ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies

Viviane Herdel[1], **Sanja Šćepanović**[2], **Edyta Bogucka**[2], **Daniele Quercia**[2,3]
[1]Ben-Gurion University of the Negev, Beer Sheva, Israel
[2]Nokia Bell Labs, Cambridge, UK,
[3]Kings College London, London, UK
herdel@post.bgu.ac.il, {sanja.scepanovic, edyta.bogucka, daniele.quercia}@nokia-bell-labs.com

Abstract

Responsible AI design is increasingly seen as an imperative by both AI developers and AI compliance experts. One of the key tasks is envisioning AI technology uses and risks. Recent studies on the model and data cards reveal that AI practitioners struggle with this task due to its inherently challenging nature. Here, we demonstrate that leveraging a Large Language Model (LLM) can support AI practitioners in this task by enabling reflexivity, brainstorming, and deliberation, especially in the early design stages of the AI development process. We developed an LLM framework, *ExploreGen*, which generates realistic and varied uses of AI technology, including those overlooked by research, and classifies their risk level based on the EU AI Act regulation. We evaluated our framework using the case of Facial Recognition and Analysis technology in nine user studies with 25 AI practitioners. Our findings show that *ExploreGen* is helpful to both developers and compliance experts. They rated the uses as realistic and their risk classification as accurate (94.5%). Moreover, while unfamiliar with many of the uses, they rated them as having high adoption potential and transformational impact.

Introduction

In today's fast-paced tech world, balancing innovation with responsibility is essential (Sraml Gonzalez and Gulbrandsen 2022; Owen and Pansera 2019). As Artificial Intelligence (AI) spreads across areas like healthcare and finance, it is crucial to understand its uses and potential risks relating, e.g., to data privacy, security, and fairness (Davenport and Kalakota 2019; Goodell et al. 2021; Dignum 2019; Tahaei et al. 2023). Business developers and engineers seek opportunities to employ the latest AI trends ahead of their competitors (Phaal, Farrukh, and Probert 2004), while researchers take part in a similarly fast-paced environment to publish their latest AI discoveries. In both roles, these AI practitioners are faced with increased need to envision potential uses, as well as risks and benefits of the technologies they are developing, and to produce impact assessment reports (Stahl et al. 2023). Given the increasing number of AI regulations (Smuha 2021), compliance experts also face the task of supporting their colleagues in assessing the regulatory risks and compliance of AI technologies. The process of cataloging AI uses and associated risks is both challenging and time-consuming (Moraes, Almeida, and de Pereira 2021; Hassel and Özkiziltan 2023). Recent research shows that AI developers struggle with detailing uses and impacts for model cards (Liang et al. 2024) and data cards (Yang, Liang, and Zou 2024), as well as for the broader societal impacts sections now mandated by some of the top AI conferences (Nanayakkara, Hullman, and Diakopoulos 2021; Prunkl et al. 2021; Ashurst et al. 2022). Recommendations to support AI practitioners with envisioning the impacts of their technology include encouraging reflexivity, including constructive and data-driven deliberation (Ashurst et al. 2022; Prunkl et al. 2021; Yang, Liang, and Zou 2024).

Our research responds to this challenge by exploring the use of Large Language Models (LLMs) to generate AI technology uses and their risk assessments based on the EU AI Act (European Comission 2024). This aims to support AI practitioners during the initial phases of the AI design process, including reflexivity, brainstorming, and deliberation. While LLMs have demonstrated utility in diverse applications (Gilardi, Alizadeh, and Kubli 2023; Wu, Terry, and Cai 2022; Dowling and Lucey 2023; Byun, Vasicek, and Seppi 2023), their suitability for two specific tasks—identifying potential uses of a given AI technology and conducting legal risk assessments of its uses—remains an open question. Our aim is *not* to produce an exhaustive list of uses for a given AI technology, nor to provide a definitive risk classification. Instead, we aim to investigate whether LLMs can generate outputs of sufficient quality to support AI practitioners in envisioning the impacts of their technology, particularly focusing on *less well-researched uses*. On one hand, LLMs might generate unrealistic use cases or ones that practitioners are already familiar with. On the other hand, the extent to which LLMs can accurately map legal regulations to specific AI uses, if at all, is yet to be substantiated.

This paper aims to evaluate LLMs for these specific goals. We explored them using OpenAI's GPT-4 (OpenAI 2023), making two main contributions (Figure 1):

1. We designed an LLM framework (*ExploreGen*) incorporating novel prompt elements—a set of curated *domains* to generate a variety of uses, and *risk concepts* proposed by Golpayegani, Pandit, and Lewis (2023), framing each use along these concepts for risk assessment (*UsesGen*). *UsesGen* classifies generated uses into realistic (existing

Project website: https://social-dynamics.net/explore-gen

and upcoming) and unlikely (hallucinations) with Chain-of-Thought (CoT) reasoning (Wei et al. 2022), retaining only *realistic* ones. These uses are then classified into prohibited, high-risk, and limited or low-risk categories according to the EU AI Act (*RiskLabelling*). Additionally, we processed 3M Semantic Scholar papers, to uncover ∼12% among the identified uses, which were overlooked by the scientific literature (*Literature Coverage Analyzer*).

2. Using Facial Recognition and Analysis (FRA) technology as a use case, we *evaluated* our framework on six aspects: *(I)* whether it generates realistic uses, *(II)* literature coverage of the generated uses, *(III)* familiarity of AI practitioners with these uses, *(IV)* adoption potential, *(V)* transformational impact, and *(VI)* accuracy of risk classification and perceived riskiness by the practitioners.

   To perform the evaluation, we conducted a scoping literature review, and 9 user studies with 25 AI practitioners (12 AI developers and 13 AI compliance experts). We found that *UsesGen* generated realistic uses, covering 96% of the literature uses identified through the scoping review *(I-II)*. AI practitioners reported low familiarity with the uses, especially the overlooked ones *(III)*. They considered the uses somewhat to very likely to be adopted *(IV)* and to have a high transformational impact on business operations or people's lives *(V)*. Compliance experts found that *RiskLabelling* correctly classified the risk of uses based on the EU AI Act with a 94.5% accuracy. Although over 50% of the FRA uses were classified as high risk or prohibited, AI developers, who were not presented with the classification, perceived most uses as only slightly risky for society and not at all for the environment. Lastly, thematic analysis of open-ended responses during in-person interviews revealed that both AI developers and compliance experts found *ExploreGen* helpful for ideation, brainstorming, and deliberation of AI uses and their risks and benefits. Compliance experts found it directly useful, while developers recommended adjustments to better suit their needs.

Background & Related Work

First, we present background on assessing AI impacts, followed by a glimpse of emergent AI regulations, and finish with prior work leveraging LLMs for various tasks.

Assessing Impacts of AI Technology

AI impact assessments (AIIAs) are recommended as a tool to recognize both the beneficial and adverse effects early in the AI technology development process, aiming to predict and evaluate the impact that new digital technologies have on all stakeholders. Stahl et al. (2023) reviewed literature and identified 38 proposed AIIAs, including DataSheets for Datasets (Gebru et al. 2021) and methods inspired by environmental impact assessments (Calvo, Peters, and Cave 2020). However, despite the proliferation of proposed AIIAs, developer teams often encounter difficulties initiating AI impact assessments (Buçinca et al. 2023) and require additional guidance throughout this process (Wang et al. 2023).

