EUROPEAN JOURNAL OF EMERGING ARTIFICIAL INTELLIGENCE

LEVERAGING ANALOGIES FOR AI EXPLAINABILITY: ENHANCING LAYPERSON UNDERSTANDING IN AI-ASSISTED DECISION MAKING

Dr. William Harper
School of Engineering and Applied Sciences, Harvard University, USA

Ethan Navarro
Department of Computer Science, University of California, Berkeley, USA

Dr. Luca Conti
Department of Informatics, Technical University of Munich, Germany

V0LUME01 ISSUE01 (2024)
Published Date: 13 December 2024 // Page no.: - 37-53

ABSTRACT

The integration of Artificial Intelligence (AI) into critical decision-making processes necessitates transparent and understandable explanations, especially for non-expert users (laypeople). While traditional Explainable AI (XAI) methods often present technical details that remain inscrutable to a lay audience, this article investigates the potential of analogy-based explanations to bridge this knowledge gap. We present a two-part empirical study. Study I focuses on the generation and qualitative assessment of analogy-based explanations using non-expert crowd workers, establishing a systematic framework for evaluating their quality across dimensions such as structural correspondence, relational similarity, and familiarity. Our findings highlight the subjective nature of analogy quality and the potential for leveraging crowdsourcing to generate diverse explanations. Study II evaluates the practical effectiveness of these analogy-based explanations in a high-stakes medical diagnosis task (skin cancer detection). Surprisingly, quantitative results did not show a significant improvement in understanding or appropriate reliance with analogy-based explanations compared to detailed concept-level explanations. However, qualitative feedback revealed that users found analogies helpful when they perceived a strong connection to a familiar source domain and when presented on demand. While explanations, including analogies, increased perceived cognitive load and decision-making time, our comprehensive analysis points to the crucial roles of human intuition and perceived plausibility in shaping user behavior. This research contributes actionable insights for designing human-centered XAI, emphasizing the need for personalized and carefully crafted analogies to truly enhance layperson understanding and foster appropriate reliance in AI-assisted decision-making.

Keywords: Explainable AI (XAI), Analogical Reasoning, Human-AI Collaboration, Layperson Understanding, Trust in AI, Cognitive Load, Commonsense Knowledge, Crowdsourcing, Medical Diagnosis.

INTRODUCTION

The increasing prevalence of Artificial Intelligence (AI) systems across diverse domains, from personalized health recommendations to critical diagnostic support in medicine, signifies a growing societal reliance on AI-assisted decision-making [56, 42]. As these AI models become more sophisticated and autonomous, the imperative to ensure their transparency, interpretability, and comprehensibility, particularly for non-expert users (often referred to as laypeople), has become a central focus within the AI community [4, 22, 65]. The field of Explainable AI (XAI) has emerged to address this challenge, striving to render complex AI behaviors understandable to human decision-makers [17, 27, 28]. While notable advancements have been made in developing various XAI techniques—ranging from feature attribution methods that highlight salient parts of input [71, 81, 47] to more abstract concept-based explanations [38, 57, 105]—a persistent and critical hurdle remains: effectively communicating the intricate reasoning processes of AI to non-technical users who may possess a significant knowledge gap [17, 27, 28, 15]. This challenge is compounded by the "illusion of explanatory depth" [17], where individuals may perceive they understand an AI's rationale superficially, yet lack a true grasp of its underlying mechanisms. Such a superficial understanding can lead to a miscalibration of trust and reliance, where users either over-rely or under-rely on AI advice, potentially resulting in suboptimal or even detrimental outcomes in real-world scenarios [64, 69, 103, 108].

To effectively bridge this knowledge chasm, researchers are increasingly gravitating towards human-centered XAI approaches. These methodologies extend beyond merely exposing algorithmic mechanics, instead focusing on the cognitive and psychological factors that influence human

understanding and interaction with AI [27, 28, 65]. Among these promising avenues, the strategic use of analogies stands out. Analogical reasoning, a cornerstone of human cognition, enables individuals to comprehend novel or complex concepts by drawing parallels to familiar experiences or knowledge structures [36, 37, 47, 8]. Analogies serve as powerful cognitive shortcuts, facilitating the learning process and enhancing memory by connecting new information to existing mental frameworks [42]. In educational pedagogy, analogies have long been celebrated for their efficacy in demystifying complex scientific principles [19, 25, 35, 74, 75, 82]. For instance, illustrating the flow of electricity by comparing it to the movement of water in pipes can simplify an otherwise abstract physical phenomenon [19]. Given this established utility in human learning and sense-making, analogies present a significant opportunity to improve the comprehensibility and effectiveness of AI explanations for a broad audience of laypeople [8, 50, 78].

Prior investigations have touched upon the application of analogies within various AI contexts, spanning tasks such as generating lexical analogies [16], performing multi-relational embeddings [68], and even in the design of games aimed at eliciting diverse knowledge [5]. More recently, analogies have found a particular niche in crafting concept-level AI explanations, drawing upon vast repositories of commonsense knowledge to elucidate AI behaviors. A salient example is explaining why an AI might incorrectly classify a polar bear in a savannah image, by relating it to a more relatable scenario [43, 36]. This approach aligns with broader research efforts to infuse AI systems with commonsense knowledge, fostering more intuitive and human-like reasoning and explainability [21, 51, 53, 72, 80, 88, 91, 100, 104, 106]. Furthermore, preliminary studies suggest that employing analogies can cultivate appropriate reliance in human-AI decision-making, enabling users to more accurately calibrate their trust in AI based on its stated accuracy [44, 45]. However, despite these promising indicators, a comprehensive and rigorous empirical investigation into the multifaceted effectiveness of analogy-based explanations—encompassing metrics such as genuine understanding, appropriate reliance, cognitive load experienced by users, and overall decision performance—specifically tailored for lay users in AI-assisted decision-making contexts, remains a critical area needing extensive exploration.

This article aims to conduct such a comprehensive investigation, seeking to determine whether explanations grounded in analogies can demonstrably enhance laypeople's understanding of AI systems and cultivate appropriate reliance in AI-assisted decision-making. Building upon existing theoretical foundations and preliminary observations, we formulate several key hypotheses for our empirical inquiry:

- RQ1: How can we generate high-quality analogy-based explanations using non-experts? This research question explores the practical methods for creating effective analogies, leveraging the collective intelligence of crowd workers.

- RQ2: How can we systematically assess the quality of analogy-based explanations? This question focuses on developing and validating a structured framework for evaluating the conceptual quality of generated analogies.

- RQ3: How do analogies for concept-level explanations shape the understanding of an AI system among non-expert users? This question investigates the direct impact of analogy-based explanations on user comprehension.

- RQ4: How do analogy-based explanations affect user reliance on AI systems? This question probes the influence of analogies on how users calibrate their trust and decisions when collaborating with AI.

We hypothesize that analogy-based explanations, by leveraging familiar conceptual frameworks, will lead to a deeper understanding of AI reasoning, foster better calibrated reliance patterns, reduce the cognitive burden on users, and ultimately improve the efficacy of human-AI collaborative decision outcomes compared to traditional, non-analogical explanations. By meticulously examining the mechanisms through which analogies function within the realm of XAI, this research endeavors to make significant contributions to the evolving body of knowledge on human-centered AI design, thereby facilitating more effective, trustworthy, and intuitive partnerships between humans and AI systems. This manuscript represents an extended and more detailed exploration of previous work [43], incorporating new research questions, hypotheses, and an extensive empirical study on skin cancer detection, alongside synthesized guidelines for future XAI research.

METHODS

To systematically and rigorously evaluate the multifaceted impact of analogy-based explanations on laypeople's interactions within AI-assisted decision-making scenarios, a meticulously designed and controlled online user study was executed. This study employed a robust between-subjects experimental design, facilitating a direct comparison of participants' understanding, reliance behaviors, perceived cognitive load, and ultimate decision performance across distinct explanation conditions. The methodology was structured into two primary studies: Study I focused on the generation and expert evaluation of analogy-based explanations, while Study II investigated their effectiveness in a practical medical diagnosis task.

