Atlas of AI Risks: Enhancing Public Understanding of AI Risks

Edyta Bogucka[1], **Sanja Šćepanović**[1], **Daniele Quercia**[1,2]

[1]Nokia Bell Labs, Cambridge, UK
[2]Kings College London, London UK
edyta.bogucka@nokia-bell-labs.com.com, marios.constantinides@nokia-bell-labs.com,
sanja.scepanovic@nokia-bell-labs.com, daniele.quercia@nokia-bell-labs.com

Abstract

The prevailing methodologies for visualizing AI risks have focused on technical issues such as data biases and model inaccuracies, often overlooking broader societal risks like job loss and surveillance. Moreover, these visualizations are typically designed for tech-savvy individuals, neglecting those with limited technical skills. To address these challenges, we propose the Atlas of AI Risks—a narrative-style tool designed to map the broad risks associated with various AI technologies in a way that is understandable to non-technical individuals as well. To both develop and evaluate this tool, we conducted two crowdsourcing studies. The first, involving 40 participants, identified the design requirements for visualizing AI risks for decision-making and guided the development of the Atlas. The second study, with 140 participants reflecting the US population in terms of age, sex, and ethnicity, assessed the usability and aesthetics of the Atlas to ensure it met those requirements. Using facial recognition technology as a case study, we found that the Atlas is more user-friendly than a baseline visualization, with a more classic and expressive aesthetic, and is more effective in presenting a balanced assessment of the risks and benefits of facial recognition. Finally, we discuss how our design choices make the Atlas adaptable for broader use, allowing it to generalize across the diverse range of technology applications represented in a database that reports various AI incidents.

Introduction

Effectively communicating AI risks to ordinary individuals enables them to make informed decisions, advocate for their rights and interests, and push for better regulations and safer practices (Bao et al. 2022), while also limiting unrealistic AI perceptions and expectations (Nourani, King, and Ragan 2020; Neri and Cozman 2020). Such AI risks range from biases in decision-making to effects on jobs and collective freedoms (McGregor 2021).

In order to communicate risks of an AI technology use, the process of its specific risk discovery needs to take place. Current approaches to risk discovery, targeting AI practitioners, include harm description templates (Buçinca et al. 2023), impact assessment reports (Microsoft 2022; Stahl et al. 2023), interactive risk cards (Constantinides et al. 2024a), and databases of technical risks (IBM watsonx 2023; Robust Intelligence 2023). Risk discovery and AI system's impact assessment are also required to comply with regulations and standards like the European Union AI Act (EU AI Act) (European Comission 2024) and the NIST AI Risk Management Framework (NIST AI RMF) (National Institute of Standards and Technology 2023).

However, effectively communicating AI risks to ordinary individuals, even those with a strong interest in technology, remains a significant challenge. Ojewale et al. (2024) identified only two such communication approaches among the 390 AI auditing tools they surveyed. This challenge happens because tools like the AI Incident Database (CSET 2024b), which are made for experts, focus too much on technical risks, making it hard for regular people to understand how AI risks affect their lives.

Our study aims to bridge this gap by first crowdsourcing the design requirements and then developing a tool that uses information visualization techniques to fulfil these requirements and communicate AI risks to ordinary individuals interested in technology. We made four contributions:

1. We conducted a crowdsourcing formative study with 40 participants to identify requirements for the new risk communication tool. Using facial recognition as a case study, participants generated only 22 unique uses, paired with 18 risks, 9 mitigations, and 8 benefits. From their feedback, we derived six design requirements: *multiple uses (R1)*, *balanced assessment of uses (R2)*, *structured uses (R3)*, *reduced complexity (R4)*, *broad appeal (R5)*, and *engaging exploration (R6)*.
2. To meet design requirements *R1-R2* by increasing the number and variety of generated uses, risks, mitigations, and benefits, we used a Large Language Model (LLM) and a Generative Image Model (GIM). We started by using the LLM to devise 138 facial recognition uses across application domains. We then assessed each use for its risks and benefits. For each identified risk, where applicable, we generated mitigations that could be understood by individuals regardless of their technical knowledge. We used textual data from the LLM to generate GIM illustrations for each use and evaluated the whole generated content for correctness.

Project Website and Supplementary Materials:
https://social-dynamics.net/atlas

3. To clearly communicate the evaluated content to ordinary individuals interested in technology and meet the remaining design requirements *R3-6*, we employed information visualization techniques. These included visual groupings, visual metaphors, narrative patterns, interactions, and aesthetic styling, all of which contributed to the development of our tool – the Atlas of AI Risks.
4. We evaluated the Atlas in a crowdsourcing user study with 140 participants who reflect US population in terms of age, sex, and ethnicity, and compared it against the spatial view of the AI Incident Database. We found that the Atlas met all design requirements and participants preferred it both in terms of usability and aesthetics, finding it more helpful in shaping their decision-making process about facial recognition.

Related Work

Our work draws upon literature from diverse streams, which we organized into four areas as follows.

Visualizing technological risks. AI risk assessments typically focus on specific technologies, such as facial recognition (Moraes, Almeida, and de Pereira 2021), LLMs (Weidinger et al. 2021), and generative AI (Barrett et al. 2023). Visualization plays a crucial role in making these technologies' input data, components, and outputs more interpretable, aiding in identifying technological risks like data biases (Subramonyam and Hullman 2024; Inel, Draws, and Aroyo 2023). For example, visualizing dataset disparities can reveal sampling biases (Bellamy et al. 2018), and comparing data patterns with historical data can expose historical biases (IBM watsonx 2023). Comparing inputs and outputs can show inconsistencies in decision-making (Cabrera et al. 2023), and model vulnerabilities (Sietzen et al. 2021).

Overlooking human-interaction and systemic risks. A systematic analysis of model cards (Mitchell et al. 2019), an essential tool for documenting AI models, reveals that AI developers often emphasize technological risks related to data and models while neglecting impacts on individuals and the environment (Liang et al. 2024). This gap arises from the difficulty in predicting a wide range of risks from diverse model uses, stakeholder types, and deployment contexts (Boyarskaya, Olteanu, and Crawford 2020; Buçinca et al. 2023). To address this, Weidinger et al. (2023) propose assessing risks across three sociotechnical layers: capability, human interaction, and systemic impact. The capability layer evaluates risks inherent in technical features like poor model performance, the human interaction layer addresses risks from user interactions like overreliance on AI (Boyarskaya, Olteanu, and Crawford 2020), and the systemic impact layer considers broader societal and environmental consequences, including unequal distribution of technology's benefits and risks (Bellamy et al. 2018).

