Bridging the Communication Gap: Evaluating AI Labeling Practices for Trustworthy AI Development

Raphael Fischer[1,2], **Magdalena Wischnewski**[1,3], **Alexander van der Staay**[1,4], **Katharina Poitz**[1,2], **Christian Janiesch**[1,4], **Thomas Liebig**[1,2]
[1]TU Dortmund University
Dortmund, Germany
[2]Lamarr Institute for Machine Learning and Artificial Intelligence
[3]Research Center Trustworthy Data Science and Security
[4]Chair of Enterprise Computing
{forename}.{surname}@tu-dortmund.de

Abstract

Artificial intelligence (AI) is becoming integral to economy and society. However, communication gaps between developers, users, and stakeholders hinder trust and informed decision-making. To make the behavior of AI models more transparent, high-level AI labels have been proposed, drawing inspiration from systems like energy labeling. While AI labels can already inform on performance trade-offs, for example with regard to predictive model performance and resource efficiency, the practical benefits and limitations of this communication form remain underexplored. Our study evaluates AI labeling through qualitative interviews along key research questions. Based on thematic analysis and inductive coding, we firstly identify a broad range of practitioners with diverse use cases and requirements to be interested in AI labeling. Benefits are primarily seen for bridging communication gaps and aiding non-expert decision-makers. However, our interviewees also mentioned limitations and suggestions for improvement. In comparison to other reporting formats, the reduced complexity of labels was acknowledged to benefit fast knowledge acquisition without deep technical AI expertise. Trustworthiness was found to be strongly influenced by usability and credibility, with mixed preferences for self-certification versus third-party certification. Our insights specifically highlight that AI labels pose a trade-off between simplicity and complexity, address diverse user needs, and nudge interviewee priorities toward sustainability. As such, our study validates AI labels as a valuable tool for enhancing trust and communication in AI, offering actionable guidelines for their refinement and standardization.

Introduction

As artificial intelligence (AI) technology advances, it holds strong economic promises and becomes integral for various business cases (Lins et al. 2021). Adopting AI however involves various stakeholders, such as software developers, domain experts, and project leaders, who need to reach agreements despite their very different levels of expertise. To ensure the trustworthy (Li et al. 2023) and sustainable (Rohde et al. 2024) use of AI, it is imperative to bridge the communication gaps between the diverse parties that develop, use, or are affected by AI services. Examples of such gaps include limited technical understanding, which can even occur among AI developers (Kaur et al. 2020), and unrealistic expectations (Piorkowski et al. 2021), which can potentially result in AI misuse and disuse (Parasuraman and Riley 1997). As such, whether stakeholders are simply using AI services or developing custom machine learning (ML) models, sustainable development necessitates to understand AI behavior and practical implications (Chaaben 2024).

To make informed decisions, stakeholders need a comprehensible form of communication that allows to understand practical AI properties and performance trade-offs, for example regarding resource demands versus predictive quality (Fischer et al. 2024). Unfortunately, established forms of reporting mostly address experts (Fischer, Liebig, and Morik 2024) and are biased toward focusing on predictive performance (Birhane et al. 2022). To foster resource-awareness and increase transparency toward audiences that are less proficient in AI, Fischer et al. therefore developed high-level AI labels. In analogy to well-known systems like the European Union (EU) energy labels (Fischer et al. 2022) or textile care labels (Morik et al. 2022), they use relative comparisons and color-coding to inform about AI models without presupposing any profound ML understanding. While AI labeling could be an "excellent tool" for sustainable development (Genovesi and Mönig 2022), an empirical evaluation is direly needed, especially because several works questioned the effectiveness of labeling systems (Solà et al. 2020; Peters and Verhagen 2024) and AI trust seals (Wischnewski et al. 2024; Scharowski et al. 2023).

With this work, we address the need for an in-depth evaluation of AI labeling by conducting an interdisciplinary user study, focusing on the following main research questions:

- Who is interested in AI labeling and what are their problems with using or developing AI? (RQ1)
- What are the practical benefits and limitations of labeling AI model behavior? (RQ2)
- How are AI labels perceived in comparison to other forms of reporting? (RQ3)
- How do AI labels and the corresponding certifying authority affect the trustworthiness of AI systems? (RQ4)

Figure 1: Prototype AI labels, as generated by `STREP` (Fischer, Liebig, and Morik 2024) and shown during interviews.

To answer them, we conducted semi-structured interviews with 16 participants from various application domains and backgrounds. Besides discussing their daily business with AI, we also confronted them with prototype AI labels as displayed in Figure 1. Facing the labels, the interviewees mentioned various advantages and limitations, and were asked to compare labels with other forms of reporting and describe their role in the context of trust. Following thematic analysis, we developed an extensive code system around our research questions, with a total of over 1000 code occurrences.

Our findings evidence practical benefits of AI labels for connecting ML experts with less-informed users despite occasional misunderstandings and concerns about technical complexity. Labeling was not only acknowledged to ease decision making, but our analysis even suggests that AI labels can act as nudges toward more informed and sustainable AI use. Our interviewees highlighted the importance of usability and customizability, indicating that labeling formats need to be tailored to stakeholder desiderata. Our work thus can be seen as a justification for establishing AI labels, but also as an implication for not understanding the current state of labeling as a 'one-fits-all' solution. To that end, our findings allow us to formulate guidelines for future refinement and standardization. With this study, the theoretical concept of AI labeling is qualitatively validated for practical feasibility, showcasing effectiveness for bridging communication gaps and facilitating sustainable AI development. Our research follows open science practices by making all supplementary materials and results, such as the transcripts and coding system, available at https://github.com/raphischer/labeling-evaluation. We start by establishing a literary background and describing our methodical approach, after which the results from our interviews will be presented and analyzed.

Related Work

On the Current Challenges in AI Development Until recently, incorporating AI into business required skilled ML engineers who analyzed the business use case and data at hand to develop custom solutions (i.e., models). Small and medium-sized enterprises often struggled to keep pace in the race to make business and profit with AI, as it required substantial upfront investments in hardware and expertise (Boag et al. 2018). The last years, however, brought forth a paradigm shift commonly referred to as AI-as-a-service (AIaaS) (Lins et al. 2021), making AI more accessible via cloud computing. The respective ML models are usually of pre-trained and generative nature (Feuerriegel et al. 2024) and can be used for a variety of tasks (Bommasani et al. 2022), while the internal complexity remains hidden behind prompt interfaces. While open source and science practices are elementary in AI evolution (Widder et al. 2022), it should be noted that the largest portion of AI technology "resides firmly in the hands of a small set of actors, companies, and countries" (Sætra 2021).

Whether it be AIaaS or custom ML models—the versatility of AI inevitably leads to knowledge and communication gaps. Classically, they occur between ML engineers and application experts, which can for example complicate the identification and prioritization of AI use cases (Fischer et al. 2023). Discussing AI communication challenges, Piorkowski et al. (2021) have singled out prominent gaps in the context of knowledge, establishing trust, and setting expectations, explicitly highlighting the importance of customized documentation. Summarizing the experiences of nearly 5000 practitioners, another study concluded that education is an important factor for the success of incorporating ML (Nahar et al. 2023). In interactive ML, non-experts were found to struggle due to misunderstanding (Yang et al. 2018), for example with respect to predictive model performance. While using AIaaS reduces the need for employing ML experts, the practical implications of services remain hard to grasp without fundamental ML knowledge. Opacity, among other potential drawbacks, represents a significant inhibiting factor to the adoption of AIaaS (Pandl et al. 2021). A designated qualitative study argues that "there is no one-size-fits-all service" and concludes that AIaaS providers need to better convey information on performance, costs, and trade-offs (Brecker et al. 2023). AI literacy also plays an important role for bridging knowledge gaps, however requires a solid conceptualization (Ng et al. 2021) that is hard to establish at the current pace of AI evolution.