Study I: Analogy Generation and Evaluation

The objective of Study I was to develop a method for generating high-quality analogy-based explanations using non-expert crowd workers (addressing RQ1) and to establish a systematic framework for assessing their

quality (addressing RQ2).

Participants (Study I)

For the analogy generation phase, a total of 100 crowd workers were recruited from Prolific, a well-regarded crowdsourcing platform recognized for its diverse participant pool and commitment to data quality [2, 23]. Participants were fairly compensated for their contributions at a rate of £1.35, equivalent to a 9-minute task at an hourly wage of £9. Strict inclusion criteria were applied: participants had to be proficient English speakers, at least 18 years of age, and maintain an approval rating of at least 90% on the Prolific platform, indicating consistent high-quality work. To counteract common challenges associated with crowdsourcing, such as repeated concepts or task domain-specific biases, a pilot study was conducted (with 7 participants from Prolific). Insights from this pilot led to refinements in the main study design: to prevent participants from using concepts from the task domain itself or generating duplicate analogies, each participant was assigned to generate analogies for tasks from only one domain (either calorie level classification or scene classification) and forbidden from using specific "taboo phrases" (words present in task descriptions or labels). This resulted in 50 workers for the calorie task and 50 for the places task, collectively generating 600 analogy-based explanations.

AI System and Decision Task (for Analogy Generation)

To ensure that the analogy generation tasks were comprehensible and relevant for non-expert crowd workers, two distinct image classification tasks were selected: Calorie Level Classification (CLC) and Scene Classification (SC). These tasks were chosen for their interpretability by laypeople and the varying explicitness of relationships between concepts and labels within their domains.

- Calorie Level Classification (CLC): This task utilized a dataset provided by Buçinca et al. [11]. Participants were presented with an image (e.g., Figure 2(a) in the original PDF, showing food) along with bounding boxes highlighting relevant concepts (e.g., chocolate, ice cream). The task required predicting one of two labels: "high calorie level" (fat more than 30%) or "low calorie level" (otherwise). In this domain, the relationships between food concepts and calorie levels are often correlational rather than strictly causal.

- Scene Classification (SC): For this task, a subset of the Places dataset [109] was used. Images depicted various scenes (e.g., Figure 2(b) in the original PDF, showing a conference room). Participants had to classify scenes into one of six place labels: living room, bathroom, hospital room, conference room, bedroom, or dining room. In contrast to CLC, the concepts within this domain (e.g., furniture, objects) typically have more explicit and discernible relationships with the labels, often described by commonsense relations like "PartOf," "SignOf," or "FoundAt," which are also present in knowledge bases such as ConceptNet [91].

For both tasks, participants were instructed to explain the relevance of the given concepts to the predicted labels, formulating their explanations using everyday concepts and pre-defined templates.

Templates for Analogy-based Explanations

To guide crowd workers in structuring their analogies and associating concepts with model predictions, a set of six distinct templates was provided. These templates were categorized based on three different "relevance levels" that reflect how a machine learning model might interpret a concept's contribution:

- Positive Evidence: Concepts that strongly indicate a particular prediction.
  - Definite Sign Of: [Concept A] is definitely a sign of [Concept B]. This is like a [trunk] is a definitely sign of [an animal being an elephant]. (Example: "Mayonnaise is definitely a sign of high calorie food. This is like a trunk is a definitely sign of an animal being an elephant.")
  - Typically Associated With: [Concept A] is typically associated with [Concept B], while rarely associated with [Concept C]. This is like [printers] can typically be associated with [offices], but it's also possible to associate [printers] with [homes]. (Example: "Chocolate is typically associated with high calorie food, while rarely associated with low calorie food.")
- Inconclusive Evidence: Concepts that are present but do not definitively point to a specific prediction.
  - Insufficient: [Concept A] is not sufficient to indicate [Concept B], as both [Concept B] and [Concept C] may contain it. This is similar to how we can find [chair] in both [a living room] and [a bedroom], you can't determine which room it is by seeing a [chair]. (Example: "Bread is not sufficient to indicate high calorie, as both high calorie food and low calorie food may contain it.")
  - Irrelevant: [Concept A] is irrelevant to indicate [Concept B]. This is similar to to how [an arbitrary stone] is irrelevant for [recognising a continent]. (Example: "A plate is irrelevant to indicate high calorie food.")
- Negative Evidence: Concepts that suggest the absence of a prediction or contradict it.
  - Seldom Found: [Concept A] can seldom be found in [Concept B]. This is like [cats] can seldom be found in [water]. (Example: "Carrots are seldom found in high calorie food.")
  - Contradict With: [Concept A] contradicts with [Concept B]. This is similar to how one cannot find [water] in [electrical appliances]. (Example: "A vegetable salad contradicts with high calorie food.")

These templates, along with illustrative examples (Table 1 in the original PDF), guided participants to fill in

placeholders with everyday concepts from a source domain that differed from the task domain.

Task Selection and Hints for Analogy Generation

To ensure a balanced distribution of generated analogies across all six relevance categories, 12 specific tasks were manually selected: 6 from the CLC domain and 6 from the SC domain. A key challenge identified during pilot testing was that crowd workers, despite their non-expert status, found it difficult to generate a continuous stream of novel analogies after an initial few attempts. To address this "analogy fatigue," participants were provided with a curated list of "hint domains" through a clickable button in the user interface. This list encompassed common, everyday categories from which participants could draw inspiration: weather, animals and plants, place, transportation, food, art, education, sports, finance, clothes, electronics, games and toys, and health. This scaffolding aimed to stimulate creativity and reduce the cognitive burden of concept generation.

Analogy Generation Procedure

The analogy generation process for each crowd worker involved several structured steps:

1. Template Selection: Participants first selected one of the six provided templates that best described the perceived relevance level between a given concept and its associated model prediction for an image.

2. Hint Reference: They could refer to the provided example analogies and the list of everyday hint domains to inspire their concept choices.

3. Placeholder Filling: Based on the chosen template, participants were instructed to fill in the placeholders with one word or a short phrase (up to five words) as concepts. A crucial constraint was that these concepts must not belong to the original task domain (e.g., no places or furniture for the Scene Classification task), ensuring that the analogy indeed served as a bridge to a different, more familiar domain.

An example of the analogy generation interface (Figure 3 in the original PDF) showed a workflow where participants: (1) selected a template, (2) referenced examples/hints, and (3) filled in concepts.

Analogy Evaluation with Experts

Following the generation phase, a critical step was to qualitatively assess the quality of the generated analogies to address RQ2. This was done through an expert evaluation process.

- Experts (Study I Evaluation): Five external experts were purposefully sampled from the authors' institution [92]. These experts possessed at least a foundational understanding of machine learning and explainable AI, making them suitable for evaluating the nuances of explanation quality.

- Sample Selection: A subset of 294 generated analogy-based explanations was selected for evaluation, comprising analogies from 23 calorie task participants and 26 place task participants (approximately half of the generated analogies). To ensure consistency and measure inter-rater agreement, a 10% overlap (29 analogy-based explanations) was ensured across all experts. Each expert, on average, evaluated 82 distinct analogies, dedicating approximately 2.5 hours to this qualitative assessment.

- Qualitative Assessment Dimensions: Based on a systematic review of existing literature on analogy quality [8, 36, 37, 48, 89, 94], a structured set of nine dimensions was synthesized for the qualitative assessment of analogy-based explanations (Table 2 in the original PDF). These dimensions were categorized into "Analogical Properties" and "Utility":

  - Analogical Properties (evaluating the core relational mapping between source and target domains):

    - Structural Correspondence: "How well can you align the properties of the explanation concepts to the properties of the concepts in the target sentence?" (5-point Likert scale).

    - Relational Similarity: "How similar do you perceive the relationship between concepts in the explanation and the relationship between concepts in the target sentence?" (5-point Likert scale).

    - Transferability: "How well can the explanation be used in other contexts?" (5-point Likert scale).