Visualizing risks for tech-savvy individuals. Eppler and Aeschimann (Eppler and Aeschimann 2009) noted that risk visualizations mainly target tech-savvy users, utilizing quantitative tools like matrices, bow-tie diagrams, and cognitive maps. AI risk visualizations often depict model inaccuracies with confusion matrices, accuracy curves, or activation maps, and are embedded in interactive systems handling diverse, high-dimensional data for real-time evaluation (Shergadwala, Lakkaraju, and Kenthapadi 2022; Subramonyam and Hullman 2024; Kwon et al. 2022; Johnson et al. 2023). Despite these tools, practitioners may misunderstand visualizations from interpretability tools (Interpret ML 2019; Lundberg and Lee 2017), overrelying on them as proof of model readiness (Kaur et al. 2020).

Overlooking visualizing risks for ordinary individuals. Effectively communicating AI risks to ordinary individuals, even those with a strong interest in technology, is challenging and requires understanding their beliefs, experiences, and media representations of AI (Buçinca et al. 2024; Cave, Dihal, and Dillon 2020). Communication techniques for this group can be categorized into three areas: *explaining*, *relating*, and *fostering* engagement with risks (Franconeri et al. 2021). To *explain* risks, visualizations should enhance comparisons and aid understanding, especially for those with low numeracy skills (Franconeri et al. 2021), using techniques like visual groupings and icon arrays. To *relate* risks to existing knowledge, visual metaphors like hazard labels can be effective (Zelenka et al. 2021), along with personalized risk presentations using individuals' own data (Luccioni et al. 2021). To *foster* engaging exploration, visualizations should employ narrative techniques, interactive features, and aesthetic styles to enhance engagement and understanding, using methods like story structures, gamified exploration, and contrasting colors (Segel and Heer 2010; Lavie and Tractinsky 2004; Crawford and Joler 2023).

Research gap. Past efforts in visualizing AI risks focused on technological aspects, neglecting human interaction and systemic risks. Additionally, these visualizations were often tailored to experts and tech-savvy users. To address this gap, we developed a new tool that shows broad AI risks and uses visualization techniques to make them accessible to ordinary individuals.

Proposing a Tool for Mapping Risks of AI Technology Uses for Ordinary Individuals

Identifying Design Requirements

The formative study aimed to generate AI technology uses, identify effective design techniques for understanding their trade-offs, and establish design requirements for the tool. We selected facial recognition technology as a case study because it is well-documented (Adjabi et al. 2020), it is relevant for both identifying humans (Moraes, Almeida, and de Pereira 2021) and animals (Roberts 2023), and it has sparked numerous public debates (Crawford 2019).

We recruited 40 participants interested in technology residing in the US through Prolific (Prolific 2014). These participants reflected the US population demographics (U.S. Census Bureau 2021, 2022) in terms of sex (21 males and 19 females) and ethnicity (24 White, 5 Black, 2 Asian, 4 Mixed or Other, and 1 Native American), with ages ranging from 19 to 63 years old. They were also digitally literate, exposed to visual media, and skilled in communication for providing feedback.

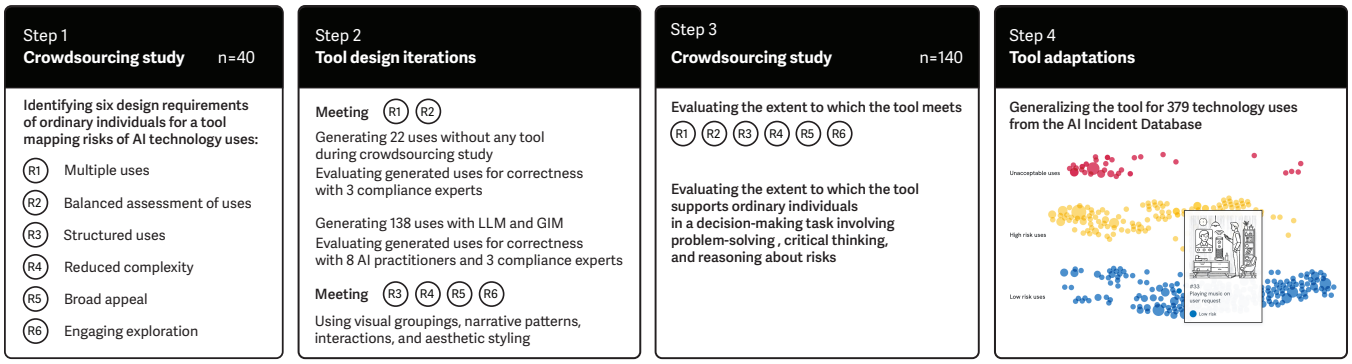

Figure 1: Proposing a tool for mapping risks of AI technology uses for ordinary individuals involved four steps. First, we conducted a crowdsourcing formative study to identify six design requirements for visualizing them. Next, using facial recognition as a case study, we asked participants of the formative study to generate its uses (including benefits, risks and mitigations) and evaluated them for correctness. Due to the low number of identified uses, we generated new uses by prompting the LLM and also evaluated them for correctness. Then, with the dataset in hand, we used information visualization techniques to build an interactive tool. We evaluated it against design requirements and its support for a decision-making task. Finally, to demonstrate how it generalizes, we visualized over 300 uses sourced from the AI Incident Database (McGregor 2021).

The study consisted of four parts. In the first part, we asked participants to rate on a scale from 1 to 5 (ranging from very low to very high) their knowledge in facial recognition, AI, and technology in general. In the second part, we asked them to write emails to regulators, requesting either a ban or further adoption of specific uses of facial recognition. They were also asked to explain their reasons by enumerating the risks, benefits, and steps to minimize the risks. In the third part, we presented participants with a list of 30 facial recognition uses (Adjabi et al. 2020; Roberts 2023), asking them how to best present their risks in an interactive tool. In the final part, participants were given three design techniques – an *explorative* dashboard, a *narrative* infographic, and a *simulative* tool with dropdowns – and asked to choose the best one and describe its pros and cons.