An important challenge faced by AI practitioners when performing AI impact assessments is mapping the intended and unintended AI uses (Prunkl et al. 2021). For example, recent research on 32K model cards posted on the HuggingFace platform (Liang et al. 2024) shows that while most cards detail *Training Information*, sections on *Intended Uses* and *Bias, Risks, and Limitations* have lower completion rates (17-23%). Similarly, Yang, Liang, and Zou (2024) found that in Data Cards also hosted on HuggingFace, the section on *Considerations for Using the Data* receives the lowest proportion of content (only 2.1% of the card's text length).

As another means of reflecting on potential positive and negative consequences of AI models, broader societal impacts are introduced as a requirement by leading AI conferences (e.g., the conference on Neural Information Processing Systems (NeurIPS)) (Nanayakkara, Hullman, and Diakopoulos 2021). However, researchers also struggle with filling in such sections due to the inherently difficult nature of the task and high opportunity costs (Prunkl et al. 2021).

Conventional methods to understand the uses and scope of AI technology include systematic and scoping reviews, which are useful for mapping fields of study (Peters et al. 2015). For instance, Moraes, Almeida, and de Pereira (2021) combined literature review with news media research to unveil FRA applications in (semi-)public spaces in Brazil and the associated risks. Similarly, Hupont et al. (2022) reviewed scientific papers and company portfolios to identify 60 facial processing applications, which were then assessed for their risk levels. However, these methods, while insightful, are resource-intensive, demanding both time and expertise (Arksey and O'Malley 2005).

Moreover, even when the uses of AI are known, they can bring unanticipated challenges, from privacy and security issues (Li et al. 2023; Ekambaranathan, Zhao, and Van Kleek 2021) to distorting human beliefs (Kidd and Birhane 2023), excessive dependence that could diminish crucial human skills (Byun, Vasicek, and Seppi 2023; Lu and Yin 2021), and negative environmental impacts (Rillig et al. 2023), as well as impacts on human rights and society (Mantelero 2022). Anticipating such challenges and broader, systemic impacts of technology remains a significant challenge for AI practitioners (Prunkl et al. 2021; Weidinger et al. 2023).

Regulating AI

The pervasiveness of AI, along with the potential risks discussed above, has intensified calls for regulatory oversight (Tahaei et al. 2023; Borenstein and Howard 2021). The first binding regulatory response is the European Commission's AI Act, which aims to balance fostering innovation with protecting rights and societal values. The Act covers a spectrum from low-risk to prohibited AI applications, prohibiting those that can harm individuals or manipulate behaviors, such as social scoring by public authorities. Other regulatory frameworks include the US Office of Science and Technology Policy (OSTP) Blueprint for an AI Bill of Rights, China's Interim Measures for the Management of Generative AI Services, and the UK's pro-innovation approach to AI regulation.

To sum up, the dynamic nature of AI poses a challenge in

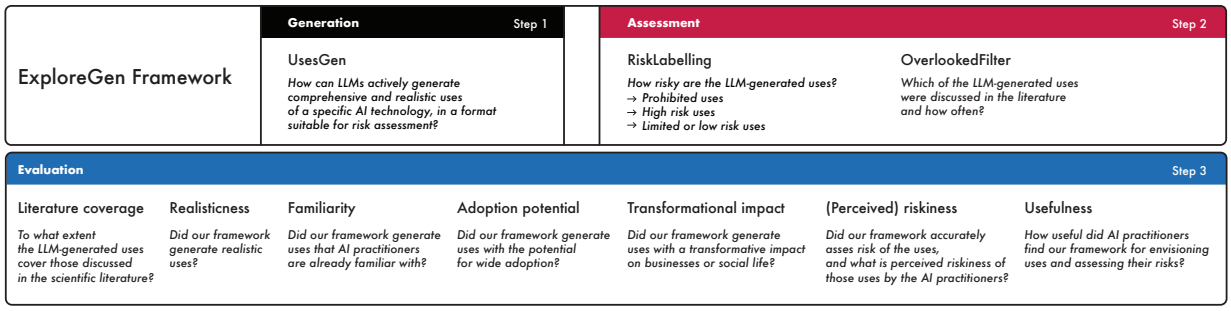

Figure 1: Our methodology consists of three steps. In the first two steps, *ExploreGen* performs *(i) generation (UsesGen)* of various uses for a given AI technology, and their *(ii) assessment (RiskLabeling, Literature Coverage Analyzer)* in terms of the risks based on the EU AI Act, and determining whether they are discussed or overlooked in previous literature. In the last step *(iii)*, we performed the *evaluation* of the generated uses and their risk classification, including the realisticness of the uses, risk assessment accuracy, and usefulness for AI practitioners in envisioning the impacts of AI technology.

its impact assessment, particularly in identifying its myriad uses and ensuring thorough risk assessments. We propose to leverage LLMs to partly tackle these challenges.

Large Language Model Applications

LLMs have already demonstrated their usefulness in a variety of tasks. These range from text annotation (Gilardi, Alizadeh, and Kubli 2023) to assisting with creative and argumentative writing (Lee, Liang, and Yang 2022) and potential for providing help for mental health issues (Sharma et al. 2023). LLMs offer insights that surpass general public knowledge (Gilardi, Alizadeh, and Kubli 2023), show promise in human-AI co-creation processes (Wu, Terry, and Cai 2022; Lee, Liang, and Yang 2022), *brainstorming assistants* (Lukowicz et al. 2023; Bouschery, Blazevic, and Piller 2024), and have the potential to support *interpreting regulatory texts* (Zheng et al. 2023; Cui et al. 2023).

To achieve the desired output from LLMs, it is important to employ best practices in prompt engineering, such as Chain-of-Thought reasoning, using appropriate roles, and providing cues and examples to guide the model's output (Wu, Terry, and Cai 2022; Shieh 2023). However, LLMs also introduce their own AI risks, including biases associated with the training data (Luccioni et al. 2024) and hallucinations (Mittelstadt, Wachter, and Russell 2023), which need to be carefully considered in each application.

Methodology

For our framework's development and assessment, we focused on Facial Recognition and Analysis (FRA), a well-established yet controversial technology due to its known risks (Zhang, Feng, and Sadeh 2021; McClurg 2007), and a contentious topic during the development of the EU AI Act (Hupont et al. 2022).

Designing ExploreGen

We selected GPT-4 due to its top-ranking performance, as shown in leaderboards (LMSYS 2024).

Full text with the Appendix: https://arxiv.org/pdf/2407.12454

Generating Uses (UsesGen). To generate a list of various uses (Figure 1, Framework, Step 1), we specified five elements in UsesGen (Appendix A, Figure 1): system role, instructions, risk concepts, definitions of being realistic, domains, and examples.

The *system role* has been shown to improve the quality of the output, as it allows to generate content from specific perspectives (Giray 2023). We assigned the role of a "*Senior [Technology X] Specialist and Evaluator*" and described its main tasks as *"reviewing, and cataloguing the diverse applications and use cases of [Technology X] across multiple domains, and conducting exhaustive research and analysis"*.

