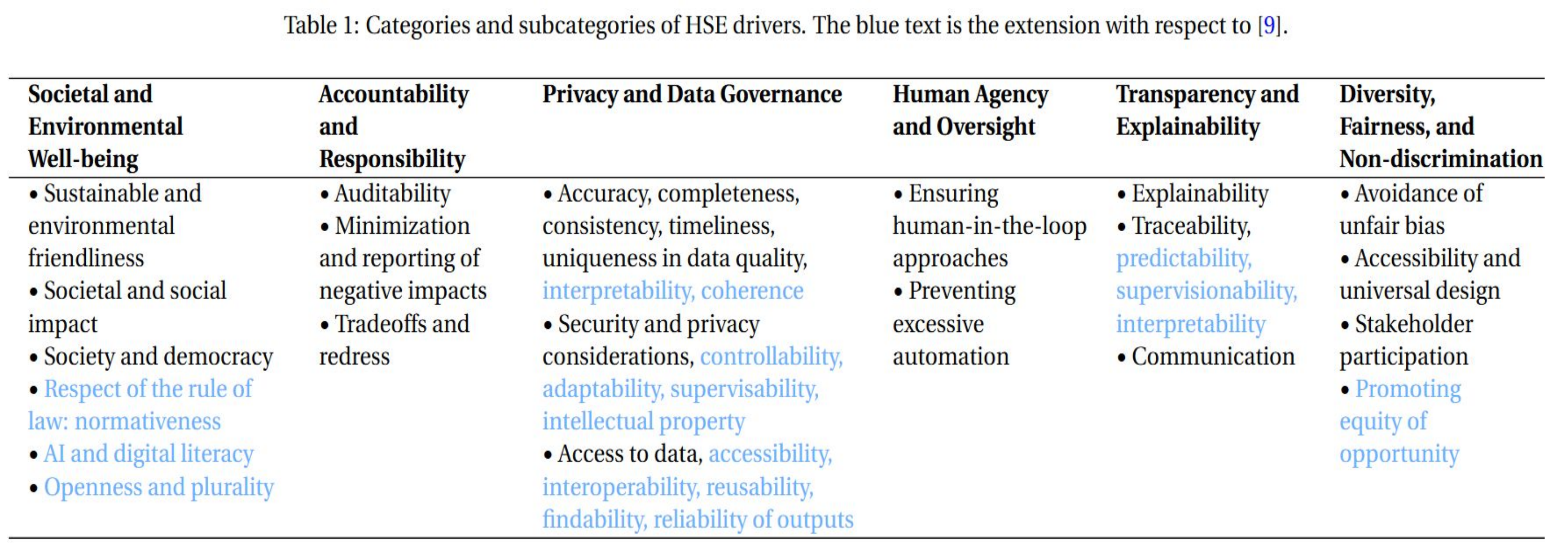