The study was approximately 30-45 minutes long and participants were paid on average about $12 (USD) per hour. We then conducted inductive thematic analysis (Miles and Huberman 1994), examining the uses mentioned in participant's emails and their recommendations for the tool. Participants self-reported slightly above-average knowledge in technology in general ($\mu = 3.8$), followed by average knowledge in AI ($\mu = 3.2$) and facial recognition ($\mu = 2.9$). Their recommendations (quotes are marked with FP) resulted in the following six design requirements for the tool:

(R1) Multiple uses. The tool should help participants to learn about a variety of uses instead of a limited number of them, as stated by FP30: *"it should prompt me to consider what this technology can do"*.

(R2) Balanced assessment of uses. The tool should present each use with its risks, benefits, and mitigation strategies. This can be achieved by providing *"concrete examples"* (FP13) and *'distinguishing between personal and societal risks and benefits"* (FP22).

(R3) Structured uses. The tool should categorize uses for better understanding, as FP14 stated, *"to help me get a clearer picture of the fields that use facial recognition"*.

(R4) Reduced complexity. The tool should present data on uses, risks, and mitigations, but its sheer volume can overwhelm users with limited technical backgrounds. To minimize this effect, the visualization should *"offer different depth levels"* (FP5) and *"break down the information into pieces to make it easier to come up with an opinion"* (FP21).

(R5) Broad appeal. The tool should make the uses, risks, benefits, and mitigation strategies accessible and relevant to individuals interested in technology, regardless of their technical background. As stated by FP20, *"examples should relate to issues and concerns that people commonly have about AI"*. The uses should be visualized in *"a less complicated way, not like designs for a technical audience"* (FP40).

(R6) Engaging exploration. The tool should engage users with *"many interactive elements to allow for deeper exploration"* (FP5) or a *"guided tour"* (FP28).

Over half of participants ($n = 22$) preferred the narrative technique for the tool because it closely matched R4 by simplifying complex data into understandable snippets, and R6 by engaging users with a coherent flow of information, which helps maintain the viewer's interest and attention. However, participants raised potential concerns that it could limit user flexibility to explore the data independently and introduce bias if it overfocuses on risks or benefits.

Meeting the First and Second Design Requirement

We outline the design decisions we made, informed by previous research and expert feedback, to ensure the tool effectively presents **(R1) Multiple uses** and **(R2) Balanced assessments of uses**.

Generating uses without any tool. To generate many uses of facial recognition and identify their associated risks, mitigations, and benefits, we adopted the EU AI Act's five-component definition of use (Golpayegani, Pandit, and Lewis 2023) and thematically analysed the participants' emails (Miles and Huberman 1994).

First, we identified phrases in the emails related to one of eight categories: *purpose* (the AI's end goal, e.g., verifying traveler identity at border controls), *capability* (technological solution behind the AI, e.g., matching faces to criminal databases), *AI subject* (those impacted by the AI, e.g., travelers), *AI user* (the entity managing the AI, e.g., border control agency), *domain* (the specific sector where the AI is applied, e.g., border control management), *risk* (e.g., infringing on the right to privacy), *mitigation* (e.g., implementing an opt-out option), and *benefit* (e.g., improving security measures). Second, we organized the phrases by category, grouping similar ones together. We summarize all the generated uses, risks, mitigations, and benefits in Supplementary Materials, Appendix A.

Evaluating uses generated without any tool. To evaluate the correctness and variety of generated uses, we introduced four quantitative metrics. The first metric assessed the number of correct uses by implementation potential and risk level per the EU AI Act. We defined correct use as technically feasible, considering its applicability and usability, and categorized its implementation potential into existing (currently in use), upcoming (in development or early prototype stage), and unlikely (lacking applicability and usability). The second, third, and fourth metrics assessend the number of correct risks, mitigations, and benefits, defined as those that are realistic and likely to occur or succeed when implemented. We categorized these into three types—technical capability, human interaction, and systemic impact—based on the existing taxonomy of socio-technical evaluations (Weidinger et al. 2023).

We independently assessed each generated use to first determine its correctness and implementation potential and then agreed on the final assessment. To assess the risk level of use, we recruited three AI compliance experts from our company who classified uses as unacceptable risk, high-risk, or low-risk, with justifications for each label.

Participants jointly identified 22 correct uses: 21 existing and 1 upcoming (Supplementary Materials, Appendix A). The top three uses mentioned were unlocking devices (n = 8), identifying suspects for crime prevention (n = 5), and apprehending individuals on the run (n = 4). Four uses were considered unacceptable (e.g., tracking citizens in public spaces for law enforcement), 15 high-risk, 3 low-risk, and 1 varied between high-risk and unacceptable depending on context. Participants identified 18 correct risks, 8 correct benefits, and 9 correct mitigations, mostly focusing on systemic impacts (n = 11, n = 6, and n = 4, respectively).

Generating uses with the LLM. Formative study participants primarily generated high-risk uses, overemphasized risks compared to benefits and mitigations, and barely addressed risks related to technical capability and human interaction, likely due to limited knowledge of AI development. These are, however, the most common risks resulting in incidents (McGregor 2021). To ensure the tool meets the two design requirements, we needed to increase the variety of generated uses, risks, mitigations, and benefits. We achieved this by using four prompts for LLM-assisted AI impact assessment — *ExploreGen* (Herdel et al. 2024), *RiskGen*, *BenefitGen*, and *MitigationGen* (Constantinides et al. 2024b; Bogucka et al. 2024) — and engineering one prompt for a generative image model — *IllustrationGen*. We report the prompts' content in Supplementary Materials, Appendix B.

Evaluating uses generated with the LLM. Similarly to the previous evaluation of uses generated without any tool, we introduced five quantitative metrics to evaluate the correctness and variety of generated uses. In addition to the already used four metrics – the number of correct risks, mitigations, and benefits by type – we evaluated each illustration for correctness of use depiction, defined as recognizable, relatable, and free of visual stereotypes.

Each metric was evaluated using a two-step process involving two authors and external experts. For assessing correctness, implementation potential, and depiction, we first independently evaluated each generated use, then discussed and agreed on the final assessment with the research team. For risk level, we familiarized ourselves with the EU AI Act (European Comission 2024) and randomly sampled 18 uses. We then recruited three AI compliance experts from our company to annotate these uses with risk labels and provide justifications. Afterwards, we independently assessed the remaining 120 uses for risk level, reaching a final consensus through discussions with the research team. For the number of correct risks, mitigations, and benefits by type, we recruited 8 raters: two authors and six industry researchers and developers from our company, all experienced in responsible AI. We randomly assigned 46 uses to the raters through an interactive survey, ensuring each use was evaluated by three raters. The raters reviewed the descriptions of the uses, along with three lists of risks, benefits, and mitigations. They marked whether they agreed with each list and indicated which items should be removed if they disagreed. We then compared the correctness scores for each list.