On Trustworthiness and the Potential of AI Labeling Closely connected to these problems is the ongoing endeavor to align AI development with human rights and ethical values, which entails the question of AI trustworthiness (Chatila et al. 2021). From an interdisciplinary viewpoint, trust is a key prerequisite for not only the general uptake but also the appropriate use of technology (Lee and See 2004; Parasuraman and Riley 1997). For the "appropriate reliance" on automated systems and AI, users' trust should be calibrated to the system's actual trustworthiness (Wischnewski, Krämer, and Müller 2023; Mehrotra et al. 2024). This necessitates a "harmonized interplay between product and organizational perspectives", which could possibly be established via systematic quality assurance (Schmitz et al. 2022). For cloud-based offerings such as AIaaS, a duality of trust can be observed, meaning that trust should be complemented by organizational trust in service providers (Lansing and Sunyaev 2016), also referred to as as initial trust (Mck-

night et al. 2011). Signaling theory suggests that trust could be boosted by addressing the information asymmetry between two parties (e.g., ML engineers and AI users), in order to signal transparency intentions and reveal important information (Kirmani and Rao 2000). Unfortunately, this concept does not always translate to digital services, as for example Kim, Ferrin, and Rao found high-level communication about web services via trust seals to not increase trust in electronic commerce (2008). In contrast, a later study reported a high effectiveness of trust sealing, depending on "national or cultural characteristics of consumers" (Kim et al. 2016).

Coming back to AI, trust can also stem from knowledge about aspects (Li et al. 2023) and properties (Fischer, Liebig, and Morik 2024) of models and services, or respective requirements (Kaur et al. 2022). This entails matters of transparency and accountability (Hauer, Krafft, and Zweig 2023), responsibility (Baum et al. 2022; Dignum 2019), or explainability (Langer et al. 2021), which are explored in designated research communities. These discussions fueled the first wave of regulations including the EU AI Act, which prominently promotes "the uptake of human-centric and trustworthy AI" (European Parliament and Council 2024). While the well-established EU General Data Protection Regulation already manifested a "right to explanation in the context of automated decision-making" (Selbst and Powles 2018), empiric research suggests that accurate performance matters more than transparency and interpretability (Nussberger et al. 2022). Lastly, matters of AI ethics and trustworthiness are also connected to the greater scheme of sustainability (Fischer 2025), as AI possesses the power to make our world more sustainable (Vinuesa et al. 2020), however also negatively impacts our society and planet (van Wynsberghe 2021), for example due to high CO_2 emissions (Luccioni, Trevelin, and Mitchell 2024). To that end, it is important to note that the ML research field was shown to follow values "that disfavor societal needs" (Birhane et al. 2022), for example evidenced by the focus on predictive performance (Wanner et al. 2020) and under-reporting of limitations and environmental impact (Liang et al. 2024). Moreover, the corporate concentration of AI power fosters "unrivalled inequality and environmental degradation", a commonly observable phenomenon in the age of "surveillance capitalism" (Sætra 2021).

By now, the importance of AI reporting and documentation (Arnold et al. 2024) for bridging communication gaps, establishing trust, and facilitating sustainable development should have become clear. Fischer, Liebig, and Morik (2024) have categorized reporting types into (1) scientific publications, (2) gray literature such as blogs and software documentation, (3) benchmarks and open databases like *Papers With Code* (Stojnic et al. 2018) or *OpenML* (Vanschoren et al. 2014), and (4) model-specific documentation such as *model cards* (Mitchell et al. 2019), *fact sheets* (Arnold et al. 2019), and (5) high-level *labels* (Fischer et al. 2022; Morik et al. 2022). While all approaches have individual pros and cons, the potential for establishing "communication between ML scientists and stakeholders" (Morik et al. 2022) make labeling particularly interesting. In analogy with the EU energy label (European Commission 2024) or Nutri-scoring (The European Consumer Organisation 2022), AI labels could be designed to conceal the complexity of ML and only report on the relevant practical aspects of trained models, for example with regard to the quality versus resource consumption trade-off (Fischer et al. 2022). As such, AI labeling holds promises for trustworthy and sustainable AI development (Fischer, Liebig, and Morik 2024). The conceptual idea was positively received in the research community (Hanna et al. 2025; Pimenow, Pimenowa, and Prus 2024; Heaton and Fung 2023; Genovesi and Mönig 2022) and some first adaptions have already been proposed (Luccioni et al. 2025; Duran et al. 2024; Castaño et al. 2023), however practicability and effectiveness remain unclear. For once, well-established role model systems are not without limitations—for example, codes and standards were argued to be hard to implement for energy labeling (Solà et al. 2020), and others claim that "there is insufficient scientific evidence" for the effectiveness of Nutri-scoring (Peters and Verhagen 2024). Similar results can be found for trust sealing and certificates in the context of AI: A mixed-methods study found positive effects in "both low- and high-stakes scenarios" (Scharowski et al. 2023), while Wischnewski et al. (2024) did not find any such evidence. In conclusion, AI labeling holds strong potential but also necessitates careful evaluation, which motivates our study.

Methods

We started our evaluation study by formulating the central research questions (see Introduction) and obtaining ethical approval from the ethics committee of the University of Duisburg-Essen Computer Science faculty, also part of the RC Trustworthy Data Science and Security (ID: 2407SPWM1293). To actively foster reproducible research, we make all results, including the interview guide, transcripts, and code system, publicly available at https://github.com/raphischer/labeling-evaluation.

Approach and Recruitment We acquired participants for our semi-structured interviews via a public campaign, which was spread via mailing lists and social media posts on *LinkedIn*, *WhatsApp*, *X*, and *Instagram*. We particularly invited developers of AI systems but also indicated openness to anyone generally interested in the concept of AI labeling, which was abstractly teased with an exemplary figure. While our social networks are naturally biased toward the research community, we were successful in recruiting a total of 16 practitioners from different fields, backgrounds, and levels of AI experience[1]—an overview is given in Table 1. Their level of AI skill was determined via self-assessment, based on some orientation help in our application form: The *beginner* (1 person) has a general idea of but no practical experience with AI, *users* (4) have practical experience with AIaaS, *engineers* (8) have performed basic ML on custom data, and *experts* (3) have extensively trained and deployed ML models[2]. Two of our interviewees hold a doctorate as

[1]Five participants are former colleagues of at least one of the authors, however they are new to the concept of AI labeling.

[2]A *novice* option was offered for candidates without any AI understanding, however no respective application was received.

ID	Job Title	Company Type	Employees	Gender	Age	AI Skills
I1	AI Manager	Industrial Manufacturer	5000-10000	male	—	Engineer
I2	Researcher	Research Service Provider	51-200	male	—	Engineer
I3	Software Developer & Student	IT Service Provider	5000-10000	male	20	Engineer
I4	Student	University	40000-45000	male	21	Beginner
I5	Software Developer & Student	IT Services (self-employed)	1	male	22	Engineer
I6	Solution Engineer	IT Service Provider	50-200	male	28	User
I7	Startup CEO & AI Developer	AI Service Provider	2-10	male	29	Expert
I8	Analytics Platform Manager	Public Service Provider	1000-5000	male	30	Engineer
I9	Software Developer	Lottery Service Provider	50-150	male	31	Expert
I10	Software Developer	IT Services (self-employed)	1	male	31	Expert
I11	Data Scientist	IT & AI Service Provider	11-50	female	32	Engineer
I12	Researcher	Telecommunications	201-500	female	32	User
I13	Development Engineer	Industrial Manufacturer	5000-10000	male	43	Engineer
I14	Maintenance Manager	Public Service Provider	5000-10000	male	46	User
I15	Principal Cloud Engineer	IT Service Provider	51-200	male	47	User
I16	Software Architect	IT Services (self-employed)	1	male	48	Engineer

Table 1: Overview of Interview Participants and Their Jobs and Skills

the highest professional qualification, eight have completed a full master's (or diploma) degree, three have graduated as bachelors (or are certified specialists), and the rest have successfully graduated from high school. Every registered application resulted in an interview, for which the participants were compensated with 15€. After informing them about our anonymization procedure and obtaining written consent, the interviews were conducted via *Zoom* from which the audio data was saved and analyzed. Recruitment was stopped once we collected sufficient information power (Malterud, Siersma, and Guassora 2016). After the interviews, all participants received occasional updates on our study progress.