We then followed with the three-part *instruction*: (i) to create a comprehensive and self-explanatory JSON (JavaScript Object Notation) list detailing particular use cases or applications of [Technology X], (ii) to provide precise descriptions for each concept, and (iii) to categorise the LLM-generated uses into 1) *already existent*, 2) *upcoming*, and 3) *unlikely*, along with a one-sentence justification for each use categorization (enacting the *CoT reasoning*).

We asked for each use to be generated along the five concepts proposed by Golpayegani, Pandit, and Lewis (2023):

1. *Domain*: "The area or sector the AI system is intended to be used in" (e.g., education).
2. *Purpose*: "The objective that is intended to be accomplished by using an AI system" (e.g., attendance tracking).
3. *Capability*: "The capability of the AI system that enables the realisation of its purpose and reflects the technological capability" (e.g., identify students' faces and match them with database).
4. *AI user*: "The entity or individual in charge of deploying and managing the AI system, including individuals, organisations, corporations, public authorities, and agencies responsible for its operation and management" (e.g., schools).
5. *AI subject*: "The individual directly affected by the use of the AI system, experiencing its effects and consequences.

They interact with or are impacted by the AI system's processes, decisions, or outcomes" (e.g., students).

To aid the realisticness categorisation, we also provided the *definitions* of the three categories of being realistic. Already existent uses were defined as currently implemented and well-established uses. Upcoming uses were defined as being under current development, being researched, or subject to discussions without being implemented or being severely limited in practice due to various reasons. Lastly, unlikely uses, introduced to capture hallucinations, lack value, usability, applicability, or practicality, or are deemed unnecessary, impossible, incoherent, or unrealistic.

To further guide UsesGen we requested the AI technology uses across a broad set of *domains*. Without such a request, the uses generated by the LLM would encompass the most common and well-known FRA uses, since LLMs suffer from exposure bias (Wu, Terry, and Cai 2022). The domains served as a *cue* in our prompt. Our procedure for listing a broad set of domains was as follows. First, domains were derived from the EU AI Act's Annex III (e.g., "Education and vocational training"), along with 32 domains that were not explicitly listed but were mentioned in the EU AI Act text or its Amendments (e.g., "Social Media" from Amendment 51 stating: *"The indiscriminate and untargeted scraping of biometric data from social media [...] add to the feeling of mass surveillance [...]"*). Moreover, we derived additional domains from a focus group using a think aloud protocol (N=8) to ensure capturing all significant domains beyond the EU AI Act. The session was with our research group (3F, 5M, mean age: 31.8, SD: 6.74, range: 22-45). We used a Miro board and asked the participants to think of domains that affect their lives along the five levels of the Social-Ecological Model (Golden et al. 2015): individual, interpersonal, institutional, community, and public policy. This resulted in an additional 6 domains that were not yet covered by the previous list of 40 domains, resulting in the final 46 diverse domains (Appendix B, Table 1).

To complete the prompt, we carefully crafted five *examples* (employing *few-shot* learning (Brown et al. 2020)) striking a balance between providing a diverse range of examples and keeping the prompt at a manageable context length (Liu et al. 2024). The output also requires the label for the realisticness of the use. For example, "FRA for medical diagnosis" was categorised as an *upcoming* use, along with the justification saying that it *has the potential to revolutionise healthcare, yet successful integration depends on resolving privacy, regulatory, and trust-related issues.* We placed the examples section at the end of the prompt, as examples can not only illustrate the desired input-output relationships, but also aid the model's context comprehension and response expectations (Brown et al. 2020).

Assessing the Risk of Generated Uses (RiskLabelling).

To enable risk assessment as per the EU AI Act (Figure 1, Framework, Step 2), we specified five elements (Appendix D, Figure 2): system role, instructions, legal documents, placeholder for a list of uses, and output structure.

We started the prompt by selecting the *system role* of an *"Experienced Judge who works in the field of AI technology regulation"*, and described the role further: *"You are thoughtful, decisive, experienced and conscientious. You have access to the entirety of the EU AI Act"*.

Use ID. Use Description
27. Validate remote worker identity online.
52. Recognize customers, tailor services.
68. Identify watchlisted individuals at borders.
69. Verify asylum seeker identities.
70. Prevent voter fraud via identity verification.
80. Authenticate energy facility personnel access.
83. Verify military personnel identities.
84. Identify threats in crowds by military.
88. Identify citizens for personalized services.
91. Secure embassies by identifying visitors.
98. Authenticate emergency responders' identities.
104. Verify cargo access by authorized personnel.
108. Control access to restricted urban areas.
114. Verify access to protected environmental areas.
118. Identify illegal loggers.
120. Verify access to climate-sensitive areas.

Table 1: LLM-generated uses overlooked by the research literature. For full details, see Appendix F, Table 3.

We then provided the *instructions* to classify the uses by utilising the *CoT reasoning* by requesting to first expand the concise FRA use into a description of a hypothetical AI system that employs it. We then followed with the requests to consider the EU AI Act and its amendments (European Comission 2024) provided in *input*, and to classify the system as "prohibited", or "high risk", or, otherwise, as "limited or low risk".

The prompt was then provided with the *placeholder* for AI technology uses for which the risk assessment should be performed.

Finally, we requested the *output structure* of the risk classification to encompass:

1. *Description*: Provides a clear understanding of the intended use of the AI system.
2. *Classification*: Outcome of the classification which can be either prohibited, high risk, or limited or low risk.
3. *Relevant Text from the Act*: If applicable, a quote from the EU AI Act is included, along with a relevant amendment or section to provide legal context.
4. *Reasoning*: Explanation that rationalises the specific risk classification of the inputted AI use.

Assessing the Literature Coverage of Generated Uses (Literature Coverage Analyzer). To assess which of the LLM-generated uses were discussed in the literature (Figure 1, Framework, Step 2), and possibly uncover overlooked ones by the literature, we collated all the 200M papers from Semantic Scholar's May 2023 dump. We then filtered the papers to those being written in English, and having both the title and abstract fields available, resulting in 3M papers.

Next, we embedded the `title + abstract` field for each of the articles, as well as the description of each of

api.semanticscholar.org/api-docs/datasets

the LLM-generated use using *all-mpnet-base-v2* sentence-transformers (Reimers and Gurevych 2019) model. This model is trained using a self-supervised contrastive learning, by fine tuning the pretrained *microsoft/mpnet-base* model on above 1 billion sentences. Upon pairing each use with the paper with the maximum similarity of their embeddings, we then manually explored which similarity threshold will yield use-paper pairs such that the paper's abstract indeed discusses the use. We explored $\{95^{th}, 99^{th}, 99.5^{th}, 99.9^{th}\}$ percentile thresholds, until we concluded that the 99.9^{th} percentile one yielded $3,295$ papers, which indeed discussed paired FRA uses.

The top frequent venues in which these papers are published include: arXiv.org, International Journal for Research in Applied Science and Engineering Technology, IEEE International Conference on Systems, Man and Cybernetics, ACM Multimedia, Interspeech, PLoS ONE, IEEE/ACM International Conference on Human-Robot Interaction, and Computer. The most commonly discussed uses are: *secure access control, use #1* discussed by 291 articles, *detecting driver fatigue through facial analysis, use #134* discussed by 251, and *using diverse facial data to refine algorithms, use #60*, discussed by 189 articles.