All 138 generated uses to be correct, with 91 (66%) already existing, 39 (28%) being upcoming, and 8 (6%) being unlikely. The agreement between the authors and experts and the LLM's classification was 91%. Disagreements were about 12 uses related to environmental sustainability, agriculture, farming, and climate change mitigation, which LLM marked as existing, and the experts as unlikely. 10 uses (7%) were identified as unacceptable, 66 (48%) as high risk, and 62 (45%) as limited or low risk. The agreement between the authors and experts and the LLM's classification was 90%. Disagreements were about 14 uses, for which experts found insufficient information to derive risk label, while the LLM classified them as low risk.

By comparing the correctness scores, we found that 93% of the risks, 95% of the mitigations, and 82% of the benefits were correct. The average agreements across each set of three annotators labeling the same parts of the data, as measured using the intraclass correlation coefficient, were 25% for the risks (fair), 23% for the mitigations (fair), and 47% for the benefits (moderate agreement). As for the illustrations of uses, 126 (91%) were correct, while 12 (9%) needed re-generation due to references to national symbols (n = 2), incorrect cultural depictions (n = 5), and insensitive gender role representations (n = 5).

Meeting the Remaining Four Design Requirements

We describe the design decisions, informed by previous research on effective communication with the public, made to ensure the tool meets the remaining design requirements.

(R3) Structured uses. We use a map atlas metaphor, with each use as a dot in a two-dimensional space. This visually emphasizes the need to assess risks per use and helps people understand the overall probability of risks of facial recognition (Franconeri et al. 2021). We then adopted three methods for varying analytical skills: spatial distribution based on semantic similarity (for a continuous view), functionality to split uses by risk (for group comparison), and new data dimensions for color-coding uses (for a discrete view). First, to spatially distribute the uses (Figure 2a), we created sentence-level BERT (SBERT) embeddings for each use description using the *paraphrase-distilroberta-base-v2* model. We used SBERT's standard settings, and showed the resulting embeddings using a JavaScript-based t-distributed stochastic neighbor embedding algorithm. Second, to facilitate comparisons between the high-risk and low-risk uses, we split their dots vertically (Figure 2b) and added the interaction to group the uses back together (Franconeri et al. 2021). Third, to color-code the data, we used the existing AI Harm Taxonomy for the AI Incident Database (CSET 2024a). We manually label each use case based on its area of application, the types of affected subjects and supervising users, as well as its impacts on critical infrastructure, children, entertainment, and the public sector (Figure 2c).

(R4) Reduced complexity. We adopted a frame-based Martini Glass narrative structure (Segel and Heer 2010) and a progressive disclosure design (Norman and Draper 1986) to reveal information gradually.

First, we unfold the complexities of assessing risks through five distinct story sections (Figure 2, R4). The first section explains the technological features of facial recognition and introduces its 138 uses, each represented by a dot. The second section highlights the dots representing daily uses of facial recognition. The third section shows how uses can vary in risk, illustrated through an animated transition that categorizes the dots into unacceptable, high, or low risk groups. The fourth section explains the common characteristics of each risk group, emphasizing through color-coding how, despite regulations, all uses can still pose harm. The final section introduces a dashboard that encourages users to explore the uses and to find ways to mitigate risks.

Second, we paired each use with an impact assessment card that comes in two versions: a brief tooltip accessible with mouseover (Figure 3a), and a detailed profile accessible with a click (Figure 3b). The tooltip includes an illustration of the use, its short description, and a tag indicating its level of risk according to the EU AI Act. The profile includes four sections, starting with a use summary box that contains an illustration, a long description, and the overall risk level. The subsequent three sections detail the benefits, risks, and mitigations, with benefits and risks grouped by technical capability, human interactions, and social impact, and checkboxes indicating who is affected.

(R5) Broad appeal. We used different colors to separate the uses into two groups: daily (like unlocking smartphones) and non-daily (like tracking illegal poaching). Nearly half of the dots represented daily uses, showing how facial recognition technology is part of everyday life and making the data more relatable (Figure 2, R5). Additionally, we carefully phrased the mitigations in the impact assessment card to be understandable regardless of (non)technical background.

(R6) Engaging exploration. The final dashboard includes interactions for atlas browsing, onboarding, and exploration tracking (Figure 3). Users can browse the atlas using the search box for keyword matching (Figure 3c) and the filter box for color-coding dots based on ten categories (Figure 3d). The onboarding guide covers the interface, tooltips, profiles, search, and filtering options (Figure 3e). For exploration tracking, micro-interactions include a counter that changes color as more uses are explored (Figure 3f) and animations for already explored uses (Figure 3g).

Evaluating the Tool Mapping Risks of AI Technology Uses

The goal of the study was to evaluate if our Atlas met the design requirements and how effectively it communicated the broad risks associated with the diverse uses of facial recognition to ordinary individuals. Next, we describe our study's design (i.e., metrics), setup, execution, and results.

Metrics

We created a task for our study that reflects how people usually handle and question AI decisions in real life, such as writing an email to challenge or praise a system(Alfrink et al. 2023). Our participants might have done similar things before, like giving feedback to a company about its smart devices or checking their town's AI city register, where local governments list technologies used in their cities (City of Helsinki 2023). Specifically, we instructed participants to "Write a brief email to the AI policymakers and ask them to stop and approve some uses of facial recognition. For each use, argue by explaining its risks or benefits". This task also links the risks in the Atlas to three higher-order decision-making skills: problem-solving (identifying reasonable actions), critical thinking (evaluating and synthesizing information on risks, mitigation, and benefits), and reasoning (constructing logical arguments to support actions).

We outlined six questions to evaluate whether Atlas met our design criteria (R1-6) and how effectively it supported individuals in completing the task, asking, "How successful was the tool in...

Q1 ...communicating multiple uses (R1)?
Q2 ...providing a balanced assessment of uses (R2)?
Q3 ...structuring uses (R3)?
Q4 ...reducing complexity (R4)?
Q5 ...achieving a broad appeal (R5)?
Q6 ...engaging individuals in exploration? (R6)?

We then defined four quantitative and four qualitative metrics to answer these questions.