Interview Structure Guided by our research questions, we developed an interview guide that consisted of four parts. At the beginning of each interview (part one), we asked participants to introduce themselves, explain how their work relates to AI, and describe difficulties they face in their daily business, in order to later answer RQ1. In the second part, the interviewees were presented with prototype AI labels generated with the `STREP` software (Fischer, Liebig, and Morik 2024). It uses a relative scaling and rating approach to visualize the model performance with regard to quantifiable metrics as color-coded icons, and moreover aggregates the performance into a compound score. Additionally, the respective labeling framework also allows to consider user preferences, as the importance of individual performance metrics can be adjusted for compound scoring and visualization. We first depicted a label for `MobileNetV3Small` (Howard et al. 2019), a popular image classification model that is usually used in pretrained form. After discussing first impressions and explaining some of the labeling concepts (if required), we subsequently showed participants a second label featuring an `EfficientNet` variant (Tan and Le 2019). Given both labels, as displayed in Figure 1, participants were asked to compare the information on both labels and explain which aspects they found helpful and confusing, thus relating to RQ2. In the third part, we investigated how labels compare to other types of reporting, as characterized in the Related Work. For answering RQ3, we thus presented and discussed other reports about `MobileNetV3`, particularly the associated research paper (Howard et al. 2019), the model card on *Hugging Face*[3], a blog article[4], the *Keras* software documentation[5], a *Papers With Code* overview[6] (Stojnic et al. 2018), and a comparable *IBM* fact sheet[7] (only available for *IBM* products, so not describing `MobileNetV3`). At the end of the interview, we opened a conversation around trustworthiness in the context of AI labeling, leading into RQ4. The more specific questions discussed possible providers or authorities for labeling models (i.e., who to trust with such a process) as well as the interviewees' position toward certification and regulation. For each interview, at least two interviewers from the author pool were present, benefitting from the multidisciplinarity of our team.

Transcription and Coding In a first step, the recordings were transcribed with the `whisper-large-v3` speech recognition model (Radford et al. 2023), which was locally deployed via the *Shoutout* tool[8] and manually revised for errors. For our qualitative analysis, we efficiently identified recurring patterns and topics by performing inductive, thematic coding with *MAXQDA*[9]. We iteratively discussed individually coded interviews and mutually refined the hierarchically organized coding system. The final system encom-

[3]https://huggingface.co/qualcomm/MobileNet-v3-Small

[4]https://towardsdatascience.com/everything-you-need-to-know-about-mobilenetv3-and-its-comparison-with-previous-versions-a5d5e5a6eeaa

[5]https://keras.io/api/applications/mobilenet/

[6]https://paperswithcode.com/method/mobilenetv3

[7]https://aifs360.res.ibm.com/examples/max_object_detector

[8]https://github.com/RWTH-TIME/shoutout

[9]https://maxqda.com (Version 24.5)

Code Family	RQ	Size	Occ	Quote
General Codes	RQ1	8	63	"To use AI [. . .] to counter the shortage of skilled workers" (I14, p. 4)
Types of Daily Work	RQ1	9	64	"develop an app to detect tolerable products in the supermarket" (I10, p. 4)
AI Use Cases	RQ1	10	41	"monitoring the machine condition such that we can make predictions" (I14, p. 4)
ML Methods	RQ1	7	64	"the AI evaluates whether the typed text contains specific data" (I3, p. 44)
ML Tools & Brands	RQ1	8	36	"I used scikit-learn models and also worked with TensorFlow" (I13, p. 4)
Requirements on AI	RQ1	13	118	"My boss doesn't care much about the process, he wants results" (I13, p. 160)
Benefits	RQ2	12	140	"Your label helps me to decide immediately, it saves a lot of time" (I9, p. 219)
Limitations	RQ2	21	205	"I don't get how the value is included in the overall scoring" (I16, p. 58)
Property Importance	RQ2	5	64	"the primary objectives: reducing time and enhancing accuracy" (I7, p. 98)
Associations	RQ2	3	31	"like I'm looking for a washing machine at the DIY store" (I3, p. 84)
Target Audience	RQ2	1	10	"the addressees are likely to be people who are intensively involved" (I14, p. 64)
Workflows and Use	RQ3	12	61	"different agendas and newsletters as a regular source of information" (I16, p. 70)
General Comparison	RQ3	4	46	"It is time-consuming – that is the disadvantage of other approaches" (I9, p. 219)
Who Needs Trust	RQ4	3	26	"it helps to understand how the model works if you are a developer" (I13, p. 20)
Reasons for Trust	RQ4	9	91	"if it has a university stamp on it, it seems more trustworthy" (I11, p. 144)
Dimensions of Trust	RQ4	11	66	"trust in AI, or trust in a label – these are two different things" (I11, p. 152)
Total		136	1130	

Table 2: Overview of the derived code system with number of occurrences and exemplary quotes

passes 136 codes assigned to a total of 1130 text passages, for which an overview is given in Table 2. It summarizes the number of associated codes (Size) and occurrences for each of the top-level families as well as corresponding exemplary quotes, while the full complexity of our code system can be explored in our supplementary material repository. At the final stage of our code system, we had all authors do a re-coding of two interviews reaching about 90% intercoder agreeability (obtained by the `MAXQDA` code frequency agreement analysis).

Results

Having introduced the details of our study methods, we now present our thematic analysis to answer RQ1–RQ4. For that we investigate the opinions and statements encountered during the respective interview parts, as visually summarized in Figure 2. Each verbatim quote includes interviewee and paragraph references for easy location in the respective transcripts.

Who Is Interested in AI Labeling and What Are Their Problems With AI Technology? (RQ1)

To assess possible user groups of AI labels and their needs, we asked participants to describe their daily AI-related work, which were analyzed toward general problems, types of work, and requirements.

Problems and Types of AI Work Many interviewees mentioned communication and knowledge gaps between different stakeholders as a frequent general problem AI development. I1 described, for example, that it is difficult to "get employees on board so that they can actually use the new [AI] tools" (p. 26), and I11 mentioned issues with "customer communication and expectation management" (p. 50). Despite the overall assessment of AI as a business enhancing technology, interviewees themselves voiced insecurity concerning the use of AI, such as with I15, who mentioned having "quite a few concerns, but on the other hand, I find AI very convenient" (p. 26). Moreover, the interviewees differentiated between AI tools utilized internally during work processes and those integrated into developed products (I12, p. 28), highlighting that AI functionality is frequently employed by employees while simultaneously being offered to customers. This emphasized the scale of a possible labeling approach by also stressing the breadth of applications that needs to be covered.

On similar lines, we encountered a broad spectrum of daily work—software development, infrastructure and operationalization, consulting, as well as data exploration and analysis being most frequently mentioned. Relating to the use of AI within these different streams of work, AIaaS was more frequently encountered while only few interviewees train models on custom data in-house, indicating a shift from a predominantly developer- to a customer-based perspective on AI. This speculation is also supported by some of our interviewees, such as I7, who stated that "when people talk about AI today, they no longer mean deep learning, they mean solely and exclusively large language models" (p. 4). The described AI use cases for in-house service applications are likely linked to this phenomenon. I7 also mentioned that "many companies are not yet ready to implement their own deep learning projects" (p. 8), which explains why tools and brands like `scikit-learn` (for traditional ML, outside of deep learning) and `OpenAI` (for AIaaS) were frequently mentioned.

Requirements on AI

In line with the diversity of possible applications, participants also named a wide range of important requirements for AI systems, with "whether the result works or whether the