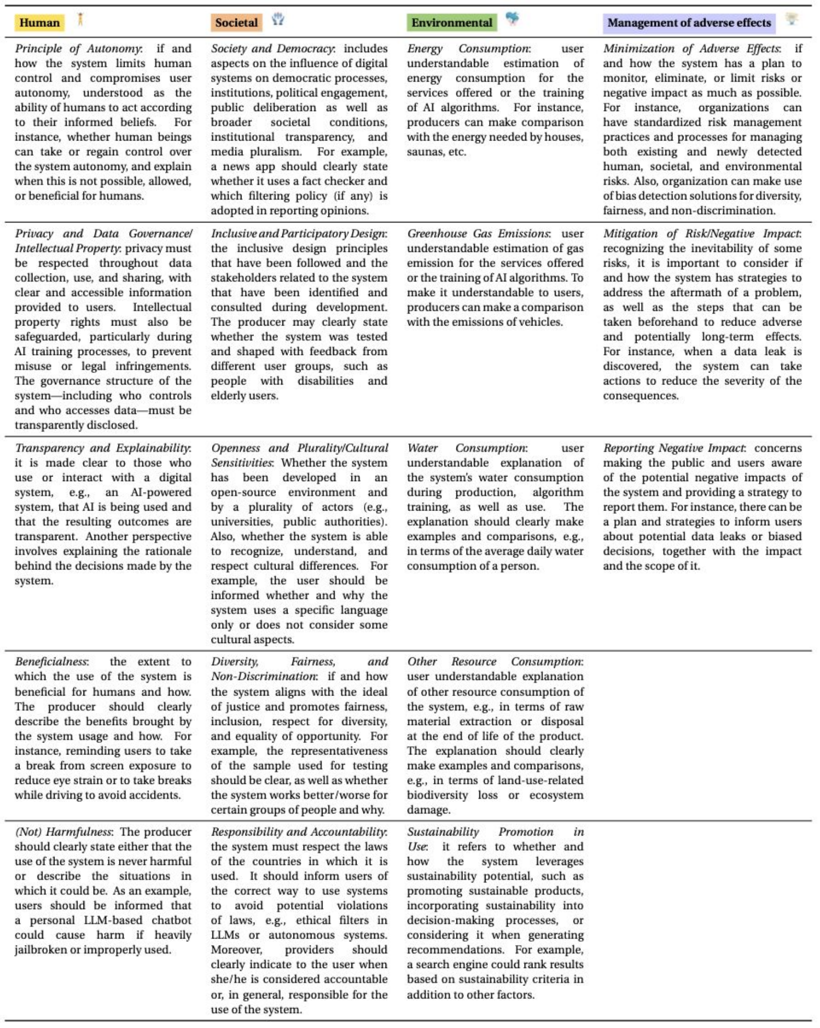