    - Helpfulness: "How helpful is this explanation for you to understand the target sentence?" (5-point Likert scale). This corresponds to the "purpose" in Holyoak and Thagard's multiconstraint theory [48].

  - Utility (evaluating inherent qualities of the generated commonsense explanations themselves):

    - Familiarity: "How familiar are you with the concepts in the explanation?" (5-point Likert scale). This acknowledges that effective analogies require a familiar source domain [34, 94].

    - Simplicity: "Do you think the explanation is simple enough for others to understand?" (5-point Likert scale), reflecting ease of interpretation [94].

    - Misunderstanding: "Do you think this explanation lead to more than single interpretation?" ({Yes, No}). This addresses ambiguity.

    - Syntactic Correctness: "Whether the analogy sentence is syntactically correct?" ({Yes, No}).

    - Factual Correctness: "Whether it describes a fact about real world? Can we switch it to make it factual?" ({Yes w/o switch, Yes & switch, No}). This ensures the truthfulness of the analogy's premise.

- Annotation Rubrics: An iterative coding process [93] was used to develop detailed annotation rules to guide experts' assessments, ensuring consistency:

- Invalid analogies (those using concepts from the target domain) were skipped.
- For Factual Correctness, examples were provided (e.g., "The pink feather is definitely a sign of flamingo" could be corrected by switching "pink feather" and "flamingo").
- An analogy with potential misunderstanding was considered factually correct if at least one interpretation was true.
- Helpfulness and Transferability were automatically assigned '1' if the analogy was factually incorrect.

● Procedure (Study I Evaluation): Each expert received an annotation manual detailing the dimensions and rules. They spent approximately 10 minutes reviewing the manual and clarifying any ambiguities before independently assessing their assigned samples.

● Annotation Agreement: Krippendorff's α scores were calculated based on the 22 valid overlapping analogy explanations. While syntactic correctness showed relatively high agreement (0.64), other dimensions like Structural Correspondence (0.15), Relational Similarity (0.17), Familiarity (0.03), Helpfulness (0.14), Transferability (0.11), and Simplicity (0.14) showed lower agreement. This highlights the subjective nature of evaluating analogy quality, which can vary based on individual experience and interpretation [12]. An example illustrating disagreement among experts was provided (Table 3 in the original PDF), where "Lemon is seldom found in high calorie food. This is similar to how having hair is irrelevant for recognising a human" received divergent scores across dimensions like Structural Correspondence and Relational Similarity, due to varying personal interpretations and abstract thinking.

Study II: Effectiveness of Analogy-based Explanations in Medical Diagnosis

The second empirical study (Study II) aimed to move beyond conceptual quality and investigate the practical effectiveness of analogy-based explanations in a real-world, high-stakes human-AI decision-making scenario, addressing RQ3 and RQ4. All hypotheses and experimental setups for Study II were preregistered to ensure methodological rigor.

Hypotheses (Study II)

Building upon the findings of Study I and existing literature, the following hypotheses were formulated for empirical testing:

● H1: Using analogy-based explanations can help users better understand AI systems, compared to conventional concept-based explanations. This hypothesis posits that analogies, by simplifying complex AI concepts, will lead to a deeper user understanding.

● H2: Using analogy-based explanations can facilitate appropriate reliance on AI systems, compared to conventional concept-based explanations. This hypothesis suggests that improved understanding through analogies will translate into more calibrated trust and reliance behaviors.

● H3: Analogy-based explanations can reduce the perceived cognitive load of users in their decision making process. This hypothesis anticipates that analogies, by making information more accessible, will reduce the mental effort required for decision-making [82].

● H4: Providing analogy-based explanations on demand can improve users' efficiency in their decision making process. This hypothesis explores the efficiency benefits of user-controlled access to explanations.

Task (Study II)

A real-world medical diagnosis scenario—skin cancer detection based on skin lesions—was chosen as the testbed. This task was selected for several reasons:

1. Realism and Accountability: It represents a realistic human-AI collaboration context where humans retain final decision-making authority due to accountability concerns.

2. Cognitive Challenge for Laypeople: Medical concepts related to skin lesions are often challenging for non-experts to grasp, aligning with the study's motivation to provide commonsense analogy-based explanations.

3. Practical Need for AI Assistance: There is a significant need for AI support in medical diagnosis due to the increasing volume of images requiring analysis [56].

All task data were sourced from the HAM10000 dataset [96], a large collection of dermatoscopic images of common pigmented skin lesions. Participants were asked to classify images as depicting either 'malignant' or 'benign' skin lesions.

Medical Concepts

To aid participants, eight specific medical concepts relevant to skin cancer diagnosis were adopted from previous work [105]: Blue-Whitish Veil, Regular Dots & Globules, Irregular Dots & Globules, Regression Structures, Irregular Streaks, Regular Streaks, Atypical Pigment Network, and Typical Pigment Network. While these concepts initially contain descriptive terms like "Irregular" or "Atypical" that might hint at their correlation with malignant/benign labels, these hints were replaced with neutral abstractions ("type 1" and "type 2") to ensure that any learning effect stemmed from the explanations themselves, not from the concept names. An overview figure illustrating these eight concepts with example images (Figure 6 in the original PDF) was provided to participants to facilitate comprehension. A clickable button below the concept-level explanations allowed participants to access this overview figure on demand.

Selection of Tasks (Study II)

To ensure diversity and representative coverage of different medical concepts and AI performance scenarios, 14 tasks were carefully selected from the HAM10000 dataset, spanning seven fine-grained categories (Table 5 in the original PDF, which provides descriptive statistics of the dataset and AI performance per category). The selection process mirrored the performance of a post-hoc concept bottleneck model [105] on the HAM10000 validation set: 10 tasks with correct AI predictions and 4 tasks with incorrect AI predictions. This resulted in an overall AI system accuracy of 71.4% (10 correct out of 14 tasks) within the study.

Pilot Study (Study II)

A pilot study involving 20 participants from Prolific was conducted to assess the capabilities of non-expert crowd workers in this medical diagnosis task. Participants were compensated £2 (£8 per hour) for completing 14 trial tasks independently. After filtering out three outliers who spent less than 5 minutes on the tasks, the remaining 17 participants achieved an average accuracy of 59.2%, which was lower than the AI's 71.4% accuracy. This finding underscored the potential benefit of AI assistance in improving team performance, thereby justifying the inclusion of an AI system in the main study.

Experimental Setup

The main study employed a between-subjects design with five distinct experimental conditions, varying in the type and presentation of explanations provided alongside the AI's advice. Participants in all conditions received the AI's diagnosis (malignant/benign).

- Control: Participants received only the AI's prediction with no additional explanation.
- Concept: Participants received a concise, concept-based explanation from a post-hoc Concept Bottleneck Model [105], similar to the ExAID framework [70]. An example might be: "absence of Streaks - type 1: strong evidence."
- Concept-Imp: This condition provided more detailed information about the importance of each concept, representing the target domain of our analogy-based explanations. An example: "Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign." (See Table 6 in original PDF).
- Analogy: Participants received both the detailed concept-based explanation (as in Concept-Imp) AND an analogy-based explanation for each concept. For instance: "Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign. This is like how a beak is a definite sign of a bird." (See Table 6 in original PDF).
- Analogy-OD (On Demand): This condition initially displayed the same explanations as Concept-Imp. An analogy was provided on demand when the user explicitly requested further clarification by clicking a "Clarify" button. (See Table 6 in original PDF).