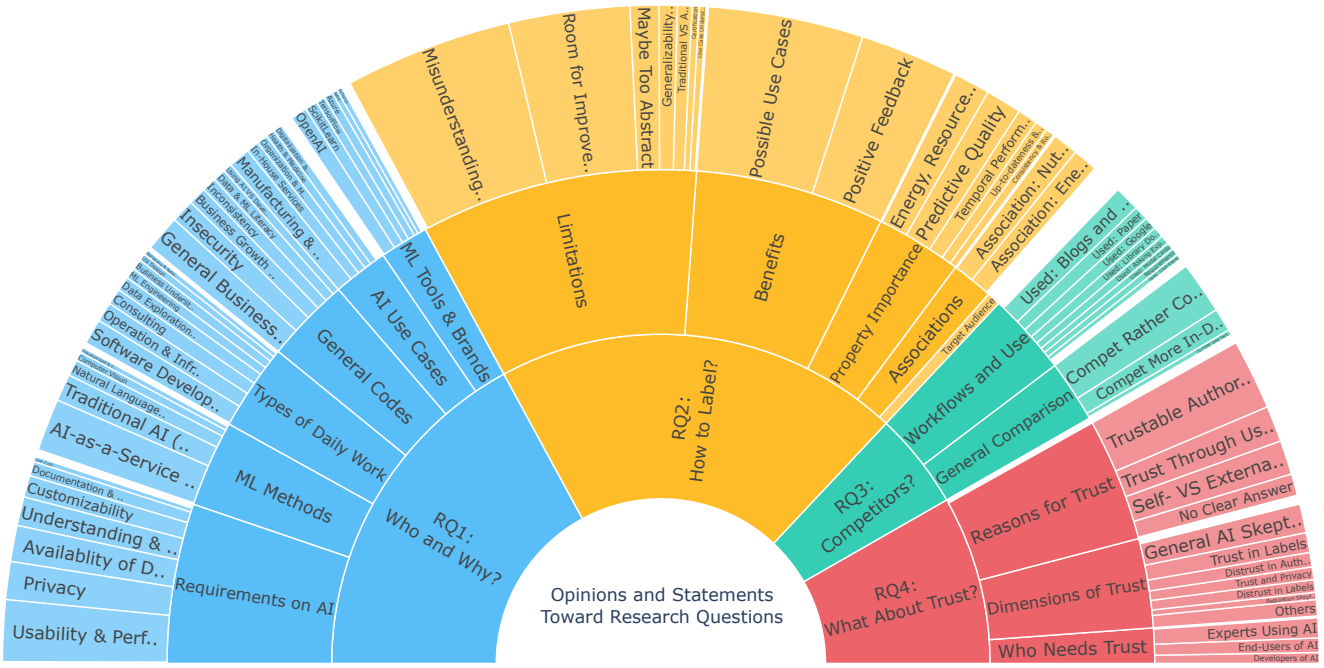

Figure 2: Summary of encountered and analyzed opinions with regard to research questions.

AI itself is bad" (I5, p. 98), or in other words, usability and performance, being especially noteworthy. In this context, most comments referred to performance in terms of predictive quality, while consistency and robustness as well as fast response time received less attention. Privacy also played a major role, with many interviewees discussing the importance of data protection and I11 noting that for this reason her company does not develop AI services that have to deal "with personal data" (p. 14). For AI developers, the importance of data availability was emphasized—"the biggest step in ML and training is data collection, and then, feature engineering and data processing" (I9, p. 40). Additionally, regarding communication about specific models, we found understanding and transparency, customizability, as well as documentation and reporting frequently mentioned as requirements beyond model performance. This includes the importance of "carrying out educational work" for employees (I1, p. 30), archiving "an understanding in terms of what is happening, so that we don't get a black box" (I8, p. 8), and "finding out what is the right model for my area of application" (I4, p. 40).

Overall, we conclude that just as user groups are highly diverse, so are their individual challenges when working with AI. Practitioners come from diverse backgrounds, and while some train custom models, a large amount today simply uses AIaaS. For the context of AI labeling, this indicates the need to consider a diverse range of stakeholders, including users of AI-driven tools, managers overseeing AI implementation, customers interacting with AI-powered solutions, experts configuring and deploying AIaaS offerings, and developers creating new AI models. Each of these practitioner types come with different levels of AI expertise and individual problems in their daily use of AI.

What Are the Practical Benefits and Limitations of Labeling AI Model Behavior? (RQ2)

As a possible solution to some of the discussed challenges, we next presented AI labels and asked interviewees for feedback. The following analyzes the mentioned labeling benefits, limitations, and suggestions for improvements. Generally, we did not encounter strong positions *against* labeling, so critical comments should be understood as evidencing both interest in the idea but also the importance of further refinement.

A Trade-off Question: Simplicity Over Complexity We observed the general tendency that, on the one hand, interviewees deemed the simplicity of the labels helpful and necessary for decision-making and knowledge transfer, while, on the other hand, they also missed more detailed information. For example, I14 perceived our labels as "informative at a glance" (p. 56) and interviewee I1 stated that labels would "help in any case—as there are more and more models, it is increasingly difficult to keep an overview, and the more compact the information is, the better" (I1, p. 100). We encountered over 50 remarks that described benefits for decision processes, in which labels could be used for "comparing different models with each other and seeing how well they perform" (I2, p. 70). Interviewees also saw use cases in the context of communication and knowledge transfer, with I7 mentioning the "greatest added value for customer presentations" (p. 174) and I11 stating that customers "really like pictures and colors [. . .], so you can always catch people with categories like red, yellow and green" (p. 106). On similar lines, one interviewee stated that users "like to have

the basic information presented immediately. And the AI label, with the color-coded scale, presents an excellent way of doing this" (I14, p. 170). Additionally, some interviewees mentioned benefits for transparency (I1, I3, I8, I14), advertisement (I1, I3, I14, I15), and validation or certification of ML results and models (I12, I13, I14).

However contrary to the indicated benefits, some comments indicated that the abstraction of more complex information, while improving accessibility and fast understanding, could result in confusion. Based on their strong expertise with AI, I11 for example anticipates complex factors around models "which I don't think you can necessarily cover in a label" (p. 164). Moreover, the understanding of displayed label metrics were questioned in context of the labeling target audience—interviewees believed that stakeholders without an ML background "likely do not really understand" the rather technical presentation of information, which is, instead, only "interesting for developers" (I3, p. 216). Importantly, our interviewees did not reach a consensus on who our shown labels address: benefits were seen for "people who want to use AI" (I5, p. 228), "decision-makers and customers" (I7, p. 74), or "people who are intensively involved [with AI]" (I14, p. 64).

Adaptability Possibly bridging the tensions of simplicity and in-depth information, interviewees of all proficiency levels appreciated the aforementioned possibility of incorporating the individual user priorities: I16 for example states that "with a label like this, you have very good opportunities to demonstrate what is important to the developer" (p. 50). This interactivity was seen as a helpful means to align development with management expectations: "if I knew what my boss wanted, I would go to the website, set the weighting, press enter and then I would pull out your label" (I13, p. 108). Taking this a step further, our participants also suggested possible improvements to ease the trade-off question, for example, using a fully interactive dashboard "that exactly displays those things that are relevant to you" (I7, p. 166). I13 also emphasized the need for guiding users with their specific use cases and answer questions like "Which model should I use? How do I find my way around?" (p. 72). On similar terms, I2 (as well as I5 and I8) suggested to better inform on "the combination of model quality and application area" (p. 118).

Design Benefits, Limitations, and Improvements Addressing the label design more directly, interviewees described the labels as generally "very consumer-friendly" (I3, p. 84), and "optically appealing" (I7, p. 94). However, participants also voiced various cases of misunderstanding and confusion, mostly in relation to the displayed performance criteria (lower part of labels), and specifically, regarding robustness and accuracy. Here, interviewees could "not really imagine what it means" (I4, p.48) or even mistook robustness (which here relates to adversarial input perturbations (Croce et al. 2021)) for a metric describing "that the model hallucinates very little" (I6, p. 86), relating to a common AIaaS problem that however does not occur with the labeled image classifiers. I7 and I12 even questioned whether the displayed metrics can actually capture the "real experience" of using the model (I12, p. 178) and highlighted the importance of a "qualitative, subjective" (I7, p. 66) evaluation, relating to the common problem of choosing good evaluation metrics (Naidu, Zuva, and Sibanda 2023). Moving beyond the individual metrics, confusion was also voiced with regard to the compound scoring, for example I9 questioned "what weights does this have, for example accuracy or power draw, to decide which categories it belongs to" (p. 134). As a solution to these misunderstandings, participants suggested customizing the labeling procedure (20) by adding a second page "where the evaluation metrics or the parameters are more clearly explained" (I2, p. 122).

Despite various positive comments about the color-coded performance icons, we also encountered some confusion around the underlying relative scoring methodology, for example voiced by I15: "what does it mean for me when it is green?" (p. 48). Interestingly, several interviewees only noticed the color-coding of icons when the second label was visualized, such as I7, who "didn't realize before that the icons at the bottom were color-coded" (p. 78). Moreover, I10 raised concerns for visually impaired people and suggested using different letter sizes instead of color-coded icons. Beyond these points, there was also some misconception regarding the model (`MobileNetV3Small`) and evaluation data (`ImageNet`), which are well-known in visual computing, however, not so much outside of the domain. As other possible improvements, interviewees suggested to feature information on the financial cost of deploying the model, and more prominently display the up-to-dateness of evaluated model results.