Table 1: Categories and subcategories of HSE drivers. The blue text is the extension with respect to [9].

Societal and Environmental Well-being	Accountability and Responsibility	Privacy and Data Governance	Human Agency and Oversight	Transparency and Explainability	Diversity, Fairness, and Non-discrimination
• Sustainable and environmental friendliness • Societal and social impact • Society and democracy • Respect of the rule of law: normativeness • AI and digital literacy • Openness and plurality	• Auditability • Minimization and reporting of negative impacts • Tradeoffs and redress	• Accuracy, completeness, consistency, timeliness, uniqueness in data quality, interpretability, coherence • Security and privacy considerations, controllability, adaptability, supervisability, intellectual property • Access to data, accessibility, interoperability, reusability, findability, reliability of outputs	• Ensuring human-in-the-loop approaches • Preventing excessive automation	• Explainability • Traceability, predictability, supervisionability, interpretability • Communication	• Avoidance of unfair bias • Accessibility and universal design • Stakeholder participation • Promoting equity of opportunity

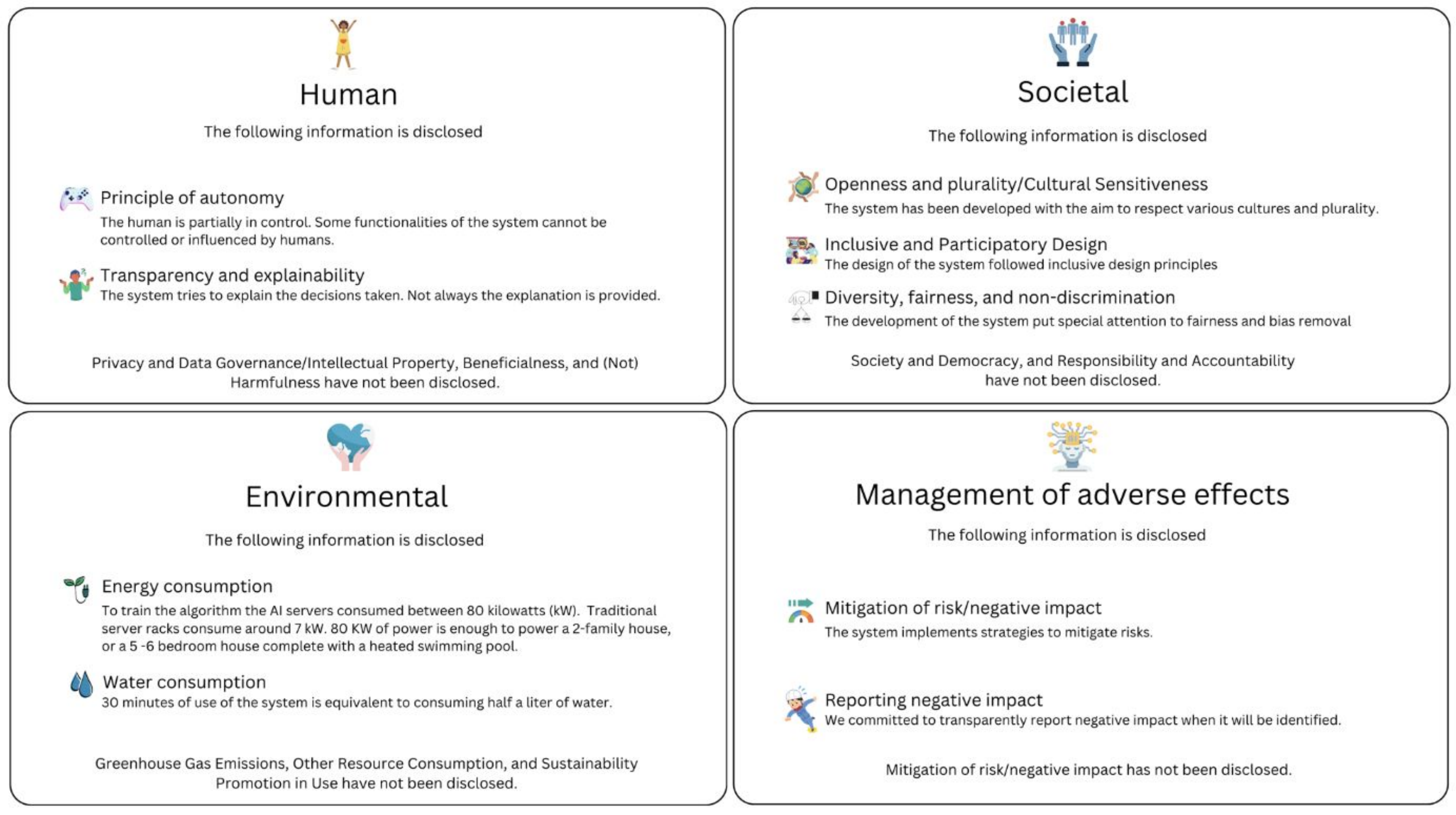
Ethics label

Each column represents a *Card* (e.g., Human), made by a set of *characteristics* (e.g., principle of autonomy, etc.)