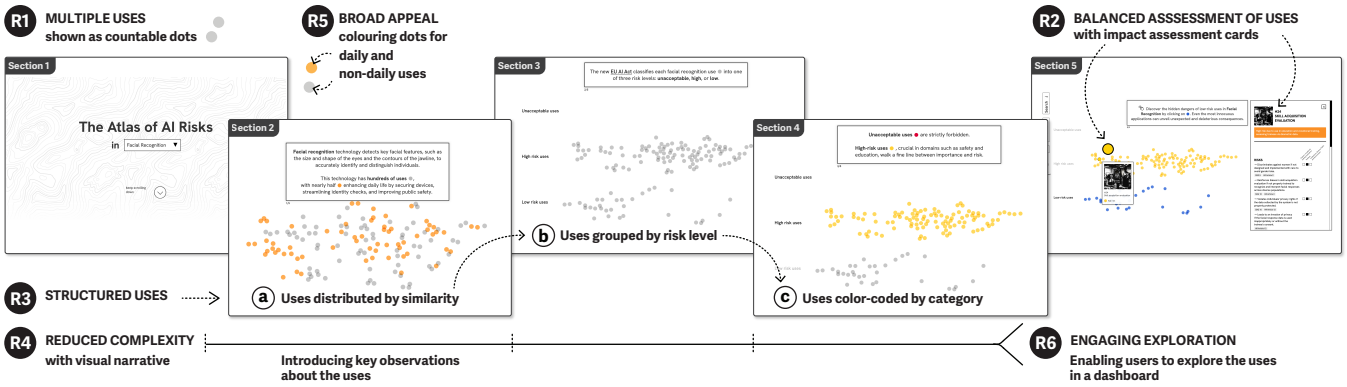

Figure 2: The interface of the Atlas of AI Risks meets six design requirements: mapping many uses of technology (R1), presenting a balanced assessment of their risks and benefits (R2) categorizing them for better understanding (R3), reducing their complexity (R4), making them relevant to ordinary individuals (R5), and making their exploration engaging (R6).

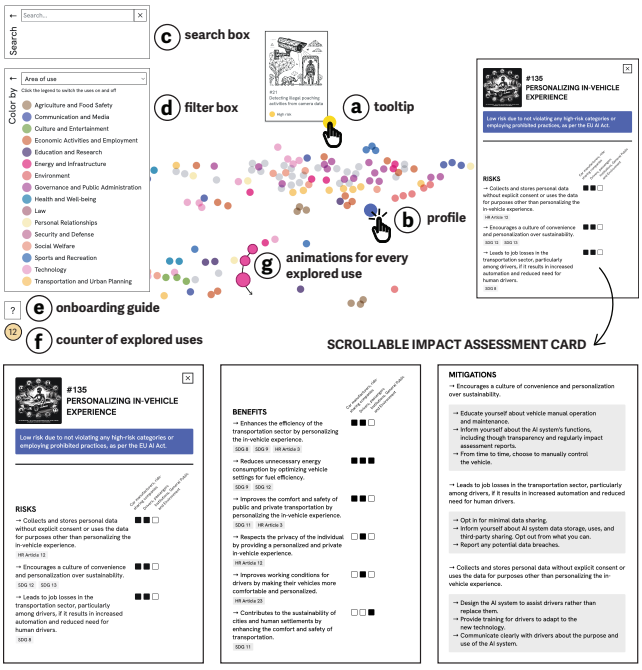

Figure 3: The final dashboard for use exploration. It includes an impact assessment card available in two versions—a brief tooltip (a) and a detailed profile (b) listing risks, benefits, and mitigations—as well as interactions for use browsing, onboarding, and exploration tracking (c-g).

Quantitative metrics. The first quantitative metric assessed the extent to which the tool provided a balanced assessment of uses. It was measured as the percentage of participants who agreed with the statement: *"The tool helped me to understand both the risks and benefits of facial recognition"*. Three metrics measured how successful the tool was in engaging users in exploration.

The first metric was the about the *usability* of the tool and it was measured through the System Usability Scale (Brooke 1996). The second metric was about the *visual aesthetics*, understood as classic aesthetics (for clarity and order), expressive aesthetics (for originality), and pleasurable interaction (for user enjoyment), three factors crucial for making technology use exploration intuitive and engaging for ordinary individuals. We measured them through the Perceived Visual Aesthetics scale (Lavie and Tractinsky 2004). The third metric was about the *exploration time* required for a participant to complete the task.

Qualitative metrics. Qualitative metrics were captured through four open-ended questions. The first assessed the effectiveness in communicating multiple uses, by asking: *"How successful was the tool in helping you learn about a variety of uses?"* The second measured the usefulness of structuring uses, by asking *"How useful were for you the categories of uses shown in the tool?"*. The third measured effectiveness in reducing complexity by asking: *"How successful was the tool in simplifying the information?"* The fourth assessed the *relevance* of presented information to ordinary individuals: *"How successful was the tool in identifying uses, risks, benefits, and mitigations relevant to you?"*.

To assess the impact of knowledge on responses, participants self-evaluated whether they were more skilled or knowledgeable than the average person in the task, technology, facial recognition, and AI.

Setup

The study consisted of seven steps (Supplementary Materials, Appendix C). In the first step, participants were introduced to the first reflective judgment task and were asked to write the first email to the regulators without using any tools. In the second step, participants completed the self-assessment control questions, alongside the first randomly assigned attention-check. Moving to the third step, participants interacted with either our Atlas (treatment) or the baseline (control), analyzed the technology's uses in the visualization and completed the second reflective judgment task. After interaction with the treatment or control, participants evaluated its usability and completed the second attention-check sentence in the fourth step. In the fifth step, participants assessed the visualization's visual aesthetics.

This step also included the third attention check. In the sixth step, participants explained if the tool helped them to learn about multiple uses of facial recognition, if the proposed categories of uses were useful, and if the content presented in the visualization was relevant to them. Finally, in the last, seventh step participants provided recommendations for future development of the visualization.

Baseline. The control group's visualization mimicked state-of-the-art AI risk visualization – a dashboard (Supplementary Materials, Appendix D, Figure 10). It displayed technology uses grouped by similarity on the left, pop-ups for the uses in the center, and a dropdown menu with a legend on the right. Like the Atlas, it displayed various uses (R1), categorized them (R2), and allowed exploration (R6). However, it differed in balancing risks and benefits (R2), reducing complexity (R4), and appealing to audience (R5).