Labels as Nudges Facing the labels, we also asked our participants to rate the importance of the displayed properties, namely energy (i.e., resource) consumption, predictive quality (accuracy), temporal performance, and robustness. Ideally, one would want to have a "good mix of all points" (I6, p. 226), however the first two were generally mentioned to be most relevant. Many participants brought up the earlier-mentioned problem of balancing ML quality with resource consumption—on the one hand, "power costs money when [the system] is operationalized" (I13, p. 72), however "when accuracy is key, I would also have to accept higher power draw" (I15, p. 52). These statements can also be connected to the aforementioned AI requirements discussed for RQ1, where predictive quality was generally commented to be most important in the context of AI usability and performance. Interestingly, resource efficiency was hardly mentioned in the more general discussion, however when facing the labels, the interviewees became more resource-aware. They now discussed model performance trade-offs, with I1 reporting sustainability to be "the central driver of [their] corporate strategy" (p. 54) and I2 advocating a general "frugality of system complexity" (p. 74). Echoing the earlier conclusion that labels can aid decision-making processes, we can now add that it is of central importance *which* features a label displays. If decisions are based on all available information at a glance, then including information such as energy efficiency will inadvertently nudge users' attention to these features.

To summarize, the main benefits of labeling lie in the help with decision processes and knowledge transfer. The shown labels were generally well-received but can be further improved—for non-experts, they are too technical and confusing, while for experts, they do not provide the necessary level of depth. As a central takeaway, customizability of labeling is key for benefitting the various target audiences.

How Are AI Labels Perceived in Comparison With Other Forms of Reporting? (RQ3)

Discussing other actively used types of reporting re-emphasized the diversity of AI practitioners (RQ1). Interviewees from all levels of expertise reported to use high-level media and journalistic text to learn about AI, whereas the depth of academic papers was specifically important for developers. This clearly reflects the previously mentioned simplicity versus complexity trade-off. Supporting low-key access to information, one interviewee mentioned that blogs like *Medium* are very popular "because it's often a practical example that is well explained and easy to work with" (I8, p. 116). I13 and I14 highlighted that educational videos help a lot with getting started with ML and I12 explained that blogs and articles help with learning about "trends" and seeing "what others do" (p. 186). Conversely, participants also mentioned issues of trust and reliability for this report type: "What bothers me about *Medium* is that [. . .] anyone can write anything" (I6, p. 182).

Pros and Cons for Different Stakeholders When comparing AI labeling with other reporting types, we found that the fast information access offered by labels was especially pronounced. I9 for example stated that the "disadvantage of all other approaches" is their "time-consuming" nature (p. 219), while labels only showcase the most important points (as also mentioned for RQ2). The competitors are seen as showing "significantly more text, significantly more data" (I3, p. 168), which must be consumed and understood. Even the AI engineers regard this as a drawback: "there is quite a lot of complexity involved and I would first have to have a pretty good understanding of it" (I8, p. 104). In contrast, the interviewees appreciated that once you are familiar with the label, "you don't even have to read [the sources] anymore—you just know how good [the models] are." (I7, p. 110). Having accessible information appears to be of central importance for daily business and I3 highlighted that "none of these [forms of reporting] are as easy to understand as the label" (p. 168).

Contrasting this need for information-at-a-glance, some interviewees, particularly engineers and experts, highlighted the importance of in-depth information. I3 stated, for example, that you "have to look at [the papers], in any case" and I7 specifically scans them for comparisons "with benchmarks and other models" (p. 114). Another engineer mentioned that other reports are "particularly relevant when depth is needed" (I16, p. 86), or as phrased by I3: "When I read a paper about an AI, I can probably understand it much better than if I just look at the label" (p. 172). I2 stated that "we need them all—the difference is, which target groups do I face?" (p. 110) and continued to give examples like papers for scientists, fact sheets and model cards for developers, or blogs for users. This is in line with the argumentation of Fischer, Liebig, and Morik (2024), who understand labeling as an addition to the other "highly important" forms of reporting. Another interviewee suggested seeing labels as some kind of "intermediate solution" (I12, p. 186) for connecting to different groups of people: On the one hand, those who create "technical solutions", on the other hand, "product people", who make them marketable.

In conclusion, the advantages and limitations of different reporting types depend on whether detailed information or just a brief overview is required—labels in this context are a useful addition as they allow for very fast information intake but can also point to other report forms.

How Do AI Labels and the Corresponding Certifying Authority Affect the Trustworthiness of AI Systems? (RQ4)

Establishing trust in AI systems plays a central role when considering the communicative process of labeling and reporting. We generally observed two perspectives in the discussions on trust: On the one hand, interviewees mentioned trust in the labels themselves, as well as in possible issuing authorities. On the other hand, interviewees mentioned the suitability of labels for establishing trust in AI systems, aptly summarized by I11, who differentiated between "trust in AI in general, and trust in a label in terms of a model's performance" (p. 152). For both perspectives, various reasons or origins for trust were named, such as the trustworthiness of the responsible authority, or trust gained from using the system (whether it be the AI itself or the communicating about it via a label). Beyond the direct context of AI labeling, interviewees also mentioned general AI skepticism and regulation skepticism. As an example, I1 reported very different views on AI in their company, "from euphorically enthusiastic to rather skeptically rejecting" (p. 30).

Boosting Trust via Labeling For making AI more trustworthy with the help of labels, most comments were positive, for example I13 thinks that "it would help, yes. Because [the AI label] is approved by professionals and trust is created" (p. 52). However, concerns were also raised, for example I11 doubts that "performance parameters help with such a question of trust" (p. 164), and I7 questioned whether he can truly "rely on such a label because that is such a specific thing" (p. 82), and said he would rather "test [models] himself" (p. 82). In line with the idea of epistemic trust, i.e., the importance of institutions for increasing trust (Wischnewski et al. 2024), are the many comments on suitable labeling authorities. Many interviewees appreciated the idea of having AI labels produced by a "central" (I8, p. 124), "official" (I14, p. 146), and "independent" (I15, p. 122) authority, however struggled to give a clear answer as to who could take this position. Opposing this, nearly half of our interviewees raised concerns of subjectivity, as authorities could be "bribed" (I4, p- 160) or possibly trick the labeling system for a more positive outcome, as it has happened with organic labels (I8, p. 124). Upon the question who should then certify AI, I9 responded fittingly: "quite democratically, the users"

(p. 227). This approach was also greeted by mixed feelings, which we coded as remarks on self-certification versus third-party involvement—I3 for example believes that "there must be something centralized, such that not everyone is allowed to make up their own label" (p. 184), yet others stated that having access to an open certification framework "creates transparency and you can check [...] if it works as I imagine it will" (I5, p. 252). I8 takes a slightly different perspective on epistemic trust, instead focusing on practical performance and believing that "what works well, what is well explained, counts more than who published it" (p. 116). This is also supported by I4, who argues that "most people probably don't care [about properly understanding the system]. The main thing is that the end result is correct" (p. 180), mirroring empirical results from related work (Nussberger et al. 2022).

Dimensions of Trust Exploring whether AI labels can be a means to create trust in AI necessitates to distinguish between different target audiences. We found that our participants anticipated different trust requirements, depending on the trustees' levels of AI proficiency. I11 highlights the perspective of end-users, who have minimal influence on AI model functionality, as particularly significant, noting that "as a user of an AI, then of course I have the least trust" (p. 186). This is contradicted by I4, who assumes that for such stakeholders the AI details are "so far in the background that most people probably don't care" (p. 180). In this context, the potential of labels on fostering trust was perceived as twofold: On one hand, concerns were raised that AI end-users might feel overwhelmed or disengaged due to technical details (I15, p. 106; I1, p. 96). On the other hand, I6 positively noted that displayed metrics could make AI model performance more comprehensible and tangible (p. 210).

Similarly to end-users, experts who proficiently use AIaaS but do not develop AI models themselves were reported to experience trust issues when working with pre-trained models offered as AIaaS (I11, p. 178). Regarding labels, I13 here emphasized that understanding the intricate functionality of a model may not be necessary: "You have to ask yourself again, to what extent do I need to know (the model's functionality)? And to what extent only the application of the model. Since you don't have to develop it further" (p. 20). This suggests that AI labels could be a helpful tool for informing AIaaS users in order to boost trust. In contrast to users, AI developers were noted as inherently trusting the systems they build: "I don't have a problem with trust in the sense that I'm the person who decides what kind of model to use" (I11, p. 178). However, I8 for example also reported that developers rely on implementing use-case-specific explainability methods tailored to their needs in order to ensure confidence in their model's outputs.