Human	Societal	Environmental	Management of adverse effects
Principle of Autonomy: if and how the system limits human control and compromises user autonomy, understood as the ability of humans to act according to their informed beliefs. For instance, whether human beings can take or regain control over the system autonomy, and explain when this is not possible, allowed, or beneficial for humans.	*Society and Democracy*: includes aspects on the influence of digital systems on democratic processes, institutions, political engagement, public deliberation as well as broader societal conditions, institutional transparency, and media pluralism. For example, a news app should clearly state whether it uses a fact checker and which filtering policy (if any) is adopted in reporting opinions.	*Energy Consumption*: user understandable estimation of energy consumption for the services offered or the training of AI algorithms. For instance, producers can make comparison with the energy needed by houses, saunas, etc.	*Minimization of Adverse Effects*: if and how the system has a plan to monitor, eliminate, or limit risks or negative impact as much as possible. For instance, organizations can have standardized risk management practices and processes for managing both existing and newly detected human, societal, and environmental risks. Also, organization can make use of bias detection solutions for diversity, fairness, and non-discrimination.
Privacy and Data Governance/ Intellectual Property: privacy must be respected throughout data collection, use, and sharing, with clear and accessible information provided to users. Intellectual property rights must also be safeguarded, particularly during AI training processes, to prevent misuse or legal infringements. The governance structure of the system—including who controls and who accesses data—must be transparently disclosed.	*Inclusive and Participatory Design*: the inclusive design principles that have been followed and the stakeholders related to the system that have been identified and consulted during development. The producer may clearly state whether the system was tested and shaped with feedback from different user groups, such as people with disabilities and elderly users.	*Greenhouse Gas Emissions*: user understandable estimation of gas emission for the services offered or the training of AI algorithms. To make it understandable to users, producers can make a comparison with the emissions of vehicles.	*Mitigation of Risk/Negative Impact*: recognizing the inevitability of some risks, it is important to consider if and how the system has strategies to address the aftermath of a problem, as well as the steps that can be taken beforehand to reduce adverse and potentially long-term effects. For instance, when a data leak is discovered, the system can take actions to reduce the severity of the consequences.
Transparency and Explainability: it is made clear to those who use or interact with a digital system, e.g., an AI-powered system, that AI is being used and that the resulting outcomes are transparent. Another perspective involves explaining the rationale behind the decisions made by the system.	*Openness and Plurality/Cultural Sensitivities*: Whether the system has been developed in an open-source environment and by a plurality of actors (e.g., universities, public authorities). Also, whether the system is able to recognize, understand, and respect cultural differences. For example, the user should be informed whether and why the system uses a specific language only or does not consider some cultural aspects.	*Water Consumption*: user understandable explanation of the system's water consumption during production, algorithm training, as well as use. The explanation should clearly make examples and comparisons, e.g., in terms of the average daily water consumption of a person.	*Reporting Negative Impact*: concerns making the public and users aware of the potential negative impacts of the system and providing a strategy to report them. For instance, there can be a plan and strategies to inform users about potential data leaks or biased decisions, together with the impact and the scope of it.
Beneficialness: the extent to which the use of the system is beneficial for humans and how. The producer should clearly describe the benefits brought by the system usage and how. For instance, reminding users to take a break from screen exposure to reduce eye strain or to take breaks while driving to avoid accidents.	*Diversity, Fairness, and Non-Discrimination*: if and how the system aligns with the ideal of justice and promotes fairness, inclusion, respect for diversity, and equality of opportunity. For example, the representativeness of the sample used for testing should be clear, as well as whether the system works better/worse for certain groups of people and why.	*Other Resource Consumption*: user understandable explanation of other resource consumption of the system, e.g., in terms of raw material extraction or disposal at the end of life of the product. The explanation should clearly make examples and comparisons, e.g., in terms of land-use-related biodiversity loss or ecosystem damage.	
(Not) Harmfulness: The producer should clearly state either that the use of the system is never harmful or describe the situations in which it could be. As an example, users should be informed that a personal LLM-based chatbot could cause harm if heavily jailbroken or improperly used.	*Responsibility and Accountability*: the system must respect the laws of the countries in which it is used. It should inform users of the correct way to use systems to avoid potential violations of laws, e.g., ethical filters in LLMs or autonomous systems. Moreover, providers should clearly indicate to the user when she/he is considered accountable or, in general, responsible for the use of the system.	*Sustainability Promotion in Use*: it refers to whether and how the system leverages sustainability potential, such as promoting sustainable products, incorporating sustainability into decision-making processes, or considering it when generating recommendations. For example, a search engine could rank results based on sustainability criteria in addition to other factors.	

Ethics label: an example



Human

The following information is disclosed

Principle of autonomy
The human is partially in control. Some functionalities of the system cannot be controlled or influenced by humans.

Transparency and explainability
The system tries to explain the decisions taken. Not always the explanation is provided.