Explanation Generation (for Study II)

The AI system utilized in Study II was based on a post-hoc concept bottleneck model [105], trained following its official implementation. This model is known for providing concept-based explanations aligned with medical knowledge. The process for generating explanations involved:

1. Concept Activation Vectors: The model learned concept activation vectors for skin lesions based on concept banks from the Derm7pt dataset [55].
2. Linear Classifier: A linear classifier was trained to make binary predictions (malignant/benign).
3. Contribution-Based Explanations: For each image, concept-level explanations were generated based on the contribution ($s_i = w_i * c_i$) of each concept (c_i) to the final prediction, where w_i is the linear layer weight.
4. Heuristic Simplification (Concept condition): Two thresholds were set ($\epsilon_1 = 0.5, \epsilon_2 = 0.1$) to classify evidence strength as "strong," "moderate," or "ignore." Positive s_i indicated a tendency towards malignant, negative towards benign.
5. Target Domain Generation (Concept-Imp condition): These explanations followed the templates from Study I, clarifying the contribution of concepts, including cases where the absence of a concept was indicative. For clarity, double negative expressions were avoided.
6. Analogy Generation (Analogy & Analogy-OD conditions): To ensure high-quality analogies for Study II, a two-stage filtering process was applied based on the results from Study I's expert evaluation:
   - Stage 1: Only analogies that were syntactically correct, factually correct, and easy to understand (Simplicity score > 3) were selected.
   - Stage 2: The remaining analogies were manually curated, resulting in 37 valid analogies distributed across the different template types: 11 for "Definite Sign Of," 9 for "Typically Associated With," 9 for "Seldom Found At," and 8 for "Contradict With." These valid candidates were then randomly sampled based on the concept's contribution and mapped to appropriate templates.

Measures and Variables (Study II)

A comprehensive set of variables was employed to capture the various facets of human-AI decision-making (Table 7 in original PDF summarizes these variables):

- Dependent Variables:
  - Learning Effect (H1): Assessed by calculating F1 measures (weighted average of F1 for malignant and

benign concepts) based on participants' ability to identify concepts positively associated with malignant/benign labels in a post-task questionnaire.

○ Performance (Overall Accuracy): Binary accuracy of participants' final diagnoses across all 14 cases.

○ Appropriate Reliance (H2): Quantified using several metrics:

■ Agreement Fraction: Proportion of cases where human and AI decisions align.

■ Switch Fraction: Proportion of initial human decisions changed after seeing AI advice.

■ Relative Positive AI Reliance (RAIR): Measures appropriate adoption of AI advice [86].

■ Relative Positive Self-Reliance (RSR): Measures appropriate insistence on one's own decision [86].

■ Accuracy-wid (Accuracy with initial disagreement): Accuracy when the participant initially disagreed with the AI, indicating how well they resolve disagreements.

○ Trust: Measured using adapted subscales from the Trust in Automation (TiA) questionnaire [58]: Reliability/Competence (TiA-R/C), Understanding/Predictability (TiA-U/P), Intention of Developers (TiA-IoD), and overall Trust in Automation (TiA-Trust). All on a 5-point Likert scale.

○ Cognitive Load (H3): Assessed using the NASA-TLX questionnaire [18], including dimensions like Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. All on a -7 to 7 scale.

○ Efficiency (H4): Measured as the average time (in seconds) participants spent on each decision task.

● Covariates: These variables were collected to account for potential confounding factors:

○ Affinity for Technology Interaction (ATI): Measured using the ATI scale [32] (6-point Likert scale).

○ Propensity to Trust: From the TiA questionnaire [58] (5-point Likert scale).

○ Familiarity (with AI): From the TiA questionnaire [58] (5-point Likert scale).

○ General Medical Expertise: Self-reported on a 5-point Likert scale ("To what extent are you knowledgeable about medical diagnosis?").

○ Skin Cancer Expertise: Self-reported on a 5-point Likert scale ("Do you have any experience or knowledge about skin cancer?").

● Other Variables:

○ Helpfulness of Explanation/Analogy: Self-reported on a 5-point Likert scale, with open-text fields for reasons.

○ User Experience (with skin lesions): "Have you ever had this or seen it on others?" ({Yes, No}).

○ Confidence: Self-reported for each decision on a 5-point Likert scale.

Participants (Study II)

● Sample Size Estimation: A power analysis using G*Power [30] determined a required sample size of 265 participants to detect a moderate effect size (f = 0.25) with 80% power at a Bonferroni-adjusted significance threshold of α=0.0125 (0.05/4 hypotheses), considering five experimental conditions.

● Recruitment and Compensation: 486 participants were initially recruited from Prolific [2, 23] to allow for potential exclusions. Each participant received £2 (hourly wage of £8) for the estimated 15-minute task, plus an additional £0.1 bonus for every correct decision in the 14 trial cases. This monetary incentive was used to encourage genuine effort and appropriate system reliance [64].

● Filter Criteria: Participants were excluded if they failed any of the attention checks or had missing responses. The final analytical sample comprised 280 participants. The average age was 37 (SD = 13.0), with a balanced gender distribution (51.4% female, 48.6% male).

Procedure (Study II)

The study procedure (illustrated in Figure 7 in the original PDF) involved several distinct phases:

1. Instructions and Consent: All participants first reviewed basic instructions and provided informed consent.

2. Pre-task Questionnaire: Participants completed a questionnaire to assess their ATI, general medical expertise, and specific skin cancer expertise.

3. Onboarding and Medical Concepts Overview: To familiarize participants with the skin cancer detection task, two examples of benign and malignant skin lesions were presented. Following this, all participants, except those in the Control condition, received an overview of the eight medical concepts relevant to the task (Figure 6 in the original PDF).

4. Trial Tasks (Two-Stage Decision Making): Participants proceeded to complete 14 trial tasks, each involving a two-stage decision-making process [41, 20]:

○ Stage 1 (Initial Decision): Participants viewed a lesion image (Figure 8 in original PDF) and made an initial diagnosis (malignant/benign) without any AI advice or explanation.

○ Stage 2 (AI Advice and Explanation): After their initial decision, participants were shown the AI's prediction and, depending on their assigned condition, the corresponding explanation (no explanation, concept-based, concept-important, analogy-based, or on-demand analogy). They then had the opportunity to revise their

initial decision (Figure 9 in original PDF). Confidence levels were also collected at this stage.

5. Post-task Questionnaires: Upon completing all 14 tasks, participants filled out questionnaires to assess their cognitive load (NASA-TLX), trust in the AI system (TiA), and provided open-text feedback on their decision-making criteria. Participants in conditions with explanations also responded to questions about the perceived helpfulness of explanations and provided open-text reasons. Those in Analogy and Analogy-OD conditions additionally commented on the analogies.

6. Attention Checks: Three attention check questions were strategically placed within the pre-task questionnaire, task phase, and post-task questionnaire to ensure data reliability [73, 33].

Ethical Considerations

The entire study protocol was rigorously reviewed and approved by the Institutional Review Board of the researchers' institution. All participants provided explicit informed consent prior to engaging in the study, and confidentiality and anonymity were strictly maintained throughout the data collection and analysis processes. Participants were explicitly informed of their right to withdraw from the study at any point without penalty. All data were securely collected, stored, and managed in full compliance with relevant privacy regulations and ethical guidelines.

Results

The comprehensive analysis of the collected data yielded significant findings regarding the effectiveness of analogy-based explanations in the context of AI-assisted decision-making for laypeople. The results address the hypotheses (H1-H4) and provide valuable insights into user behavior and perceptions.

Descriptive Statistics

A total of 280 participants who successfully passed all attention checks were included in the final analysis. These participants were distributed relatively evenly across the five experimental conditions: 55 in Control, 55 in Concept, 55 in Concept-Imp, 53 in Analogy, and 62 in Analogy-OD.

The distribution of covariates indicated the general characteristics of the participant pool:

- Affinity for Technology Interaction (ATI): Mean (M) = 3.87, Standard Deviation (SD) = 0.87 (on a 6-point Likert scale, 1: low, 6: high).
- Medical Diagnosis Expertise: M = 1.47, SD = 0.81 (on a 5-point Likert scale, 1: no expertise, 5: extensive expertise).
- Skin Cancer Expertise: M = 1.59, SD = 0.81 (on a 5-point Likert scale, 1: no expertise, 5: extensive expertise).
- TiA-Propensity to Trust: M = 2.76, SD = 0.57 (on a 5-point Likert scale, 1: tend to distrust, 5: tend to trust).
- TiA-Familiarity (with AI): M = 2.31, SD = 1.05 (on a 5-point Likert scale, 1: unfamiliar, 5: familiar).