Evaluating ExploreGen

This section outlines the process of evaluating our ExploreGen framework (Figure 1, Framework, Step 3). The goal of our framework was to generate realistic uses of a given AI technology, such that AI practitioners are not familiar with all of them, and to accurately classify their risks based on the regulation. Moreover, the generated uses should exhibit potential for adoption and transformational impact.

To ascertain the effectiveness of the framework at meeting this goal, our evaluation ought to answer seven questions:

I. *Literature coverage.* To what extent the generated uses cover those discussed in the scientific literature?

II. *Realisticness.* Did our framework generate realistic uses?

III. *Familiarity.* Did our framework generate uses AI practitioners are familiar with?

IV. *Adoption potential.* Did our framework generate uses that have a potential for adoption?

V. *Transformational impact.* Did our framework generate uses that have a transformation impact?

VI. *(Perceived) riskiness.* Did our framework accurately asses risk of the uses, and what is perceived riskiness of those uses by the AI practitioners?

VII. *Usefulness.* How useful did the AI practitioners find our framework in assisting with their tasks of envisioning AI uses and assessing associated risks?

Metrics. We then defined six quantitative and one qualitative metric to answer these questions.

The first metric assessed the *coverage* of the generated uses in relation to those discussed in the literature. It was measured as the percentage of matches with the ground truth (GT), which we derived from a scoping review of FRA use cases (Appendix C). Two authors independently conducted a manual assessment, categorizing each generated use as either matching or not matching the ground truth list.

The second metric assessed the *realisticness* of the generated uses. We measured it by calculating the agreement between the realism labels assigned by the LLM and those given by the participants in the user study.

The third metric assessed participants' *familiarity* with the generated uses. It was measured through a question: *"How frequently do you encounter references to this use in your professional life?"* evaluated on a 7-point Likert scale from 'rarely' to 'always'.

The fourth metric assessed practitioners' perceptions about the real-life *adoption potential* of the LLM-generated uses. It was measured through a question: *"How likely it is that this use will be widely adopted in the near future?"* evaluated on a 7-point Likert scale from 'very unlikely' to 'very likely'.

The fifth metric assessed AI practitioners' perceptions of the potential *transformational impact* of the LLM-generated use cases. It was measured by asking, *"How likely is it that this use will fundamentally change the way businesses operate or people live?"*. Participants rated this on a 7-point Likert scale from 'very unlikely' to 'very likely'.

The sixth metric assessed AI practitioners' perceptions of the *riskiness* of the use cases in terms of their potential societal and environmental adverse impacts. It was measured by asking both AI developers and compliance experts to answer how risky do they consider the use *"for society as a whole"* as well as *"for the environment"*. These two questions were rated on a 7-point Likert scale from 'not risky at all' to 'unacceptably risky'. Additionally, to validate *RiskLabelling*'s classification outputs, we provided the compliance experts with both the classification and the LLM's justification and measured their agreement. If they disagreed with the classification, they could select the correct classification (including the option of 'insufficient information to assess the use'). If they disagreed with the justification, they could provide their own reasoning.

The last, seventh metric was about the *usefulness* of our framework, captured through three open-ended questions: *"How useful is this framework for envisioning uses of technology?"*, *"How useful is this framework for understanding the risks and benefits of each use?"*, and *"At what stage in your assessment process would you use this framework?"* .

Setup. To derive the first metric *(literature coverage)*, we performed a scoping review. To derive the remaining six metrics *(realisticness, familiarity, adoption potential, transformational impact, perceived riskiness, usefulness)*, we conducted nine user studies with 25 AI practitioners in total (12 AI developers, and 13 AI compliance experts).

Scoping Review. To obtain a list of a FRA uses discussed in the literature, we performed the scoping review in accordance with the 5-stage guidelines (Arksey and O'Malley 2005):

1. *Identifying research question.* "What are the researched

huggingface.co/sentence-transformers/all-mpnet-base-v2

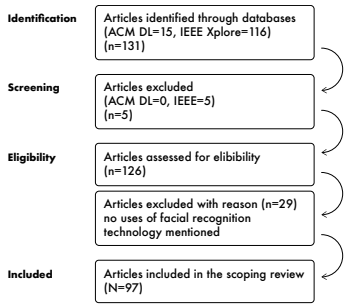

Figure 2: *The scoping review:* identification, screening, and assessment for eligibility of articles. Starting with 131 initial papers identified, a total of 97 were included. From these papers, 75 unique FRA uses were identified (Appendix C).

or proposed uses of FRA as found in the literature?"

2. *Identifying relevant articles.* In consultancy with the research team, we selected the ACM Digital Library (https://dlnext.acm.org) and IEEE Xplore (http://ieeexplore.ieee.org) as our databases, which correspond to the main Computer Science and Engineering digital libraries, likely to cover a broad spectrum of research on FRA technology. We used the following Query (Q) = [[Title: "face recognition"] OR [Title: "facial recognition"]] AND [Full Text: "use case*"] (Figure 2, *Identification*).
3. *Selecting articles.* We included peer-reviewed articles as well as larger scholarly works, such as extended abstracts (e.g., posters and demos) and workshop papers. All selected works are referred to as *articles*. For all identified articles, we applied the following inclusion criteria: (1) written in English and (2) describing, studying, or envisioning at least one use of facial recognition technology. In the initial search, 131 articles were identified from the ACM and IEEE databases. As no duplicates were found, all 131 articles were screened based on titles and abstracts. Five articles were removed as they did not discuss an FRA use (Figure 2, *Screening*). Subsequently, 126 articles were assessed for eligibility based on their full text (Figure 2, *Eligibility*), resulting in a final selection of 97 relevant articles (Figure 2, *Included*). The lead author performed the article selection process.
4. *Charting the data.* The lead author began reading the articles and simultaneously developed a coding system for the FRA uses described, studied, and envisioned in the articles. As the lead author read the articles, they color-coded the FRA uses and extracted them. Each time a new FRA use was identified, it was added to the coding system. Any ambiguities—though rare due to the straightforward nature of the FRA uses mentioned—were discussed and resolved between the lead and second author.
5. *Collating, summarising, and reporting results.* The scoping review resulted in 97 articles from which we identified 75 unique uses of FRA, listed in Appendix C.

User Studies with AI Practitioners. We conducted seven in-person studies involving 3 AI developers (30 minutes each) and 4 AI compliance experts (45 minutes each), complemented by two online studies on Prolific: one with 9 AI developers, and another one with 9 AI compliance experts.

The in-person studies consisted of four steps. First, we asked participants about their current practices and challenges in envisioning AI technology uses and their associated risks. Second, we presented an interactive list of 138 uses and tasked them with selecting one project that balances being interesting to develop and adhering to the company's code of conduct (Appendix G, Figure 3A), followed by questions on the usefulness of this list for envisioning technology uses and understanding the risks and benefits. Third, we presented 16 interactive assessment cards for overlooked uses and tasked them with annotating the uses for realisticness, familiarity, adoption potential, transformational impact, and perceived riskiness (Appendix G, Figure 3B). AI compliance experts also evaluated the *RiskLabelling* classification and justification, making corrections if necessary (Appendix G, Figure 3C). This allowed us to compare perceived use riskiness between developers and compliance experts. Finally, we asked participants about the framework's usefulness for envisioning technology uses, understanding risks and benefits, and identifying the stage in their assessment process where they would use this framework. Each of the 16 uses was annotated by 7 different AI practitioners: 3 AI developers and 4 AI compliance experts.