To answer our final research question, trustworthiness is a broad problem with various dimensions and perspectives. The two biggest factors for increasing trust seem to reside in responsible authorities and personal experience (i.e., from using available systems). Importantly, interviewees were not united in their positions as to who could possibly be a good labeling authority and actively discussed the trustworthiness and bias potential of candidates such as companies, researchers, open source communities, or governments.

Discussion

Our analysis revealed several key themes for AI labeling practices, which will be critically discussed in the following: The inherent trade-offs involved in designing labels (RQ2 & RQ3), the manifested understanding of labels as sustainability nudges (RQ2), the ongoing challenge of trust in labels and their certifying authorities (RQ4), and lastly, the diverse needs and expectations of practitioners (RQ1 & RQ3). Each aspect presents challenges and opportunities for improving AI transparency, communication, and trust. Moreover, we discuss limitations and ideas for future work.

A critical theme that emerged from our study revolves around the trade-off between simplicity and depth, as discussed for RQ2. Across all levels of expertise, participants agreed on the need for simplification of AI information to facilitate quick decision-making and communication. However, interviewees also expressed concerns of oversimplification and acknowledged the limitations of high-level labeling, especially when it comes to capturing the nuances of model performance and application suitability. While labels are meant to distill complex, often highly technical information into digestible, accessible formats (Morik et al. 2022), our study evidences that this abstraction could also negatively impact stakeholders' understanding and trust. The resulting tension highlights a central challenge for how to design AI labeling systems:

Labels must strike a balance between providing an overview that is both accessible and meaningful without sacrificing important detail. In that context it is of utmost importance to acknowledge and characterize the "desiderata of the various classes of stakeholders", as for example also mentioned in the context of explainability (Langer et al. 2021). If this is established, interactivity could provide a potential solution to balancing simplicity and complexity, allowing for a more dynamic, user-driven label experience. In addition, it is important to acknowledge the connections between labeling and other forms of reporting (RQ3)—by linking multiple representations (Fischer, Liebig, and Morik 2024), interested users can dive deeper into the intricacies which might reinforce the effectiveness and trustworthiness of labels.

While not explicitly focused in our study, it became clear that another important role of AI labels is their potential as nudges, influencing user decisions by emphasizing certain aspects of model performance (RQ2). Our results indicate that labels can function as a tool for guiding decision-making by drawing attention to key trade-offs between model attributes such as accuracy and energy consumption. This connects to psychological literature, such that labels could function as "signals which are actions that parties take to reveal their true type" (Kirmani and Rao 2000), and moreover can be interpreted as persuasive cues, either peripherally or centrally (Petty, Cacioppo et al. 1984). This entails:

Labels do more than simply present information—they actively shape the decision-making process by highlighting the factors deemed most important. In the sus-

tainability context, this could potentially shift the questionable focus on predictive capabilities (Birhane et al. 2022; Wanner et al. 2020) toward environmental awareness (Liang et al. 2024; Luccioni, Trevelin, and Mitchell 2024), and thus, trustworthiness (Fischer, Liebig, and Morik 2024).

Indeed, our findings manifest the close connection between AI labeling and trustworthiness, however also reinforce the ambivalence of this relation. Our analysis evidences that trust in AI and trust in AI labels are linked, but should be investigated and fostered separately. Regarding the label's trustworthiness, participants highlighted the importance of clear, reliable metrics, but also expressed skepticism about the adequacy of labels to fully represent the complexity of AI models. For experts, labels serve as a starting point for decision-making, but they are not a substitute for hands-on testing or exploring technical details. As in the related literature (Wischnewski et al. 2024), the question of authoritative responsibility was contentiously discussed—participants suggested that a neutral, centralized authority (e.g., independent regulatory bodies or academia) could lend legitimacy to AI labels, however concerns were raised regarding subjectivity and potential for bias. Others advocated a more democratized approach, suggesting that developers themselves could play a key role in evaluating and certifying AI models, linking to the idea of open science and open source AI (Widder et al. 2022). We understand the variety of these opinions to be a symptom originating from "surveillance capitalism" (Sætra 2021), as big tech companies currently yield the power of AI, however might not be considered trustworthy in doing so. Consequently, we posit:

Labels need to be seen as part of a larger trust-building process that involves transparency, verifiability, and user experience. While not a guarantee for boosting trust, refining the conceptual label idea under consideration of stakeholder diversity could be valuable for appropriate trust calibration (Wischnewski, Krämer, and Müller 2023; Mehrotra et al. 2024). As others have postulated in the context of sustainability (Luccioni, Trevelin, and Mitchell 2024), we believe that this evolution requires a "a variety of technical, behavioral and organizational interventions", connecting authorities such as researchers, AIaaS providers, and policy makers.

Lastly, the striking diversity in participants' backgrounds, roles, and expertise underscores the necessity for AI labels to be adaptable to different user groups and contexts. Our analysis showcased that AI proficiency acts as a moderating effect on positions toward AI labeling, connecting to similar findings in the context of chatbot adoption (Saihi, Ben-Daya, and Moncer 2024) and trust sealing, where the effectiveness is impacted by consumer characteristics (Kim et al. 2016). While labels were claimed to be a promising tool for simplifying AI-related communication (Morik et al. 2022), the broad spectrum of users, from technical experts to non-technical stakeholders, indicates that a "one-size-fits-all" approach will also likely fall short, as it is the case with different AI services (Brecker et al. 2023). Accordingly, we argue that any unified approach needs to acknowledge diversity:

Labels must allow for customization, ensuring that different audiences can extract the information they need. This need for adaptability aligns with previous research stating that AI reporting must consider varying audience expertise levels (Fischer, Liebig, and Morik 2024) and that AI documentation practices should be holistic and adaptable (Arnold et al. 2024). Another connection can be made toward AI explainability, where Miller (2019) argued that "explanations are social" and thus require evaluation and interactiveness. In our opinion, this also applies to labeling, which could even be understood as a global explanation of important model properties. In practice, future AI labeling frameworks therefore should be designed flexible, for example allowing users to navigate the level of detail they wish to see. As another example, AI engineers, who inherently trust the models the built, could potentially boost external trust by transparently communicating the exact performance aspects that users are interested in. This connects to the challenges of communicating performance aspects via trust seals (Scharowski et al. 2023) and evaluating the suitability of ML metrics (Naidu, Zuva, and Sibanda 2023) for manifesting trust (Li et al. 2023). We conclude that an interactive AI labeling framework could potentially and facilitate sustainable development by properly connecting engineers, AIaaS experts, users, and other stakeholders.

While our study provides many valuable insights, we want to lastly address some potential limitations (see also the author statements below). Recruiting participants via social media might have caused a sampling bias due to attracting those already interested in AI labeling while excluding skeptics. Additionally, creating and presenting specific labels may have introduced response bias due to social desirability. The visual similarities to energy and Nutri-score labels could have further influenced participants based on prior experiences with these systems. We also see large potential for future work—for example, our results allow to postulate design principles and develop better labeling frameworks (van der Staay et al. 2025). Putting our qualitative findings to the test, future research could also test AI labels in real-world contexts, possibly exploring alternative designs or skeptical perspectives and focusing on practical implementation and impact (Gansky and McDonald 2022). This validation could be conducted quantitatively, for example using recently proposed metrics for trustworthiness (Rutinowski et al. 2024).

Conclusion

Our study highlights the multifaceted role of AI labeling for fostering trust and informed decision-making across different user groups. We found evidence that AI labels can be valuable due to their accessibility and potential to transfer knowledge, however must overcome challenges related to audience diversity and technical comprehension. To maximize their impact, AI labeling systems should incorporate interactive features that allow for customization based on stakeholder priorities and knowledge. Moreover, independent certification processes are essential to bolster trustworthiness of labeling. By integrating these improvements, AI labels can establish transparency and benefit AI sustainability, ultimately aligning technical advancements with societal expectations.

Ethical Statement

Our research follows a proactive approach to address ethical concerns, as we obtained an ethical approval from the ethics committee of the University of Duisburg-Essen Computer Science faculty (ID: 2407SPWM1293). Participation in the interviews study was open and voluntary, and we were upfront about our personal intentions and the study progress up to the current point of publication. We also made sure to obtain written consent from all interviewees before recording, and only published fully anonymized versions of the interview transcripts. Embracing the idea of reproducible research, we make all materials for our analysis publicly available, fostering transparency and allowing for deeper insights.