Execution

Participants. We recruited participants from Prolific (Prolific 2014), aiming for a sufficiently large number in both the treatment and baseline groups to achieve statistical significance. The study focused on individuals interested in technology, who are representative of the general U.S. population, for two key reasons. First, recruiting English native speakers ensured comprehension of the study materials, enhancing results reliability. Second, exposing US participants to a risk-based approach to AI regulation is relevant, as both the US and EU adopt similar frameworks, with crosswalk documents showing their compatibility. Recruited study participants were compensated approx. $12 (USD) per hour.

Procedure. We developed a web-based survey that included a reflective judgment task and administered it on Prolific. The survey comprised seven pages, each corresponding to a setup step plus a final confirmation screen. In step three, participants were randomly presented with a link to access either the treatment or the control in a new browser window.

To ensure response quality, we implemented four survey design features. First, we disabled pasting text from external sources and prohibited editing previous responses to encourage original answers and maintain survey flow. Second, we set word ranges for open-ended questions (50-250 words (Liang et al. 2024)) and validated them in real-time to prevent survey fatigue. Third, we used a click tracker to verify participants accessed the treatment or baseline link. Fourth, we randomly administered one of three attention checks. To be included, participants had to correctly respond to at least two attention checks and click the link.

Analysis. We performed both quantitative and qualitative analyses. For the quantitative analysis, we measured four metrics: the share of participants who found the visualization helpful for understanding both the risks and benefits; the average SUS usability score; the average value of perceived aesthetics across classic aesthetics, expressive aesthetics, and pleasurable interaction; and the average time to complete the task. For the qualitative analysis, we thematically analyzed the open-ended questions (Miles and Huberman 1994) to identify repeating themes within design requirements and key factors affecting decision-making.

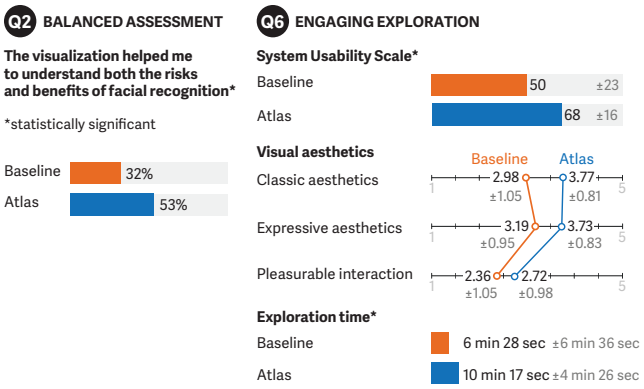

Figure 4: The Atlas outperformed baseline across all quantitative metrics. It offered a more balanced assessment of uses, scored higher in usability, visual aesthetics, and encouraged longer exploration time.

Results

We received a total of 140 responses, split equally into treatment and control (Supplementary Materials, Appendix E, Table 1). Participants reflected the US population in terms of age, sex, and ethnicity.

Quantitative results. The Atlas provided a more balanced assessment of uses, with 53% of participants finding it offered comprehensive perspectives on risks, benefits, and mitigations, compared to 32% of those who engaged with the baseline (Figure 4Q2).

Participants found the Atlas more usable than the baseline with an average SUS score of 50, while the Atlas scored 68 (Figure 4Q6). These higher scores were consistent across all levels of technological knowledge (Supplementary Materials, Appendix F, Figure 11). They also spent more time exploring the Atlas, indicating higher engagement.

Participants found the Atlas more aesthetically pleasing than the baseline and did so across all three dimensions of aesthetics (Figure 4Q6). In *classic aesthetics*, it provided a better-ordered structure. In *expressive aesthetics*, it reflected more innovative way for presenting risks. In *pleasurable interaction*, it was regarded as moderately successful, indicating room for improvement.

Qualitative results. Participants (referred to as CP) found that the tool supported task completion by reducing the complexity of information about facial recognition (n=20), providing numerous examples of uses (n = 18), listing their pros and cons (n = 17), and effectively presenting them visually in the tool (n = 17 mentions). As summarized by CP22, it *"gave both quick tidbits, but also in depth explanations"*, while *"making transitioning from one use to another feel smooth and understandable"* (CP26). The tool did not support participants who had strong pre-existing opinions about facial recognition (n = 20) and needed more details to deliberate about each use (n = 19). For example, CP9 felt they *"didn't see much relating to environment"*, and CP6 wished for a feature where *"AI can explain a picked dot like I'm 5"*.

Demonstrating the Generalizability of the Tool

To make sure the Atlas can handle any kind of AI, we use a straightforward five-part format. This format consists of: *purpose* (what the AI is meant to do), *capability* (what the AI can actually do), *AI user* (who uses the AI), *AI subject* (on whom the AI works), and *domain* (the area where the AI is used). This format is so flexible that it was first proposed to assess the risks of any AI system according to the EU AI Act (Golpayegani, Pandit, and Lewis 2023).

To then demonstrate the Atlas can handle any kind of AI beyond facial recognition, we populated it with 379 real-world AI applications that resulted in news incident reports. We sourced the incident descriptions from the AI Incident Database (AIID), a standardized and verified collection of AI-related harmful events (McGregor 2021). We populated the Atlas in five steps. First, we downloaded all 649 incident descriptions (e.g., "YouTube's recommendation algorithms exposed children to disturbing videos"). Second, we used the *ExploreGen* prompt to generate a possible AI use that could have caused the incident, breaking it down into the previously defined five components (e.g., "Purpose": "Recommending suitable videos for children", "AI Capability": "Content filtering", "AI User": "YouTube", "AI Subject": Children", "Domain": "Recommender Systems"). Third, we reviewed the formatted uses and merged the duplicate and most similar ones based on their semantic similarity and overlaps in our Atlas (e.g., 53 incident descriptions related to opearating autonomous taxis and delivery vehicles were merged into one use: "Purpose": "Operating autonomus vehicles", "AI Capability": "Acting on sensor readings for navigation", "AI User": Autonomous vehicle providers", "AI Subject": "Road users", "Domain": "Public and private transportation"). This review resulted in a final set of 379 uses. Fourth, we ran *RiskGen*, *BenefitGen*, *MitigationGen*, and *IllustrationGen* on each use to complete the content of the impact assessment card. Finally, we updated the Atlas's source code to include a dropdown menu, enabling users to explore all technologies from the incident database (Supplementary Materials, Appendix G, Figure 12) or to focus on individual technologies like mobile computing (Supplementary Materials, Appendix G, Figure 13). The extended Atlas is available online at https://social-dynamics.net/atlas.