These statistics confirm that most participants had low self-reported medical and skin cancer expertise, aligning with the "layperson" target audience.

Overall performance across all conditions showed an average accuracy of 63.3% (SD = 0.11), which was lower than the AI system's accuracy of 71.4%. The average agreement fraction (proportion of human-AI aligned decisions) was 0.79 (SD = 0.16), and the average switch fraction (proportion of initial human decisions changed after seeing AI advice) was 0.57 (SD = 0.30). These figures suggest that participants did not blindly defer to AI advice and were willing to reconsider their initial decisions, indicating active engagement. Given that most dependent variables were not normally distributed, non-parametric statistical tests were primarily used for hypothesis verification.

Performance Per Task

A detailed breakdown of accuracy and confidence per task (Table 8 in the original PDF) across the 14 skin lesion cases revealed interesting patterns. Generally, participants' accuracy increased after exposure to correct AI advice and decreased when exposed to incorrect AI advice, underscoring the AI's influence. Confidence also tended to increase after receiving AI advice, with one notable exception (task ISIC-0032557) where confidence decreased despite an initially high confidence level, possibly indicating an "illusion of competence" among participants who significantly overestimated their initial accuracy on this task (achieving only 4.3% accuracy while maintaining high initial confidence). The "Experience ratio" (proportion of participants who reported seeing similar skin lesions) was consistently low across all tasks (ranging from 0.04 to 0.14), further confirming the layperson status of the participants.

Helpfulness of Explanations and Analogies

Participants' perceived helpfulness of the provided explanations and analogies was assessed via post-task questionnaires (Figure 10 in the original PDF). Overall, 61.8% of participants who received concept-based explanations reported a positive attitude towards their helpfulness (either "somewhat helpful" or "helpful"). For participants in the Analogy and Analogy-OD conditions, 39.1% found the analogies to be helpful to some extent. This suggests that while concept-based explanations were generally well-received, the utility of analogies was perceived less universally, indicating a mixed reaction.

H1: The Impact of Analogy-based Explanations on Learning Effect

H1 hypothesized that analogy-based explanations would enhance users' understanding of AI systems compared to conventional concept-based explanations. To test this, the

weighted average F1 score (F1avg=85F1malignant+83F1benign) was calculated based on participants' ability to correctly identify concepts correlated with malignant and benign labels. A Kruskal-Wallis H-test showed no significant difference across the explanation conditions ($H(279) = 1.79$, $p = 0.616$). The mean F1 scores were: Concept ($M = 0.55$, $SD = 0.20$), Concept-Imp ($M = 0.58$, $SD = 0.19$), Analogy ($M = 0.56$, $SD = 0.21$), and Analogy-OD ($M = 0.52$, $SD = 0.22$). This lack of statistical significance indicates that the study did not find empirical support for H1; analogy-based explanations did not lead to a demonstrably improved learning effect or a deeper conceptual understanding of the AI system's inner workings.

H2: The Impact of Analogy-based Explanations on Appropriate Reliance

H2 proposed that analogy-based explanations would facilitate appropriate reliance on AI systems. Kruskal-Wallis H-tests (Table 9 in the original PDF) revealed significant differences in appropriate reliance measures across the conditions. Specifically, significant differences were found for Agreement Fraction ($H = 11.03$, $p = 0.026$), Accuracy-wid ($H = 15.81$, $p = 0.003$), and RAIR ($H = 12.77$, $p = 0.012$).

Post-hoc Mann-Whitney tests, with a Bonferroni-adjusted alpha level of 0.0125, indicated several key findings:

- Concept-Imp vs. Others: Participants in the Concept-Imp condition demonstrated significantly higher Accuracy-wid and RAIR compared to participants in the Control, Concept, and Analogy conditions. This suggests that simply providing more detailed concept-level explanations (the target domain of the analogies) was effective in promoting appropriate reliance, particularly by mitigating "under-reliance" (i.e., users inappropriately overriding correct AI advice).

- Potential for Over-reliance: The relatively low RSR in the Concept-Imp condition, when compared to other conditions, suggests that while it reduced under-reliance, it might have also triggered a tendency towards "over-reliance" (blindly following AI advice even when it's wrong).

- Analogies' Limited Impact: Surprisingly, the study found no statistically significant evidence that analogy-based explanations (in Analogy or Analogy-OD conditions) had the expected effect in facilitating appropriate reliance. While Analogy-OD showed a non-significant trend towards better appropriate reliance compared to Analogy, H2 was not empirically supported by the quantitative results.

H3: The Impact of Analogy-based Explanations on Cognitive Load

H3 hypothesized that analogy-based explanations would reduce the perceived cognitive load. A one-way ANOVA (Table 10 in the original PDF) revealed that participants who received any form of explanation (Concept, Concept-Imp, Analogy, Analogy-OD) reported a higher perceived cognitive load compared to the Control group. Significant differences were found for average cognitive load ($F = 5.81$, $p = 0.000$) and mental demand ($F = 7.01$, $p = 0.000$). Post-hoc Tukey HSD tests (alpha = 0.0125) confirmed that Control < Concept, Analogy, Concept-Imp, Analogy-OD for both average cognitive load and mental demand. This contradicts H3, indicating that explanations, including analogy-based ones, increased rather than decreased the perceived cognitive burden on users.

H4: The Impact of Analogy-based Explanations on Decision Making Efficiency

H4 posited that providing analogies on demand would improve users' efficiency in decision-making. A Kruskal-Wallis H-test on task completion time showed a significant difference across conditions ($H(279) = 23.73$, $p = 0.000$). Post-hoc Mann-Whitney tests revealed that participants who received any explanations spent significantly more time making decisions compared to the Control group (Control < Concept, Analogy, Concept-Imp, Analogy-OD). The average total time spent (in seconds) was: Control ($M = 462$, $SD = 309$), Concept ($M = 548$, $SD = 210$), Concept-Imp ($M = 575$, $SD = 209$), Analogy ($M = 574$, $SD = 242$), and Analogy-OD ($M = 658$, $SD = 341$). A more fine-grained analysis of time per correct/wrong decision (Table 11 in the original PDF) yielded consistent results. Thus, H4 was not supported; explanations, even on-demand analogies, increased decision-making time rather than improving efficiency.

Exploratory Analysis

Beyond the direct hypothesis testing, several exploratory analyses provided deeper qualitative and quantitative insights into the factors influencing human-AI interaction.

The Impact of First Impression

Prior research emphasizes the role of initial interactions in shaping user trust and reliance [76, 77, 95]. To investigate this, participants were grouped based on AI accuracy in the first five tasks ("Good First Impression" for no or one wrong AI advice, "Bad First Impression" for more errors). Kruskal-Wallis H-tests found no significant difference in participant performance or reliance behaviors based on this grouping, suggesting that the initial impression of the AI system did not have a measurable impact within the confines of this study.

Analysis of Trust and Covariates

An ANCOVA analysis across experimental conditions showed no significant difference in perceived trust in the AI system (TiA subscales) across the explanation conditions. However, examining covariates revealed important correlations (Table 12 in the original PDF).

- Propensity to Trust: TiA-Propensity to Trust (a measure of general disposition to trust) positively correlated with all trust measures (TiA-R/C: r=0.650,

p<0.000; TiA-U/P: r=0.344, p<0.000; TiA-IoD: r=0.283, p<0.000; TiA-Trust: r=0.677, p<0.000). Furthermore, it showed significant positive correlations with Agreement Fraction (r=0.227, p<0.000), Switch Fraction (r=0.220, p<0.000), and RAIR (r=0.183, p=0.002), but a negative correlation with RSR (r=-0.216, p<0.000). This suggests that individuals with a higher inherent propensity to trust were more likely to align with AI advice, which could lead to both appropriate reliance (reducing under-reliance) but also potentially over-reliance (as indicated by the negative correlation with RSR).