The online studies used a custom web-based survey consisting of five pages. The first page outlined the study's description and tasks for crowdworkers: read the definitions of 'risky' uses and annotate each use for realism, familiarity, adoption potential, transformational impact, and perceived riskiness. AI compliance experts were also asked to agree or disagree with the *RiskLabelling* classification and justification, and make corrections if necessary. The second page provided definitions of risky uses according to the EU AI Act. The third and fourth pages presented assessment cards for 46 uses (23 per page) with input boxes for annotations (Appendix G, Figures 3A,B). The final page included a confirmation note and redirected participants to Prolific. Each of the 138 uses was annotated by 6 different AI practitioners: 3 AI developers and 3 AI compliance experts.

To ensure response quality, we conducted two attention checks during the studies and implemented two deliberate survey design features. First, after reading task instructions, participants encountered one of the two attention-check sentences: *"When asked for your favorite color/city, you must select "Blue/Rome"*. We also included one prohibited use labelled as "low risk" with a false justification mimicking text from the EU AI Act. Participants had to correctly respond to two out of these three checks. Second, we disabled pasting from external sources and editing previous responses to ensure original and thoughtful answers.

Participants. For our studies, we recruited participants and surveyed them across two cohorts: *a)* AI developers and *b)* compliance experts.

For the in-person studies, we recruited participants through an internal mailing list at a large tech company,

and our professional networks. We asked for individuals currently developing AI systems using machine learning, computer vision, and image recognition. To recruit AI compliance experts, we sought individuals familiar with the EU AI Act, experienced in reviewing AI use cases, and involved in at least one ongoing AI impact assessment project.

For the online studies, we recruited participants from Prolific, controlling for their roles in the organization, the frequency of AI use in their jobs, fluency in English, and geographic location. To recruit AI developers, we selected participants who likely contribute to developing AI systems as part of their software engineering roles, using AI daily. To recruit compliance experts, we looked for participants likely involved in revising AI systems as part of their legal roles, using AI at least 2-6 times a week. We limited our participant pool to individuals residing in the European Union. All Prolific participants were paid an average of $12 USD/hour.

Analysis. We performed both quantitative and qualitative analyses. For the quantitative analysis, we measured the frequencies across six metrics: coverage, realisticness, familiarity, adoption potential, transformational impact, and perceived riskiness. For the qualitative analysis, we thematically analyzed responses to open-ended questions (Saldaña 2015; Miles and Huberman 1994; McDonald, Schoenebeck, and Forte 2019; Braun and Clarke 2006) to understand factors influencing the framework's usefulness for envisioning technology uses, assessing risks and benefits, and determining the appropriate assessment stage for its application.

Evaluation Results

UsesGen, using FRA technology as input, generated 138 uses listed in Appendix F, Table 3. According to its own realisticness label, 8 (6%) of the uses were deemed unlikely (e.g., *FRA to track the carbon footprint of individuals, use #119*, as it is unlikely to be adopted, and *detecting plant diseases and pest infestations, use #50*, as it does not employ the capabilities of FRA).

RiskLabelling classified 10 (7%) uses as prohibited, 66 (48%) as high risk, and 62 (45%) uses as limited or low risk. Example *RiskLabelling* outputs for one use per each class are shown in Appendix E, Table 2.

Literature Coverage Analyzer identified 16 out of the 138 LLM-generated uses that were not discussed in any of the 3M Semantic Scholar papers we analyzed. These uses are shown in Table 1. This indicates that while these uses are likely mentioned in news, press, or social media (and thus included in the LLM training data), they have not yet been the focus of in-depth scientific research.

I. Literature coverage. The uses were expressed differently between the GT list (Appendix C) and the LLM-generated list (Appendix F, Table 3). In the GT list, they are written as single sentences mainly describing the purpose, whereas in the LLM-generated list, they always follow a structured format based on the 5 risk concepts (e.g., AI domain, AI user). Therefore, we employed a relaxed matching approach, allowing us to count two uses with different levels of generality as a match (e.g., *detect fatigue in individuals, GT-use #69* was matched with *improving driver safety by detecting driver fatigue through facial analysis, use #134*).

The LLM-generated list covered 96% of the literature-derived ones with the only 3 GT uses not found in the LLM-generated list being: *provide real-time information about visitors in high-profile buildings, GT-use #5*, *help people recognise faces by using smart glasses to display names and social network activities of identified people, GT-use #72*, and *facilitate tourists in meeting new people, GT-use #74*.

Given the relaxed approach we applied, the high matching rate between the two lists reflects the LLM-generated list's scoping coverage of various uses discussed in the literature rather than comprehensively covering all possible uses. Given the many contexts for each use (e.g., various subjects, domains, or locations), comprehensive coverage is practically unattainable.

II. Realisticness. After excluding the 8 uses labeled by the LLM itself as unrealistic, the majority agreement across the participants in different user studies was that the remaining 130 uses were all realistic. Of these, 91 uses (70%) were labeled as already existing, and 39 (30%) as upcoming (e.g., *recognizing signs of distress or confusion for elderly care assistance, use #6* and *facilitating non-verbal communication by interpreting facial expressions and gestures for non-verbal individuals, use #77*).

The analysis of unrealistic uses revealed that some domains were more prone to hallucination, such as "Agriculture and Farming" or "Environment and Sustainability." Given that FRA has fewer applications in these domains, asking the LLM to generate uses in these areas led to hallucinations. These domains were included because they are mentioned in the EU AI Act and hold potential significance for other AI technologies (e.g., Earth Observation), where they might not lead to hallucinated uses.

III. Familiarity. As shown in Figure 3, both AI developers and compliance experts demonstrated low familiarity with the uses produced by *UsesGen*. Over 50% (48%) of these uses were reported by developers (compliance experts) as rarely encountered in their professional lives. For the overlooked uses, developers reported rarely encountering 60% of these, while compliance experts reported rarely encountering even 75%. The chi-squared test results confirmed that the distributions of familiarity scores significantly differ between all uses and overlooked uses, validating the ability of our *Literature Coverage Analyzer* to identify less well-known and understudied uses. The distribution of familiarity scores did not differ statistically significantly between the cohorts of AI developers and compliance experts.

IV. Adoption potential. AI developers thought that most of the uses are 'somewhat likely' ($\sim$27% of the uses) or 'very likely' ($\sim$25%) to be adopted, though the ratio of the 'very likely' ones was smaller for the overlooked uses ($<$15%). Compliance experts were, interestingly, scoring most of the uses, including the overlooked ones, as 'very likely' ($>$35% of uses) to be adopted. In this case, a chi-squared test results confirmed that the distributions of scores for adoption potential significantly differed between the two cohorts, with compliance experts generally giving higher scores.