Based on the successful testing of its generalizability, we suggest three scenarios of how the Atlas could facilitate public debates and advocating for regulatory changes. First, by integrating the Atlas into AI city registers, where governments catalog urban technologies (City of Helsinki 2023), citizens could better assess risks like privacy concerns before engaging in local discussions. Second, the Atlas could be integrated into consumer AI databases, such as the upcoming EU database for high-risk AI systems (European Comission 2024), allowing individuals to evaluate AI products and make informed purchases. Finally, in education, the Atlas could serve students studying AI's ethical implications across different industries (Feffer, Martelaro, and Heidari 2023).

Discussion and Conclusion

Weighing Explanation and Exploration. The goal of our tool is to effectively communicate AI uses and their risks to ordinary individuals interested in technology. This could also be achieved through other, more traditional methods, such as tutorials or training sessions. However, these approaches have two major limitations. First, they are often static and linear, limiting users' ability to explore content in a personalized way. Second, they demand significant time, which can overwhelm individuals with lower AI literacy. Instead, we developed a tool that aligns with the shift from static documentation, like impact assessment reports, to dynamic, glanceable formats like interactive model cards. By prioritizing visual narratives and aesthetic design, we encouraged users to spend more time exploring the tool and avoided overly technical designs that could alienate those who most need accessible learning tools.

Future work. We identify two areas for future work in data collection and visual presentation in similar tools. First is the exploration of complementary approaches that integrate collecting public perceptions and uses, expert assessments, and insights from LLMs. Second is the inclusion of visual features that help effectively build consensus around technology. These could include voting buttons where users can express their agreement or disagreement with certain assessments, and displaying aggregated community votes.

Limitations. This research comes with three limitations. First, LLM deployment faces issues like biases in training data that may overlook important risks and benefits (Luccioni et al. 2024), while strict safety measures can omit risky uses already known to the public from past incidents (Robust Intelligence 2023). Addressing these issues requires fine-tuning models with specific technology datasets. For the Atlas, we addressed them by validating the content and manually removing any incoherent information.

Second, the Atlas usability score partly reflects the novelty and design challenge for similar tools and audiences. Our study's sample may not completely represent ordinary individuals interested in technology, due to limited controls over their location, age, ethnicity, gender, experience with AI and preferences in data visualization. Third, the study reflects US perspectives, which may not apply globally, as perceptions of technology risks and benefits vary by country. For example, facial recognition is most accepted in China, less so in Germany, with the US and UK in between (Kostka, Steinacker, and Meckel 2021). While most Chinese citizens emphasize benefits like convenience. Americans, Germans, and Brits are more concerned about risks such as surveillance (Kostka, Steinacker, and Meckel 2023). Therefore, our findings should be interpreted with these limitations in mind.

Our work initially focused on designing and evaluating a tool for ordinary individuals interested in technology. However, we found that even non-tech savvy users could effectively engage with it when prompted, as the tool's usability remained consistent across users with different levels of technological knowledge. This demonstrates the Atlas's ability to make complex AI technology risks understandable, filling the gap in current risk communication methods.

References

Adjabi, I.; Ouahabi, A.; Benzaoui, A.; and Taleb-Ahmed, A. 2020. Past, Present, and Future of Face Recognition: A Review. *Electronics*, 9(8).

Alfrink, K.; Keller, I.; Doorn, N.; and Kortuem, G. 2023. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*.

Bao, L.; Krause, N. M.; Calice, M. N.; Scheufele, D. A.; Wirz, C. D.; Brossard, D.; Newman, T. P.; and Xenos, M. A. 2022. Whose AI? How Different Publics Think About AI and its Social Impacts. *Computers in Human Behavior*, 130: 107182.

Barrett, C.; Boyd, B.; Bursztein, E.; Carlini, N.; Chen, B.; Choi, J.; Chowdhury, A. R.; Christodorescu, M.; Datta, A.; et al. 2023. Identifying and Mitigating the Security Risks of Generative AI. *Foundations and Trends in Privacy and Security*, 6(1): 1–52.

Bellamy, R. K. E.; et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943.

Bogucka, E.; Constantinides, M.; Šćepanović, S.; and Quercia, D. 2024. AI Design: A Responsible AI Framework for Pre-filling Impact Assessment Reports. *IEEE Internet Computing*.

Boyarskaya, M.; Olteanu, A.; and Crawford, K. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. arXiv:2011.13416.

Brooke, J. 1996. SUS: A "Quick and Dirty" Usability Scale. *Usability Evaluation in Industry*, 189(3): 189–194.

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: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:2403.05911.

Cabrera, A. A.; Tulio Ribeiro, M.; Lee, B.; Deline, R.; Perer, A.; and Drucker, S. M. 2023. What Did My AI Learn? How Data Scientists Make Sense of Model Behavior. *ACM Transactions on Computer-Human Interaction*, 30(1).

Cave, S.; Dihal, K.; and Dillon, S. 2020. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford University Press. ISBN 9780198846666.

City of Helsinki. 2023. Artificial Intelligence Systems of Helsinki. https://ai.hel.fi/en/ai-register. Accessed: 2024-08-20.

Constantinides, M.; Bogucka, E.; Quercia, D.; Kallio, S.; and Tahaei, M. 2024a. RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. In *Proceedings of the ACM on Human-Computer Interaction*, CSCW, 1–28.

Constantinides, M.; Bogucka, E.; Scepanovic, S.; and Quercia, D. 2024b. Good Intentions, Risky Inventions: A Method for Assessing the Risks and Benefits of AI in Mobile and Wearable Uses. *Proceedings of the ACM on Human-Computer Interaction*, 8: 1–30.

Crawford, K. 2019. Halt the Use of Facial-Recognition Technology Until It Is Regulated. *Nature*, 572(7771): 565–566.

Crawford, K.; and Joler, V. 2023. Calculating Empires: A Genealogy of Technology and Power Since 1500. https://calculatingempires.net. Accessed: 2024-08-20.

CSET. 2024a. AI Harm Taxonomy for AIID. https://incidentdatabase.ai/taxonomy/csetv1. Accessed: 2024-08-20.

CSET. 2024b. Spatial Visualization of AI Incident Database. https://incidentdatabase.ai/summaries/spatial. Accessed: 2024-08-20.