- Other Covariates: TiA-Familiarity with AI showed positive correlations with TiA-R/C (r=0.232, p<0.000) and TiA-Trust (r=0.286, p<0.000), and ATI (Affinity for Technology Interaction) correlated positively with TiA-Trust (r=0.149, p=0.012). However, ANCOVA analysis determined their overall impact on trust was not significant. Medical diagnosis expertise did not show strong correlations, while skin cancer expertise (self-reported, mostly zero for laypeople) showed a negative correlation with Switch Fraction (r=-0.175, p=0.003).

Impact of User Opinions towards Explanations and Analogies

- Helpfulness of Explanations: Perceived helpfulness of explanations (for Concept, Concept-Imp, Analogy, Analogy-OD conditions) positively correlated with user trust in the AI system (TiA-R/C: r=0.400, p<0.000; TiA-U/P: r=0.397, p<0.000; TiA-IoD: r=0.249, p<0.000; TiA-Trust: r=0.407, p<0.000). However, no significant correlation was found between perceived helpfulness of explanations and reliance-based dependent variables.

- Helpfulness of Analogies: Similarly, for Analogy and Analogy-OD conditions, perceived helpfulness of analogies positively correlated with user trust (TiA-R/C: r=0.303, p=0.001; TiA-U/P: r=0.290, p=0.002; TiA-IoD: r=0.368, p=0.000; TiA-Trust: r=0.297, p=0.001), but again, no significant correlation was found with reliance-based variables.

Notably, 24.5% of Analogy condition participants found analogies helpful, compared to 51.6% in Analogy-OD. This disparity might partially explain why the Analogy condition exhibited slightly lower (though not significant) appropriate reliance metrics, suggesting that perceived unhelpfulness could negatively impact trust and lead to under-reliance. Conversely, the Concept-Imp condition showed very low RSR, indicative of over-reliance.

Qualitative Analysis of Feedback

Open-ended feedback from participants provided rich qualitative insights into their decision-making processes and perceptions of explanations.

- Decision Criteria: Thematic analysis of responses to "Please describe how you made your decisions" (Table 13 in the original PDF) identified five main topics:
  - Picture (91 mentions): Directly relying on visual assessment of the skin lesion images.
  - Examples (77 mentions): Referring back to the provided benign/malignant examples.
  - Explanations (77 mentions): Using the concept-level explanations to understand and refine decisions, often trusting AI more than self.
  - Intuition (68 mentions): Making decisions based on "instinct" or "gut feeling," often finding alignment with AI.
  - AI advice (62 mentions): Directly using AI recommendations, especially when confused.
- Reasons for Helpfulness/Unhelpfulness of Explanations (Table 14 in original PDF):
  - Helpful: Explanations were perceived as helpful because they "enrich the context of decision making" (32.4%), "help improve the understanding of the AI system" (18.7%), or "help confirm or validate their decision" (7.2%).
  - Unhelpful: Reasons for unhelpfulness included participants lacking knowledge to interpret (41.9%), failing to understand (16.3%), or finding explanations difficult to apply (11.6%).
- Reasons for Unhelpfulness of Analogies: Specific feedback on analogies included: "failed to connect the source domain with the target domain" (22.9%), "do not make sense" (18.6%), "concepts are not relevant" (14.3%), "failed to understand the analogies" (12.9%), or "not necessary" (10%).

Participants' comments also revealed conflicting attitudes towards analogies. Some found them "useful and helpful for getting the point across to laymen," while others found them distracting or irrelevant ("I don't get the relevance of using analogies to explain medical concepts. I also don't think they were explaining the concepts. It was essentially saying water is wet...").

Insights from Users to Improve Analogy Effectiveness

Based on user feedback, three potential directions emerged for improving analogy-based explanations:

1. Enhancing Source-Target Relation: Analogies need stronger, clearer connections between the familiar source domain and the complex target domain to ensure immediate understanding.

2. Domain Relevance: Analogies should ideally be drawn from domains that resonate with the context or general understanding of the specific task (e.g., medical analogies for medical tasks) to enhance plausibility and reduce cognitive dissonance.

3. Selective/On-Demand Provision: When the primary explanation is already clear, forced analogies can be perceived as unnecessary or even condescending, leading to annoyance and reduced trust. Providing analogies on demand empowers users to seek clarification

only when needed, potentially improving user experience and efficacy.

DISCUSSION

This comprehensive investigation provides compelling evidence that while analogy-based explanations hold significant intuitive appeal and show promise in certain contexts, their immediate effectiveness in enhancing laypeople's understanding and appropriate reliance on AI systems is more nuanced than initially hypothesized. Our findings demonstrate that compared to traditional feature-based explanations or no explanations, analogies, when meticulously crafted, can indeed improve some aspects of comprehension and influence reliance patterns, but they also present challenges related to cognitive load and decision efficiency. The summary of key findings from both Study I and Study II is presented in Table 15 (in the original PDF).

Key Findings and Implications

The insights gleaned from this research offer several critical implications for the design and deployment of human-centered XAI:

Subjectivity and Quality of Analogies

Study I highlighted the inherent subjectivity in evaluating the qualitative dimensions of analogies. The low Krippendorff's α scores for most qualitative dimensions (e.g., Structural Correspondence, Relational Similarity, Familiarity, Helpfulness, Transferability, Simplicity) among experts underscore that "quality" is not universally perceived. This disagreement stems from diverse personal experiences and interpretations of commonsense facts embedded within analogies. While this might initially appear as noise, prior work suggests that inter-rater disagreement can also be a valuable signal [3]. It reveals the ambiguity and vagueness of certain analogy-based explanations, pointing to areas for refinement. When evaluators diverge, involving additional crowd workers or incorporating iterative feedback loops could help in improving the clarity and universal appeal of analogies [52, 85].

Furthermore, the comparison between analogies generated from the calorie task (CLC) and the scene classification task (SC) revealed that superior quality on a single dimension (e.g., Relational Similarity for SC task) does not automatically translate to higher perceived helpfulness. However, if an explainer deems an analogy to be poor in relational similarity, they are more likely to find it unhelpful. This suggests a complex interplay between individual dimensions and overall perceived utility, further complicated by user-specific factors like abstract thinking, personal interpretation, and general attitudes toward explanations. This emphasizes the necessity for future research to delve deeper into the intricate relationships between user characteristics, qualitative dimensions, and the ultimate helpfulness of analogy-based explanations.

The observed expert disagreement also subtly challenges the assumption that commonsense knowledge, while seemingly universal, is uniformly accepted and understood by all humans [51]. This finding aligns with the "one-size-fits-all" problem in XAI, where a single explanation method rarely satisfies all user needs [89, 65]. Consequently, future XAI designs should consider tailoring commonsense explanations to align with the explainee's existing beliefs and experiences to maximize the effectiveness of analogical inference. This strongly advocates for the integration of personalization into the generation and delivery of commonsense explanations.

Automated Analogy Generation and Evaluation

Study I revealed practical challenges in analogy generation: approximately one-third of generated analogies were not factually correct, and workers struggled to consistently produce analogies with high Structural Correspondence and Relational Similarity. This underscores the need for sophisticated strategies to support the creation of effective analogies. A promising direction is the development of machine-in-the-loop crowdsourcing tasks. By integrating relational knowledge bases (like ConceptNet [91] or HowNet [21]) and machine learning methods, automated tools could pre-select or suggest high-quality candidate concepts that share structural similarities with the target domain. For instance, if a specific relationship (e.g., "is a sign of") is identified in the AI's reasoning, a system could query a knowledge base for everyday facts exhibiting that same relationship, providing crowd workers with more appropriate and structurally sound analogies, thereby reducing their cognitive load and improving output quality [97, 16].

Moreover, the high cost of continuous human expert evaluation for quality control necessitates the exploration of (semi-)automatic assessment methods for analogy quality. While dimensions like Syntactic Correctness could be automatically verified using grammar tools, and Simplicity/Misunderstanding assessed against curated lists of common/ambiguous concepts, more subjective dimensions remain challenging. However, emerging concepts like "jury learning" [39] propose using machine learning models to perform pseudo-human value judgments, offering a potential pathway for large-scale, automated quality assessment that could account for the inherent subjectivity of certain dimensions.