V. Transformational impact. Developers were slightly

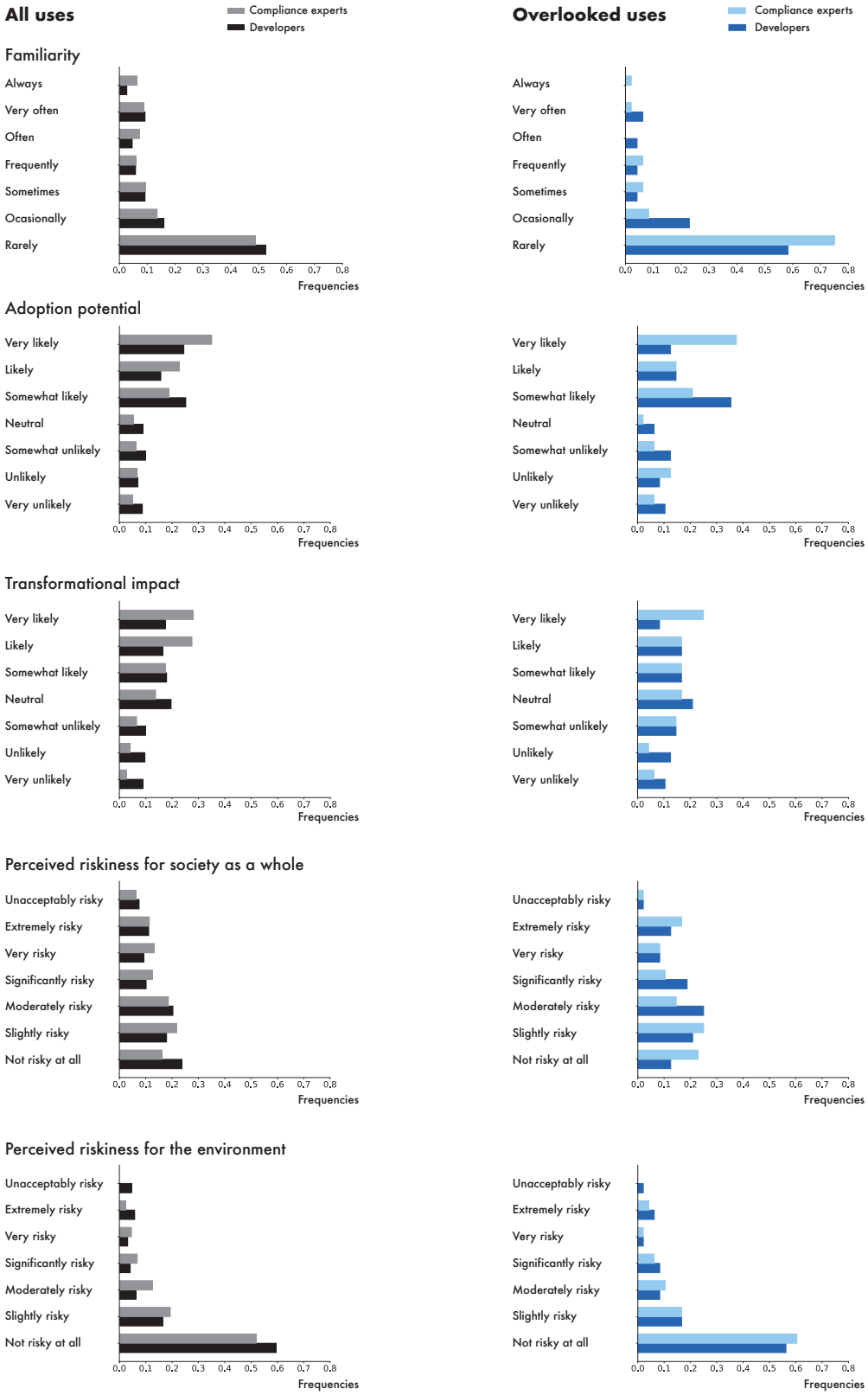

Figure 3: Evaluation results for the five quantitative metrics: familiarity with the use, its adoption potential, transformational impact, and perceived riskiness for society as a whole and for the environment.

more conservative in estimating the potential for transformative impact of the uses (Figure 3), assigning the largest proportion of uses a 'neutral' score ($\sim$20%). In contrast, compliance experts gave the highest proportion of 'very likely' scores (>25% of the uses) for both all and overlooked ones. Similarly as for the adoption potential scores, a chi-squared test results confirmed that the distributions of scores for transformational potential significantly differed between the two cohorts.

VI. (Perceived) riskiness. Each use was rated by three compliance experts. To obtain the ground truth label, we required that at least 2 of the 3 labels were aligned. By comparing these ground truth labels with the *RiskLabelling* labels, we found that 94.5% of the uses were correctly classified, with an almost perfect Cohen's Kappa agreement of 92.2%. However, the inter-rater agreement among the three annotators was only moderate, with a Fleiss' Kappa score of 49.1%, suggesting the task is challenging and that participants might have defaulted to the provided labels.

For example, participants disagreed with the LLM's limited or low-risk classification for uses such as *verifying the identity of customers during transactions by banks, use #19*, and *identifying obstacles and people to avoid collisions by robots, use #56*. For *use #19*, they commented that it should be high risk due to the *"high chance for fraud"* and the possibility that the *"AI system could see the PIN of the bank card!"*. For *use #56*, two annotators voted for a high-risk label because *"in the case of misuse or malfunctioning, the AI could lead to serious harm for individuals"* and *"[...] put human lives at risk."*

On the other hand, the participants did not agree with the high-risk classification for *assisting law enforcement agencies in criminal investigations by identifying suspects in video footage, use #85*. Two of them thought this use should be classified as *prohibited* in the EU, as it could lead to violations of privacy rights. The LLM did not classify it as such because the identification from footage is not in real-time, which is a requirement for prohibited uses specified in Article 5(1)(d). The third annotator, however, suggested downgrading the risk classification to limited or low risk because the use is *"necessary to provide proof and existence of criminal activities and facilitate law enforcement work"*. These examples demonstrate the subtleties in the risk assessment task, including the interpretation of the use context and the annotators' personal viewpoints (Hupont et al. 2022), which partly explain the lower inter-rater agreement among our participants.

As shown in Figure 3, developers thought that most of the uses are only 'slightly' to 'moderately risky' for society (approximately 20-25%), and not at all risky for the environment (approximately 50-60%). This contrasts with our risk classification finding that over 50% of the uses are either high-risk or prohibited according to the EU AI Act. These results highlight the challenge developers face in identifying and classifying the riskiness of AI uses.

VII. Usefulness. Finally, we studied the extent to which the two cohorts of AI practitioners found our framework useful in assisting with tasks such as envisioning AI uses and assessing associated risks.

AI compliance experts found *ExploreGen*'s output particularly useful. For example, L01 mentioned that a tool *"classifying [uses] in different ways and offering various uses of those [technologies], would be very useful in my job, [...] because it would help me look at things in a different way."* L03 stated, *"I enjoyed it [...] I think it's really helpful to kind of envision what will be the future use of AI and then think about how it will impact society and the environment. I think it's a good exercise for someone working in the tech space in general,"* and *"... it will also be useful for people who want to understand the technology, like people impacted by the technology and the public."* One participant from a major tech company developing FRA technologies expressed excitement upon discovering uses they are currently working on, particularly in risk and compliance assessment. They also found inspiration for new potential use cases, stating, *"We are putting more effort into going into the [domain X], and that could be a good use."* L04 was particularly engaged with the risk-classification output provided by our tool. For instance, they focused on the use *identifying personnel by logistics companies to improve the efficiency of cargo handling, use #104,* and agreed with the low-risk classification. They noted that *"[A major company] has just gotten a judgment in its favor that very far-reaching analytics in its plants in [country Y] are permissible."* L03 was also inspired to think about the risks of the presented uses. They deliberated about the use *verifying patient identity in medical settings, use #10*, which is classified as low risk, but they thought it could incur many risks as *"services like this [...] can be exclusionary to certain, especially marginalized communities."* They concluded, *"I would look into developing this, but I would consider this a high-risk use depending on the context and on the decision that's being made by verifying."*