Privacy and Data Governance/Intellectual Property, Beneficialness, and (Not) Harmfulness have not been disclosed.

Societal

The following information is disclosed

Openness and plurality/Cultural Sensitiveness
The system has been developed with the aim to respect various cultures and plurality.

Inclusive and Participatory Design
The design of the system followed inclusive design principles

Diversity, fairness, and non-discrimination
The development of the system put special attention to fairness and bias removal

Society and Democracy, and Responsibility and Accountability have not been disclosed.

Environmental

The following information is disclosed

Energy consumption
To train the algorithm the AI servers consumed between 80 kilowatts (kW). Traditional server racks consume around 7 kW. 80 KW of power is enough to power a 2-family house, or a 5 -6 bedroom house complete with a heated swimming pool.

Water consumption
30 minutes of use of the system is equivalent to consuming half a liter of water.

Greenhouse Gas Emissions, Other Resource Consumption, and Sustainability Promotion in Use have not been disclosed.

Management of adverse effects

The following information is disclosed

Mitigation of risk/negative impact
The system implements strategies to mitigate risks.

Reporting negative impact
We committed to transparently report negative impact when it will be identified.

Mitigation of risk/negative impact has not been disclosed.

Discussion

Q1: How would you describe your company in terms of its mission and activities?

Introductory questions

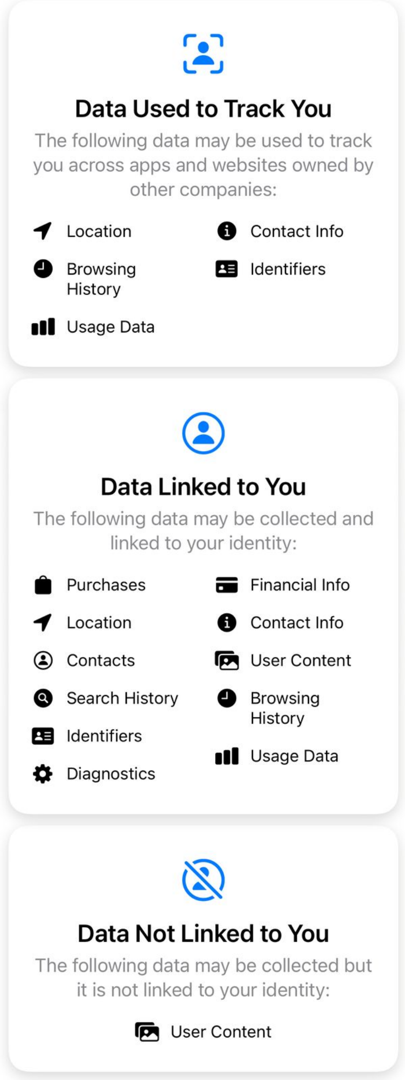
Q2: How would you describe your professional profile and background?

Introductory questions

Q3: Can you briefly describe the software product your company develops, and the role of AI within it?

Q4: Have you ever heard of privacy or nutrition labels?

Privacy Label: example



Data Used to Track You

The following data may be used to track you across apps and websites owned by other companies:

- Location
- Contact Info
- Browsing History
- Identifiers
- Usage Data

Data Linked to You

The following data may be collected and linked to your identity:

- Purchases
- Financial Info
- Location
- Contact Info
- Contacts
- User Content
- Search History
- Browsing History
- Identifiers
- Usage Data
- Diagnostics

Data Not Linked to You

The following data may be collected but it is not linked to your identity:

- User Content

Q5: Is the proposed ethics label easily understandable?

Evaluating Comprehensibility / Understandability

Q6: Does the proposed label appropriately address Human, Societal, and Environmental drivers?