With regard to the ethical situation concerning automation and AI technology, we want to once more point to the important related literature that our work builds upon (Chatila et al. 2021; Genovesi and Mönig 2022; van Wynsberghe 2021; Sætra 2021). Labeling of AI models was argued to potentially establish more transparency and trust (Fischer, Liebig, and Morik 2024), and by evaluating the concept for practicability, we hope to make an important contribution toward ethical AI development. In that context, the study also discusses specific points of improvements, such as the explicit consideration of visually impaired stakeholders.

Positionality Statement

Our individual backgrounds naturally influence our perspectives on the evaluation of labeling practices. Coming from different research disciplines (computer science, psychology, social science), we aimed at establishing a common perspective at the beginning of this study project. These initial discussions on AI and labeling in particular had the explicit goal of mitigating personal bias and ultimately resulted in the discussed research questions and publicly available interview guide.

The labels showcased during interviews were created with the help of the `STREP` framework (Fischer, Liebig, and Morik 2024), which represents an earlier work of two of the authors. While their research aims at making the field more sustainable and trustworthy, we believe that no further personal biases were introduced—the label design is mostly influenced by the well-known EU energy labels and the properties of the displayed AI image classifiers were assessed in an objective way (Fischer et al. 2022).

Not all authors were present in all interviews, but we made sure to always include a computer scientist who understands labeling on the technical level. The code analysis of the interviews included multiple points of discussion and synchronization, resulting in a strong common perspective which is evidenced by the high intercoder agreeability. Note that our personal backgrounds and positionality were also openly discussed at the start of each interviews. In addition, we embrace the spirit of open science by openly discussing the limitations of our work and making supplementary materials available to fellow researchers.

Broader Impact

Although not directly discussed within the study itself, we would like to point out that AI labeling could potentially have broader, and potentially adverse, impacts. Others have already discussed that novel AI models could for example also be used for malicious purposes (Rohde et al. 2024), which also holds true to our work. To be more specific, our study highlights that flexibility and adaptability should be embraced by future labeling systems, however this also opens opportunities for maliciously concealing specific trade-offs, which could for example result in greenwashing. If users overly trust AI labels without critical evaluation, it could potentially also result in the uncritical acceptance and use of flawed or biased AI systems. Analogously to poorly designed explainable AI methods, poorly designed labels could result in questionable model choices and even privacy violations, manipulation, and ultimately, reduced AI adoption (Martens et al. 2025). Lastly, while labeling holds promise for making negative model properties transparent, this requires to first find ways to assess and quantify them. It is therefore important to understand labels as a piece of the much larger puzzle of advancing ethical, sustainable, and trustworthy AI.

Acknowledgments

RF, KP, and TL have been funded by the German Federal Ministry of Research, Technology and Space (BMFTR) and the state of North Rhine-Westphalia as part of the *Lamarr Institute for Machine Learning and Artificial Intelligence* (`Lamarr25B`). MW has been funded by the University Alliance Ruhr as part of the *Research Center Trustworthy Data Science and Security* (https://rc-trust.ai). AV and CJ have been funded by the German Federal Ministry of Research, Technology and Space (BMFTR) within the *Zukunft der Wertschöpfung – Forschung zu Produktion, Dienstleistung und Arbeit* (`02K23A070`) and managed by Projektträger Karlsruhe (PTKA). The authors are responsible for the contents of this publication.

References

Arnold, M.; Bellamy, R. K.; Hind, M.; Houde, S.; Mehta, S.; et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5): 6–1.

Arnold, S.; Yesilbas, D.; Gröbner, R.; Riedelbauch, D.; Horn, M.; and Weinzierl, S. 2024. Documentation Practices of Artificial Intelligence.

Baum, K.; Mantel, S.; Schmidt, E.; and Speith, T. 2022. From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology*, 35(1): 12.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 173–184. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2. Event-place: Seoul, Republic of Korea.

Boag, S.; Dube, P.; El Maghraoui, K.; Herta, B.; Hummer, W.; et al. 2018. Dependability in a multi-tenant multi-framework deep learning as-a-service platform. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 43–46. IEEE.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; et al. 2022. On the Opportunities and Risks of Foundation Models.

Brecker, K.; Lins, S.; Trenz, M.; and Sunyaev, A. 2023. Artificial Intelligence as a Service: Trade-Offs Impacting Service Design and Selection. In *Proceedings of the 2023 International Conference on Information Systems (ICIS)*.

Castaño, J.; Martínez-Fernández, S.; Franch, X.; and Bogner, J. 2023. Exploring the Carbon Footprint of Hugging Face's ML Models: A Repository Mining Study. In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1–12. IEEE.

Chaaben, E. B. 2024. Exploring Human-AI Collaboration and Explainability for Sustainable ML. In Lorig, F.; Tucker, J.; Lindström, A. D.; Dignum, F.; Murukannaiah, P. K.; Theodorou, A.; and Yolum, P., eds., *HHAI 2024: Hybrid Human AI Systems for the Social Good - Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence, Malmö, Sweden, 10-14 June 2024*, volume 386 of *Frontiers in Artificial Intelligence and Applications*, 363–370. IOS Press.

Chatila, R.; Dignum, V.; Fisher, M.; Giannotti, F.; Morik, K.; et al. 2021. Trustworthy AI. *Reflections on Artificial Intelligence for Humanity*, 13–39.

Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; et al. 2021. RobustBench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Dignum, V. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 2156. Springer.

Duran, P.; Castaño, J.; Gómez, C.; and Martínez-Fernández, S. 2024. GAISSALabel: A Tool for Energy Labeling of ML Models. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, FSE 2024, 622–626. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706585. Event-place: Porto de Galinhas, Brazil.

European Commission. 2024. Understanding the Energy Label. Accessed: 2024-12-29.

European Parliament and Council. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union.

Feuerriegel, S.; Hartmann, J.; Janiesch, C.; and Zschech, P. 2024. Generative AI. *Business & Information Systems Engineering*, 66(1): 111–126.

Fischer, R. 2025. *Advancing the Sustainability of Machine Learning and Artificial Intelligence via Labeling and Meta-Learning*. Ph.D. thesis, TU Dortmund University.

Fischer, R.; Jakobs, M.; Mücke, S.; and Morik, K. 2022. A Unified Framework for Assessing Energy Efficiency of Machine Learning. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 39–54. Cham: Springer Nature Switzerland.

Fischer, R.; Liebig, T.; and Morik, K. 2024. Towards more sustainable and trustworthy reporting in machine learning. *Data Mining and Knowledge Discovery*.

Fischer, R.; Pauly, A.; Wilking, R.; Kini, A.; and Graurock, D. 2023. Prioritization of Identified Data Science Use Cases in Industrial Manufacturing via C-EDIF Scoring. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–4.

Fischer, R.; Wever, M.; Buschjäger, S.; and Liebig, T. 2024. MetaQuRe: Meta-learning from Model Quality and Resource Consumption. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, 209–226. Cham: Springer Nature Switzerland. ISBN 978-3-031-70368-3.

Gansky, B.; and McDonald, S. 2022. CounterFAccTual: How FAccT Undermines Its Organizing Principles. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1982–1992. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2. Event-place: Seoul, Republic of Korea.

Genovesi, S.; and Mönig, J. M. 2022. Acknowledging sustainability in the framework of ethical certification for AI. *Sustainability*, 14(7): 4157.

Hanna, M. G.; Pantanowitz, L.; Dash, R.; Harrison, J. H.; Deebajah, M.; Pantanowitz, J.; and Rashidi, H. H. 2025. Future of Artificial Intelligence—Machine Learning Trends in Pathology and Medicine. *Modern Pathology*, 38(4): 100705.

Hauer, M. P.; Krafft, T. D.; and Zweig, K. 2023. Overview of transparency and inspectability mechanisms to achieve accountability of artificial intelligence systems. *Data & Policy*, 5: e36. Edition: 2023/11/24 Publisher: Cambridge University Press.

Heaton, H.; and Fung, S. W. 2023. Explainable AI via learning to optimize. *Scientific Reports*, 13(1): 10103.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; et al. 2019. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Kaur, D.; Uslu, S.; Rittichier, K. J.; and Durresi, A. 2022. Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, 55(2).

Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Wortman Vaughan, J. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14.

Kim, D. J.; Ferrin, D. L.; and Rao, H. R. 2008. A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision support systems*, 44(2): 544–564.