AI developers, on the contrary, initially struggled to identify the application of our tool in their everyday work. While interested in exploring the presented uses, they frequently asked for more details and insights on specific uses. For instance, D02 expressed feeling overwhelmed by the comprehensive list of uses: *"I imagine [I am] developing that, and put a lot of cognitive load in each case and then imagining how it will work and how it will be developed."* During the interviews, it became clear that developers, especially those working on business products, have less opportunity to use a tool like *ExploreGen* because they typically do not engage in extensive brainstorming and reflexivity. Instead, they usually receive well-defined uses to develop. For example, D03 commented, *"I've been working with products and generally you start with a use that you want to develop [...] and then you work backwards and maybe a technology is not useful for that particular problem."* D03 also stated, *"For most of the people I speak with, it seems like more of an afterthought than like an active design. [You think] what could be the risks kind of post hoc?"* They added, *"But I think people are generally getting a little bit better at that now because I think people are seeing that AI is progressing quite fast..."* For these reasons, developers appreciated the color-coding of the use risk levels, as it provided a quick overview of the more or less risky domains, contexts, and uses. One

participant noted surprise at seeing a similar use having different risk levels in two domains, finding the tool helpful for educating them about the EU AI Act and its domain-based risk classification. D01, who holds the most senior role among the developer participants, stated: *"[We] have a brainstorming session, first of all understanding if AI is really needed to solve the problem or not[...]"* They added about our tool: *"It will be very helpful for me or someone in my team to get a first sense of the risks involved..."* Generally, developers preferred the second task in the study, where they could focus on a subset of uses and scrutinize them in detail, as this aligns more closely with their job responsibilities. Additionally, those in senior roles and closer to R&D found our tool more useful for brainstorming and deliberation tasks compared to junior developers and those working in business production.

Both AI developers and AI compliance experts agreed that a tool like ours would be most useful during the *design* stage of AI development. Moreover, several participants indicated they would use it throughout *all* stages, as noted by L4: *"I don't think one stage is more important than the other. I think there are different risks at different stages."*

Discussion

Findings from nine user studies have demonstrated the potential of our proposed LLM framework, *ExploreGen*, to enhance reflectivity, ideation, and deliberation for both AI developers and compliance experts. These tasks, which are increasingly essential yet often challenging, can greatly benefit from our framework (Liang et al. 2024; Prunkl et al. 2021). Given the collaborative nature of our framework, we envision that human-in-the-loop validation (i.e., Step 3) will always be integral to its application, allowing stakeholders to double-check both the LLM's outputs and each other's ideas. Our tool contributes to the growing body of research advocating for (Sherman and Eisenberg 2024) and exploring (Buçinca et al. 2023; Wang et al. 2024) the use of LLMs to support responsible AI design.

Implications

Brainstorming in AI Developer Teams. *ExploreGen* successfully generated realistic uses that practitioners were not very familiar with, many of which were rated as having high adoption potential and transformational impact. Developers found the overview of uses contextualized across various domains, along with their risk levels, to be informative. Some saw the tool's value during brainstorming meetings while deliberating on which directions for technology applications to pursue. Additionally, they expressed interest in a tool with a more in-depth analysis of specific uses, allowing to break down the associated risks of the use they are developing and be informed about similar risks faced by different uses.

Bridging Risk Perception with Compliance. Compliance experts agreed with the risk classifications provided by *RiskLabelling*, though they noted that subtle changes in the context of use might alter the classification level. Despite more than 50% of the FRA technology uses being classified as high risk or prohibited, practitioners perceived them as mostly only slightly risky for society and not at all for the environment. However, due to the size of the datasets and computational demands, energy consumption is becoming an important consideration for FRA technology (Hassel and Özkiziltan 2023), highlighting a disconnect in AI practitioners' understanding of all the technology's impacts.

Data-driven Deliberation for Compliance Experts. Compliance experts saw more direct applications of *ExploreGen* in its current form for their work, as they often explore various (often unintended or unexpected) contexts of use for a given technology. They found the tool very helpful for this task. They also appreciated the breakdown of uses across various domains and risk levels and wanted features allowing for additional breakdowns (e.g., according to the subjects or types of risk).

Limitations and Future Work

LLM Method Shortcomings. The use of LLMs presents four main challenges. First, the generated uses, and risks may be limited to the training set and biased (Luccioni et al. 2024), potentially overlooking important aspects. Enhancements could include fine-tuning (Hu et al. 2023) or augmenting with specialized datasets (e.g., from AI Incident Database (McGregor 2021)). Second, there is a risk of incorrect outputs due to LLM hallucinations (Mittelstadt, Wachter, and Russell 2023). *UsesGen* identified 6% unrealistic uses, which were removed. Future research could explore combining classifiers and manual checks to ensure accuracy (Mittelstadt, Wachter, and Russell 2023). Third, LLMs may be overly conservative, missing risky edge-case uses due to built-in guardrails. Last, presenting LLM outputs to users could create a false sense of security (Pataranutaporn et al. 2023). Ongoing research in human-AI interaction offers strategies to mitigate these issues, such as designing cognitive forcing functions (Buçinca, Malaya, and Gajos 2021) and skill improvement (Buçinca et al. 2024).

Difficulty of Risk Classification. We focused on labeling prohibited and high-risk uses, with the remainder classified as limited or low risk. However, the EU AI Act includes an additional classification label, transparency risk, which we omitted due to the task's inherent complexity arising from ambiguities in the Act's wording (Veale and Zuiderveen Borgesius 2021). These ambiguities, resulting from the interplay between technical and legal jargon, pose challenges even for professionals in the field, as reflected in the moderate inter-rater agreement among our user study participants. Additionally, while the five risk categories aid in classification, practical variations in each use ultimately determine their final classification.

Evaluation Limitations and Generalizability. Our GT list of FRA uses might not be comprehensive, as the scoping review was primarily conducted by a single author using a conservative paper selection method. Future research could include more extensive literature reviews. While we evaluated our framework with 25 AI practitioners focusing on FRA technology, future studies should test its applicability to other technologies and involve a broader group of AI practitioners, researchers, and the general public.

References

Arksey, H.; and O'Malley, L. 2005. Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, 8: 19–32.

Ashurst, C.; Hine, E.; Sedille, P.; and Carlier, A. 2022. Ai ethics statements: analysis and lessons learnt from neurips broader impact statements. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2047–2056.

Borenstein, J.; and Howard, A. 2021. Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, 1: 61–65.

Bouschery, S. G.; Blazevic, V.; and Piller, F. T. 2024. Artificial Intelligence-Augmented Brainstorming: How Humans and AI Beat Humans Alone. *Available at SSRN 4724068*.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.

Buçinca, Z.; Pham, C. M.; Jakesch, M.; Ribeiro, M. T.; Olteanu, A.; and Amershi, S. 2023. Aha!: Facilitating ai impact assessment by generating examples of harms. *arXiv preprint arXiv:2306.03280*.