HSE Drivers

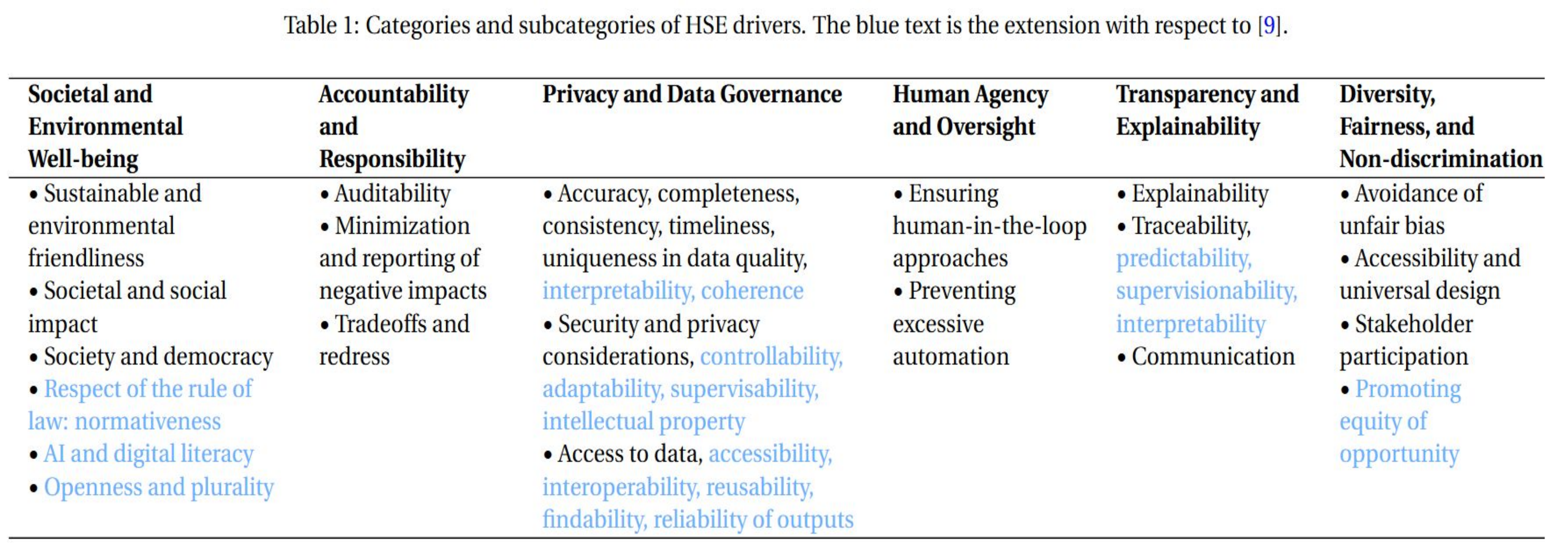


Table 1: Categories and subcategories of HSE drivers. The blue text is the extension with respect to [9].

Societal and Environmental Well-being	Accountability and Responsibility	Privacy and Data Governance	Human Agency and Oversight	Transparency and Explainability	Diversity, Fairness, and Non-discrimination
• Sustainable and environmental friendliness • Societal and social impact • Society and democracy • Respect of the rule of law: normativeness • AI and digital literacy • Openness and plurality	• Auditability • Minimization and reporting of negative impacts • Tradeoffs and redress	• Accuracy, completeness, consistency, timeliness, uniqueness in data quality, interpretability, coherence • Security and privacy considerations, controllability, adaptability, supervisability, intellectual property • Access to data, accessibility, interoperability, reusability, findability, reliability of outputs	• Ensuring human-in-the-loop approaches • Preventing excessive automation	• Explainability • Traceability, predictability, supervisionability, interpretability • Communication	• Avoidance of unfair bias • Accessibility and universal design • Stakeholder participation • Promoting equity of opportunity

Q7: From a developer's perspective, is the proposed ethics label useful for expressing the ethical quality of your product?

Q7.1: Why?

Q8: Should you label your software product using the ethics label we are proposing, which *cards* and *characteristics* would you consider displaying to the user?

Q8.1: Why?

Evaluating Utility / Usability

Anything to add?

Thanks!