The Role of Human Intuition

In Study II, a significant number of participants explicitly reported relying on their "intuition" to make final decisions. This highlights the critical, often underappreciated, role of human intuition in shaping user understanding and reliance behaviors within human-AI decision-making contexts [14, 13]. Our findings suggest that human intuition can, in some cases, facilitate complementary collaboration with AI systems.

However, intuition can also introduce biases. While the

overall Agreement Fraction was high (around 0.80), the relatively low RSR in most conditions indicated a tendency for over-reliance. This means that even when the AI advice was incorrect and participants initially disagreed, they often deferred to the AI's judgment instead of trusting their correct initial decision. This phenomenon of "automation bias" [98] is a known pitfall of XAI interventions [7, 102]. Such over-reliance can be linked to "confirmation bias" (seeking information that confirms one's existing beliefs) and the "illusion of explanatory depth" [9]. Conversely, the Control, Concept, and Analogy conditions exhibited clear "under-reliance" compared to Concept-Imp. This under-reliance might be partially explained by the Dunning-Kruger effect [59], where users overestimate their own competence, leading to an inappropriate dismissal of AI advice, as explored in recent work [46]. The strong positive correlation between perceived helpfulness of explanations/analogies and subjective trust in the AI system further suggests that when users found analogies unhelpful, their trust was negatively impacted, potentially contributing to under-reliance and suboptimal team performance. In complex AI systems, a complete understanding is often impractical, and trust then becomes the guiding factor for reliance [64]. Uncalibrated trust, stemming from perceived unhelpfulness, can lead to inappropriate reliance.

The Role of Plausibility

The empirical results of Study II indicate that while the target domain of analogy-based explanations was generally perceived as clear, additional analogies were not always helpful, particularly when participants struggled to connect them meaningfully to the target domain. This can be partly attributed to the concept of plausibility. Users implicitly assume that "plausible explanations typically imply correct decisions, and vice versa" [54]. If participants found the analogies implausible or irrelevant to the medical context, they might have perceived the AI's advice as less trustworthy, leading to reduced reliance and potentially suboptimal team performance. The Analogy condition, which showed worse RAIR than Concept-Imp, might exemplify this, as more participants in the Analogy-OD condition (who received analogies on demand) found them plausible (51.6% vs. 24.5% in Analogy). The higher Switch Fraction, Accuracy-wid, RAIR, and RSR in Analogy-OD suggest that providing analogies on demand can be a superior design choice, allowing users to opt-in only when they perceive a need for deeper understanding, thus potentially increasing the perceived plausibility and overall effectiveness of the explanations. When analogies are not used appropriately or are poorly designed, both under-reliance and over-reliance can be triggered due to a perceived lack of plausibility.

Caveats and Limitations

Despite the comprehensive nature of this research, several caveats and limitations warrant acknowledgment:

Bias in Templates

The use of six pre-defined templates for analogy generation, while facilitating the crowd-sourcing process, may have introduced biases [49, 24]. These templates inherently favor specific types of relationships, potentially limiting the creativity and diversity of analogies generated by participants. Although providing hint domains mitigated some of this limitation, the structured nature of templates might still constrain the exploration of more complex or nuanced analogical mappings.

Restricted Usage

Analogy-based explanations may not be a universal solution suitable for all application scenarios. Our study suggests specific contexts where their utility might be diminished or counterproductive:

1. Simple Tasks: When the original task is inherently simple and involves only everyday concepts already familiar to users, introducing analogies can introduce unnecessary cognitive load and create confusion rather than clarity.

2. Implicit Relationships: In domains where explicit properties and clear relationships between concepts and labels are scarce (e.g., the Calorie Level Classification task in Study I), generating effective analogies with high structural correspondence and relational similarity becomes exceptionally challenging, limiting their effectiveness for laypeople.

Cascading Effects

As analogy-based explanations are built upon underlying concept-level explanations, they are susceptible to "cascading effects." If the foundational concept-level explanations do not accurately reflect the AI system's internal state or are inherently misleading, the analogy-based explanations, even if well-crafted, will also fail to convey truthful information. Furthermore, given their familiar and often persuasive nature, effective analogy-based explanations derived from misleading concept-level explanations could potentially amplify the negative impact on user trust and decision-making, leading to greater harm.

Potential Human Biases

The crowdsourcing methodology, while offering scalability, introduces potential for human biases that can affect experiment outcomes [24]. Several biases were identified in our study:

- Overconfidence/Optimism Bias (Dunning-Kruger effect): For specific tasks (e.g., ISIC-0032557), participants exhibited high initial confidence despite very low accuracy, suggesting an illusion of competence [59, 46]. This overestimation of their own capabilities could lead to under-reliance on potentially accurate AI advice.
- Confirmation Bias: Some participants explicitly

stated that explanations helped "confirm and validate" their initial decisions, suggesting a tendency to seek information that aligns with their pre-existing judgments rather than objectively evaluating new information [9].

- Information Overload: The provision of 4-7 concept-level/analogy-based explanations per task sometimes led to self-reported information overload, which can negatively impact user trust and reliance.

- Self-interest Bias: Monetary incentives, while encouraging effort, might also lead some crowd workers to prioritize speed over thoroughness, potentially reducing the diligent examination of explanations.

Threats to Generalizability

The generalizability of our findings requires careful consideration:

- Task Complexity and Stakes: Study I used relatively simple, low-stakes image classification tasks for analogy generation. While Study II moved to a more realistic, high-stakes medical diagnosis scenario, it is unknown how these findings translate to even more complex AI applications or critical decision-making contexts (e.g., financial, legal). The effectiveness of analogies might vary significantly depending on the domain and the level of abstraction required.

- Analogy Transferability: Although the generated analogies were deemed highly transferable in Study I, their actual effectiveness when applied to a new, complex task like medical diagnosis was limited. Future work should investigate how to generate analogies that are specifically effective for highly specialized and high-stakes tasks, potentially requiring input from domain experts in the analogy generation process.

- Role of Human Intuition: The dominant role of human intuition in the skin cancer detection task might mean that findings related to reliance do not generalize to tasks where intuition is less prominent.

- Complexity of Explanations: This study focused on explaining the relevance level between concepts and predictions. Analogies can be used to convey more complex structural correspondence and relational similarities. Our findings might not directly extend to scenarios involving a greater number of concepts or more intricate relational structures within the AI's reasoning.

CONCLUSIONS AND FUTURE WORK

In this paper, we embarked on a comprehensive journey to explore the potential of elucidating concept-level AI explanations through analogical inference, leveraging commonsense knowledge to foster more meaningful collaborations between AI systems and non-expert human users. Our initial endeavor (RQ1) involved designing a novel, template-based analogy generation method, which we instantiated by engaging crowd workers across two distinct image classification tasks: calorie level classification and scene classification. To ensure the quality of these generated explanations (RQ2), we synthesized and applied a structured set of qualitative dimensions. An expert-led evaluation confirmed that our proposed method, despite the involvement of non-expert workers, could indeed yield high-quality analogy-based explanations.

Subsequently, to thoroughly investigate how analogy-based explanations influence user understanding and reliance on AI systems (RQ3 and RQ4), we conducted a rigorous follow-up empirical study focused on a skin cancer detection task. The results from this second study yielded several nuanced findings: (1) it reinforced that a lack of domain expertise significantly impedes user understanding of raw concept-level explanations; (2) while improved concept-level explanations (the target domain of our analogies) were effective in promoting appropriate reliance by mitigating under-reliance, they also demonstrated a propensity to trigger over-reliance; (3) the strategy of providing analogies on demand emerged as a potentially promising design approach for their adoption; (4) however, our findings underscore that analogy-based explanations must be meticulously designed and judiciously employed to effectively clarify concept-level explanations. The experimental results provided limited quantitative support for the hypotheses that analogy-based explanations would universally facilitate deeper user understanding of the AI system or consistently foster more appropriate reliance.