Buçinca, Z.; Swaroop, S.; Paluch, A. E.; Murphy, S. A.; and Gajos, K. Z. 2024. Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning. *arXiv preprint arXiv:2403.05911*.

Byun, C.; Vasicek, P.; and Seppi, K. 2023. Dispensing with Humans in Human-Computer Interaction Research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–26.

Calvo, R. A.; Peters, D.; and Cave, S. 2020. Advancing impact assessment for intelligent systems. *Nature Machine Intelligence*, 2(2): 89–91.

Cui, J.; Li, Z.; Yan, Y.; Chen, B.; and Yuan, L. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Davenport, T.; and Kalakota, R. 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2): 94.

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

Dowling, M.; and Lucey, B. 2023. ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, 53: 103662.

Ekambaranathan, A.; Zhao, J.; and Van Kleek, M. 2021. "Money Makes the World Go Around": Identifying Barriers to Better Privacy in Children's Apps From Developers' Perspectives. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

European Comission. 2024. Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences 120 (30)*.

Giray, L. 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering*, 1–5.

Golden, S. D.; McLeroy, K. R.; Green, L. W.; Earp, J. A. L.; and Lieberman, L. D. 2015. Upending the Social Ecological Model to Guide Health Promotion Efforts Toward Policy and Environmental Change. *Health Education & Behavior*, 42(1_suppl): 8S–14S. PMID: 25829123.

Golpayegani, D.; Pandit, H. J.; and Lewis, D. 2023. To Be High-Risk, or Not To Be–Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 905–915. New York, NY, USA: ACM.

Goodell, J. W.; Kumar, S.; Lim, W. M.; and Pattnaik, D. 2021. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32: 100577.

Hassel, A.; and Özkiziltan, D. 2023. Governing the work-related risks of AI: implications for the German government and trade unions. *Transfer: European Review of Labour and Research*, 29(1): 71–86.

Hu, Z.; Lan, Y.; Wang, L.; Xu, W.; Lim, E.-P.; Lee, R. K.-W.; Bing, L.; and Poria, S. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Hupont, I.; Tolan, S.; Gunes, H.; and Gómez, E. 2022. The landscape of facial processing applications in the context of the European AI Act and the development of trustworthy systems. *Scientific Reports*, 12(1): 10688.

Kidd, C.; and Birhane, A. 2023. How AI can distort human beliefs. *Science*, 380(6651): 1222–1223.

Lee, M.; Liang, P.; and Yang, Q. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022*

CHI Conference on Human Factors in Computing Systems, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9): 1–46.

Liang, W.; Rajani, N.; Yang, X.; Ozoani, E.; Wu, E.; Chen, Y.; Smith, D. S.; and Zou, J. 2024. What's documented in AI? Systematic Analysis of 32K AI Model Cards. *arXiv preprint arXiv:2402.05160*.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

LMSYS. 2024. LMSYS Chatbot Arena Leaderboard.

Lu, Z.; and Yin, M. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36.

Lukowicz, P.; et al. 2023. Interacting with Large Language Models: A Case Study on AI-Aided Brainstorming for Guesstimation Problems. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, volume 368, 153. IOS Press.

Mantelero, A. 2022. *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*. T.M.C. Asser Press. ISBN 9789462655317.

McClurg, A. J. 2007. In the face of danger: Facial recognition and the limits of privacy law. *Harvard Law Review*, 120(7): 1870–1891.

McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

McGregor, S. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15458–15463.

Miles, M.; and Huberman, M. 1994. *Qualitative Data Analysis: A Methods Sourcebook*. Sage.

Mittelstadt, B.; Wachter, S.; and Russell, C. 2023. To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour*, 7(11): 1830–1832.

Moraes, T. G.; Almeida, E. C.; and de Pereira, J. R. L. 2021. Smile, you are being identified! Risks and measures for the use of facial recognition in (semi-) public spaces. *AI and Ethics*, 1(2): 159–172.

Nanayakkara, P.; Hullman, J.; and Diakopoulos, N. 2021. Unpacking the expressed consequences of AI research in broader impact statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 795–806.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Owen, R.; and Pansera, M. 2019. *Responsible innovation and responsible research and innovation*. Edward Elgar Publishing.

Pataranutaporn, P.; Liu, R.; Finn, E.; and Maes, P. 2023. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10): 1076–1086.

Peters, M.; Godfrey, C.; Khalil, H.; Mcinerney, P.; Parker, D.; and Soares, C. 2015. Guidance for conducting systematic scoping reviews. *International Journal of Evidence-Based Healthcare*, 13(3).

Phaal, R.; Farrukh, C. J.; and Probert, D. R. 2004. Technology roadmapping—A planning framework for evolution and revolution. *Technological forecasting and social change*, 71(1-2): 5–26.

Prunkl, C. E.; Ashurst, C.; Anderljung, M.; Webb, H.; Leike, J.; and Dafoe, A. 2021. Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence*, 3(2): 104–110.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Rillig, M. C.; Ågerstrand, M.; Bi, M.; Gould, K. A.; and Sauerland, U. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9): 3464–3466.

Saldaña, J. 2015. *The Coding Manual for Qualitative Researchers*. Sage.

Sharma, A.; Rushton, K.; Lin, I.; Wadden, D.; Lucas, K.; Miner, A.; Nguyen, T.; and Althoff, T. 2023. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. In *ACL: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Sherman, E.; and Eisenberg, I. 2024. AI Risk Profiles: A Standards Proposal for Pre-deployment AI Risk Disclosures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23047–23052.

Shieh, J. 2023. Best practices for prompt engineering with OpenAI API.

Smuha, N. A. 2021. From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1): 57–84.

Sraml Gonzalez, J.; and Gulbrandsen, M. 2022. Innovation in established industries undergoing digital transformation: the role of collective identity and public values. *Innovation*, 24(1): 201–230.

Stahl, B. C.; Antoniou, J.; Bhalla, N.; Brooks, L.; Jansen, P.; Lindqvist, B.; Kirichenko, A.; Marchal, S.; Rodrigues, R.; Santiago, N.; et al. 2023. A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, 56(11): 12799–12831.

Tahaei, M.; Constantinides, M.; Quercia, D.; Kennedy, S.; Muller, M.; Stumpf, S.; Liao, Q. V.; Baeza-Yates, R.; Aroyo, L.; Holbrook, J.; et al. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–4.

Veale, M.; and Zuiderveen Borgesius, F. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4): 97–112.

Wang, Q.; Madaio, M.; Kane, S.; Kapania, S.; Terry, M.; and Wilcox, L. 2023. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16.

Wang, Z. J.; Kulkarni, C.; Wilcox, L.; Terry, M.; and Madaio, M. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. *arXiv preprint arXiv:2402.15350*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Wu, T.; Terry, M.; and Cai, C. J. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.

Yang, X.; Liang, W.; and Zou, J. 2024. Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on HuggingFace. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Zhang, S.; Feng, Y.; and Sadeh, N. 2021. Facial recognition: Understanding privacy concerns and attitudes across increasingly diverse deployment scenarios. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, 243–262.

Zheng, Z.; Chen, K.-Y.; Cao, X.-Y.; Lu, X.-Z.; and Lin, J.-R. 2023. Llm-funcmapper: Function identification for interpreting complex clauses in building codes via llm. *arXiv preprint arXiv:2308.08728*.