Eppler, M. J.; and Aeschimann, M. 2009. A Systematic Framework for Risk Visualization in Risk Management and Communication. *Risk Management*, 11(2): 67–89.

European Comission. 2024. Regulation of the European Parliament and of the Council on Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf. Accessed: 2024-08-20.

Feffer, M.; Martelaro, N.; and Heidari, H. 2023. The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, and Future Improvements. In *Proceedings of the ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–11.

Franconeri, S. L.; Padilla, L. M.; Shah, P.; Zacks, J. M.; and Hullman, J. 2021. The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, 22(3): 110–161.

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 ACM Conference on Fairness, Accountability, and Transparency*, 905–915.

Herdel, V.; Šćepanović, S.; Bogucka, E.; and Quercia, D. 2024. ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Hugging Face. 2024. LMSYS Chatbot Arena Leaderboard. https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard. Accessed: 2024-08-20.

IBM watsonx. 2023. AI Risk Atlas. https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/ai-risk-atlas.html?context=wx. Accessed: 2024-08-20.

Inel, O.; Draws, T.; and Aroyo, L. 2023. Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, 51–64.

Interpret ML. 2019. A Toolkit To Help Understand Models and Enable Responsible Machine Learning. https://interpret.ml. Accessed: 2024-08-20.

Johnson, N.; Cabrera, Á. A.; Plumb, G.; and Talwalkar, A. 2023. Where Does my Model Underperform? A Human Evaluation of Slice Discovery Algorithms. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, 65–76.

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 ACM Conference on Human Factors in Computing Systems*, 1–14.

Kostka, G.; Steinacker, L.; and Meckel, M. 2021. Between Security and Convenience: Facial Recognition Technology in the Eyes of Citizens in China, Germany, the United Kingdom, and the United States. *Public Understanding of Science*, 30(6): 671–690.

Kostka, G.; Steinacker, L.; and Meckel, M. 2023. Under Big Brother's Watchful Eye: Cross-Country Attitudes Toward Facial Recognition Technology. *Government Information Quarterly*, 40(1): 101761.

Kwon, B. C.; Kartoun, U.; Khurshid, S.; Yurochkin, M.; Maity, S.; et al. 2022. RMExplorer: A Visual Analytics Approach To Explore the Performance and the Fairness of Disease Risk Models on Population Subgroups. In *IEEE Visualization and Visual Analytics*, 50–54.

Lavie, T.; and Tractinsky, N. 2004. Assessing Dimensions of Perceived Visual Aesthetics of Web Sites. *International Journal of Human-Computer Studies*, 60(3): 269–298.

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:402.05160.

Luccioni, A.; Schmidt, V.; Vardanyan, V.; and Bengio, Y. 2021. Using Artificial Intelligence to Visualize the Impacts of Climate Change. *IEEE Computer Graphics and Applications*, 41(1): 8–14.

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2024. Stable Bias: Evaluating Societal Representations in Diffusion Models. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the International Conference on Neural Information Processing Systems*, 4768–4777.

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.

Microsoft. 2022. Microsoft's Framework for Building AI Systems Responsibly. https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly. Accessed: 2024-08-20.

Miles, M.; and Huberman, M. 1994. *Qualitative Data Analysis: A Methods Sourcebook*. Sage. ISBN 9781506353074.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 220–229.

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.

National Institute of Standards and Technology. 2023. AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework. Accessed: 2024-08-20.

Neri, H.; and Cozman, F. 2020. The Role of Experts in the Public Perception of Risk of Artificial Intelligence. *AI & Society*, 35: 663–673.

Norman, D. A.; and Draper, S. W. 1986. *User Centered System Design; New Perspectives on Human-Computer Interaction*. USA: L. Erlbaum Associates Inc. ISBN 0898597811.

Nourani, M.; King, J.; and Ragan, E. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 112–121.

Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. arXiv:2402.17861.

Prolific. 2014. Easily Find Vetted Research Participants and AI Taskers. https://www.prolific.com. Accessed: 2024-08-20.

Roberts, F. S. 2023. Socially Responsible Facial Recognition of Animals. *AI and Ethics*.

Robust Intelligence. 2023. AI Risk Databases. https://airisk.io. Accessed: 2024-08-20.

Segel, E.; and Heer, J. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6): 1139–1148.

Shergadwala, M. N.; Lakkaraju, H.; and Kenthapadi, K. 2022. A Human-Centric Perspective on Model Monitoring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 173–183.

Sietzen, S.; Lechner, M.; Borowski, J.; Hasani, R.; and Waldner, M. 2021. Interactive Analysis of CNN Robustness. *Computer Graphics Forum*, 40(7): 253–264.

Stahl, B. C.; Antoniou, J.; Bhalla, N.; Brooks, L.; Jansen, P.; Lindqvist, B.; Kirichenko, A.; Marchal, S.; Rodrigues, R.; Santiago, N.; Warso, Z.; and Wright, D. 2023. A Systematic Review of Artificial Intelligence Impact Assessments. *Artificial Intelligence Review*, 56(11): 12799–12831.

Subramonyam, H.; and Hullman, J. 2024. Are We Closing the Loop Yet? Gaps in the Generalizability of VIS4ML Research. *IEEE Transactions on Visualization and Computer Graphics*, 30(01): 672–682.

United Nations. 1961. Universal Declaration of Human Rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights. Accessed: 2024-08-20.

United Nations. 2023. The 17 Sustainable Development Goals. https://sdgs.un.org/goals. Accessed: 2024-08-20.

U.S. Census Bureau. 2021. Race and Ethnicity in the United States: 2010 Census and 2020 Census. https://www.census.gov/library/visualizations/interactive/race-and-ethnicity-in-the-united-state-2010-and-2020-census.html. Accessed: 2024-08-20.

U.S. Census Bureau. 2022. Population by Age and Sex. Annual Social and Economic Supplement. https://www.census.gov/library/visualizations/interactive/race-and-ethnicity-in-the-united-state-2010-and-2020-census.html. Accessed: 2024-08-20.

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 *Advances in Neural Information Processing Systems*, volume 35, 24824–24837.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986.

Weidinger, L.; et al. 2021. Ethical and Social Risks of Harm from Language Models. arXiv:2112.04359.

Zelenka, N.; Di Cara, N.; Day, H.; et al. 2021. Data Hazard Labels. https://datahazards.com. Accessed: 2024-08-20.