Kim, D. J.; Yim, M.-S.; Sugumaran, V.; and Rao, H. R. 2016. Web assurance seal services, trust and consumers' concerns: An investigation of e-commerce transaction intentions across two nations. *European Journal of Information Systems*, 25(3): 252–273.

Kirmani, A.; and Rao, A. R. 2000. No pain, no gain: A critical review of the literature on signaling unobservable product quality. *Journal of marketing*, 64(2): 66–79.

Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; et al. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296: 103473.

Lansing, J.; and Sunyaev, A. 2016. Trust in Cloud Computing: Conceptual Typology and Trust-Building Antecedents. *SIGMIS Database*, 47(2): 58–96.

Lee, J. D.; and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1): 50–80. Publisher: SAGE Publications Inc.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; et al. 2023. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55(9).

Liang, W.; Rajani, N.; Yang, X.; Ozoani, E.; Wu, E.; Chen, Y.; Smith, D. S.; and Zou, J. 2024. Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI. *Nature Machine Intelligence*, 6(7): 744–753.

Lins, S.; Pandl, K. D.; Teigeler, H.; Thiebes, S.; Bayer, C.; and Sunyaev, A. 2021. Artificial intelligence as a service: classification and research directions. *Business & Information Systems Engineering*, 63: 441–456.

Luccioni, A. S.; Gamazaychikov, B.; Strubell, E.; Hooker, S.; Jernite, Y.; Mitchell, M.; and Wu, C.-J. 2025. *AI Energy Score Documentation*.

Luccioni, A. S.; Trevelin, B.; and Mitchell, M. 2024. The Environmental Impacts of AI – Policy Primer. In *Hugging Face Blog*.

Malterud, K.; Siersma, V. D.; and Guassora, A. D. 2016. Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative Health Research*, 26(13): 1753–1760. PMID: 26613970.

Martens, D.; Shmueli, G.; Evgeniou, T.; Bauer, K.; Janiesch, C.; Feuerriegel, S.; Gabel, S.; Goethals, S.; Greene, T.; Klein, N.; et al. 2025. Beware of" explanations" of AI. *arXiv preprint arXiv:2504.06791*.

Mcknight, D. H.; Carter, M.; Thatcher, J. B.; and Clay, P. F. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, 2(2): 1–25.

Mehrotra, S.; Degachi, C.; Vereschak, O.; Jonker, C. M.; and Tielman, M. L. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM Journal on Responsible Computing*, 1(4).

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; et al. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*, 220–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Morik, K. J.; Kotthaus, H.; Fischer, R.; Mücke, S.; Jakobs, M.; et al. 2022. Yes We Care! - Certification for Machine Learning Methods Through the Care Label Framework. *Frontiers in Artificial Intelligence*, 5.

Nahar, N.; Zhang, H.; Lewis, G.; Zhou, S.; and Kästner, C. 2023. A Meta-Summary of Challenges in Building Products with ML Components – Collecting Experiences from 4758+ Practitioners. In *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, 171–183.

Naidu, G.; Zuva, T.; and Sibanda, E. M. 2023. A Review of Evaluation Metrics in Machine Learning Algorithms. In *Artificial Intelligence Application in Networks and Systems*, 15–25. Cham: Springer International Publishing. ISBN 978-3-031-35314-7.

Ng, D. T. K.; Leung, J. K. L.; Chu, S. K. W.; and Qiao, M. S. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2: 100041.

Nussberger, A.-M.; Luo, L.; Celis, L. E.; and Crockett, M. J. 2022. Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence. *Nature Communications*, 13(1): 5821.

Pandl, K. D.; Teigeler, H.; Lins, S.; Thiebes, S.; and Sunyaev, A. 2021. Drivers and Inhibitors for Organizations' Intention to Adopt Artificial Intelligence as a Service. In *Hawaii International Conference on System Sciences*.

Parasuraman, R.; and Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2): 230–253.

Peters, S.; and Verhagen, H. 2024. Publication bias and Nutri-Score: A complete literature review of the substantiation of the effectiveness of the front-of-pack logo Nutri-Score. *PharmaNutrition*, 27: 100380.

Petty, R. E.; Cacioppo, J. T.; et al. 1984. Source factors and the elaboration likelihood model of persuasion. *Advances in consumer research*, 11(1): 668–672.

Pimenow, S.; Pimenowa, O.; and Prus, P. 2024. Challenges of Artificial Intelligence Development in the Context of Energy Consumption and Impact on Climate Change. *Energies*, 17(23).

Piorkowski, D.; Park, S.; Wang, A. Y.; Wang, D.; Muller, M.; and Portnoy, F. 2021. How AI Developers Overcome Communication Challenges in a Multidisciplinary Team: A Case Study. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–25. Publisher: ACM New York, NY, USA.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; Mcleavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the*

40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, 28492–28518. PMLR.

Rohde, F.; Wagner, J.; Meyer, A.; Reinhard, P.; Voss, M.; et al. 2024. Broadening the perspective for sustainable artificial intelligence: sustainability criteria and indicators for Artificial Intelligence systems. *Current Opinion in Environmental Sustainability*, 66: 101411.

Rutinowski, J.; Klüttermann, S.; Endendyk, J.; Reining, C.; and Müller, E. 2024. Benchmarking Trust: A Metric for Trustworthy Machine Learning. In Longo, L.; Lapuschkin, S.; and Seifert, C., eds., *Explainable Artificial Intelligence*, 287–307. Cham: Springer Nature Switzerland. ISBN 978-3-031-63787-2.

Saihi, A.; Ben-Daya, M.; and Moncer, H. 2024. The moderating role of technology proficiency and academic discipline in AI-chatbot adoption within higher education: Insights from a PLS-SEM analysis. *Education and Information Technologies*, 1–39.

Scharowski, N.; Benk, M.; Kühne, S. J.; Wettstein, L.; and Brühlmann, F. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 248–260. Association for Computing Machinery. ISBN 9798400701924.

Schmitz, A.; Akila, M.; Hecker, D.; Poretschkin, M.; and Wrobel, S. 2022. The why and how of trustworthy AI. *An approach for systematic quality assurance when working with ML components*, 70(9): 793–804.

Selbst, A.; and Powles, J. 2018. "Meaningful Information" and the Right to Explanation. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 48–48. PMLR.

Solà, M. d. M.; de Ayala, A.; Galarraga, I.; and Escapa, M. 2020. Promoting energy efficiency at household level: a literature review. *Energy Efficiency*, 14(1): 6.

Stojnic, R.; Taylor, R.; Kardas, M.; Saravia, E.; Cucurull, G.; et al. 2018. Papers With Code - The latest in Machine Learning.

Sætra, H. S. 2021. AI in Context and the Sustainable Development Goals: Factoring in the Unsustainability of the Sociotechnical System. *Sustainability*, 13(4).

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6105–6114. PMLR.

The European Consumer Organisation. 2022. Revision of EU Legislation on Food Information To Consumers. Accessed: 2025-04-11.

van der Staay, A.; Fischer, R.; Wischnewski, M.; Poitz, K.; and Janiesch, C. 2025. Reflective Design Theorizing with User Interviews: A Case Study for AI Energy Labels. In Chatterjee, S.; vom Brocke, J.; and Anderson, R., eds., *Local Solutions for Global Challenges*, 66–80. Cham: Springer Nature Switzerland. ISBN 978-3-031-93979-2.

van Wynsberghe, A. 2021. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3): 213–218.

Vanschoren, J.; Van Rijn, J. N.; Bischl, B.; and Torgo, L. 2014. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*.

Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; et al. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1): 233.

Wanner, J.; Heinrich, K.; Janiesch, C.; and Zschech, P. 2020. How Much AI Do You Require? Decision Factors for Adopting AI Technology. In *Proceedings of the 41st International Conference on Information Systems*. ISBN 978-1-73363-255-3.

Widder, D. G.; Nafus, D.; Dabbish, L.; and Herbsleb, J. 2022. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 2035–2046. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Wischnewski, M.; Krämer, N.; Janiesch, C.; Müller, E.; Schnitzler, T.; and Newen, C. 2024. In Seal We Trust? Investigating the Effect of Certifications on Perceived Trustworthiness of AI Systems. *Human-Machine Communication*, 8(1): 7.

Wischnewski, M.; Krämer, N.; and Müller, E. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5. Event-place: Hamburg, Germany.

Yang, Q.; Suh, J.; Chen, N.-C.; and Ramos, G. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 designing interactive systems conference*, 573–584.