Nevertheless, we cannot discount the significant qualitative evidence that highlights the substantial potential of analogy-based explanations in assisting laypeople in effective decision-making with AI. Crucially, compared to conventional concept-level explanations, the integration of analogies did not incur a statistically significant delay in decision-making efficiency nor impose a notably higher cognitive load on users. Our findings collectively suggest that the paramount challenge lies not in the mere presence of analogies, but in the ability to consistently generate high-quality, contextually relevant analogies and, critically, in the potential for personalized delivery. Based on an in-depth qualitative analysis of participants' feedback and observed user reliance patterns, we have synthesized a set of actionable guidelines. These guidelines are crucial for informing future research and development efforts aimed at generating truly effective analogy-based explanations and ensuring their appropriate and beneficial integration into human-AI collaborative decision-making frameworks.

Looking forward, our immediate future work will focus on addressing the scalability and efficiency challenges identified in Study I. Given that both the generation and expert evaluation of high-quality analogy-based explanations are labor-intensive and time-consuming, we intend to explore the integration of machine learning algorithms and external knowledge bases to automate these tasks. This automation aims to enhance the process's

scalability and efficiency without compromising quality. Furthermore, the findings from Study II, particularly the limited quantitative support for appropriate reliance despite qualitative indications of potential, necessitate continued empirical research. This includes further exploration into the optimal characteristics of analogies (e.g., level of abstraction, domain specificity) and the conditions under which they are most effective. Finally, the observed variability in understanding of commonsense explanations based on recipient user experience strongly points towards the critical need for deeper investigation into the personalization of commonsense explanations, ensuring they resonate optimally with individual users' knowledge and cognitive styles. This continued research will contribute to building more intuitive, trustworthy, and effective AI systems for all users.

REFERENCES

1. Abdul, A., von der Weth, C., Kankanhalli, M., & Lim, B. Y. (2020). Cogam: measuring and moderating cognitive load in machine learning model explanations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14.
2. Adams, T. L., Li, Y., & Liu, H. (2020). A replication of beyond the turk: Alternative platforms for crowdsourcing behavioral research–sometimes preferable to student groups. AIS Transactions on Replication Research, 6(1), 15.
3. Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine, 36(1), 15–24.
4. Arrieta, A. B., D´ıaz-Rodr´ıguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garc´ıa, S., Gil-L´opez, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion, 58, 82–115.
5. Balayn, A., He, G., Hu, A., Yang, J., & Gadiraju, U. (2022a). Ready player one! eliciting diverse knowledge using A configurable game. In Laforest, F., Troncy, R., Simperl, E., Agarwal, D., Gionis, A., Herman, I., & M´edini, L. (Eds.), WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022, pp. 1709–1719. ACM.
6. Balayn, A., Rikalo, N., Lofi, C., Yang, J., & Bozzon, A. (2022b). How can explainability methods be used to support bug identification in computer vision models?. In CHI Conference on Human Factors in Computing Systems, pp. 1–16.
7. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16.
8. Bartha, P. (2022). Analogy and Analogical Reasoning. In Zalta, E. N. (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2022 edition). Metaphysics Research Lab, Stanford University.
9. Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect xai-assisted decision-making: A systematic review. In Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society, pp. 78–91.
10. Bounhas, M., Pirlot, M., Prade, H., & Sobrie, O. (2019). Comparison of analogy-based methods for predicting preferences. In Amor, N. B., Quost, B., & Theobald, M. (Eds.), Scalable Uncertainty Management - 13th International Conference, SUM 2019, Compi`egne, France, December 16-18, 2019, Proceedings, Vol. 11940 of Lecture Notes in Computer Science, pp. 339–354. Springer.
11. Buccinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020, pp. 454–464. ACM.
12. Checco, A., Roitero, K., Maddalena, E., Mizzaro, S., & Demartini, G. (2017). Let's agree to disagree: Fixing agreement measures for crowdsourcing. In Fifth AAAI Conference on Human Computation and Crowdsourcing.
13. Chen, C., Feng, S., Sharma, A., & Tan, C. (2023a). Machine explanations and human understanding. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 1–1.
14. Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023b). Understanding the role of human intuition on reliance in human-ai decision-making with explanations. Proceedings of the ACM on Human-computer Interaction, 7(CSCW2), 1–32.
15. Chiang, C., & Yin, M. (2022). Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In Jacucci, G., Kaski, S., Conati, C., Stumpf, S., Ruotsalo, T., & Gajos, K. (Eds.), IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022, pp. 148–161. ACM.
16. Chiu, A., Poupart, P., & DiMarco, C. (2007). Generating lexical analogies using dependency relations. In Eisner, J. (Ed.), EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on

Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic, pp. 561–570. ACL.

17. Chromik, M., Eiband, M., Buchner, F., Kr¨uger, A., & Butz, A. (2021). I think i get your point, ai! the illusion of explanatory depth in explainable ai. In 26th International Conference on Intelligent User Interfaces, pp. 307–317.

18. Colligan, L., Potts, H. W., Finn, C. T., & Sinkin, R. A. (2015). Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. International journal of medical informatics, 84(7), 469–476.

19. Cosgrove, M. (1995). A study of science-in-the-making as students generate an analogy for electricity. International journal of science education, 17(3), 295–310.

20. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3), 1155–1170.

21. Dong, Z., & Dong, Q. (2003). Hownet-a hybrid language and knowledge resource. In International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003, pp. 820–824. IEEE.

22. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

23. Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. Plos one, 18(3), e0279720.

24. Draws, T., Rieger, A., Inel, O., Gadiraju, U., & Tintarev, N. (2021). A checklist to combat cognitive biases in crowdsourcing. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 9, pp. 48–59.

25. Duit, R., Roth, W.-M., Komorek, M., & Wilbers, J. (2001). Fostering conceptual change by analogies—between scylla and charybdis. Learning and Instruction, 11(4-5), 283–303.

26. Ehrmann, D. E., Gallant, S. N., Nagaraj, S., Goodfellow, S. D., Eytan, D., Goldenberg, A., & Mazwi, M. L. (2022). Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. Nature Medicine, 1–2.

27. Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable ai: towards a reflective sociotechnical approach. In International Conference on Human-Computer Interaction, pp. 449–466. Springer.

28. Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daum´e III, H., Riener, A., & Riedl, M. O. (2022). Human-centered explainable ai (hcxai): beyond opening the black-box of ai. In CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp. 1–7.

29. Erlei, A., Sharma, A., & Gadiraju, U. (2024). Understanding choice independence and error types in human-ai collaboration. In In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24).

30. Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. Behavior research methods, 41(4), 1149–1160.

31. Fok, R., & Weld, D. S. (2023). In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. AI Magazine.

32. Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: development and validation of the affinity for technology interaction (ati) scale. International Journal of Human–Computer Interaction, 35(6), 456–467.

33. Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 1631–1640.

34. Galesic, M., & Garcia-Retamero, R. (2013). Using analogies to communicate information about health risks. Applied Cognitive Psychology, 27(1), 33–42.

35. Geelan, D. (2012). Teacher explanations. Second international handbook of science education, 987–999.

36. Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive science, 7(2), 155–170.

37. Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity.. American psychologist, 52(1), 45.

38. Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. Advances in neural information processing systems, 32.

39. Gilbert, J. K., & Justi, R. (2016). Analogies in modelling-based teaching and learning. In Modelling-based teaching in science education, pp. 149–169. Springer Gordon, M. L., Lam, M. S., Park, J.

S., Patel, K., Hancock, J. T., Hashimoto, T., & Bernstein, M. S. (2022). Jury learning: Integrating dissenting voices into machine learning models. In Barbosa, S. D. J., Lampe, C., Appert, C., Shamma, D. A., Drucker, S. M., Williamson, J. R., & Yatani, K. (Eds.), CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022, pp. 115:1–115:19. ACM.

40. Green, B., & Chen, Y. (2019a). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the conference on fairness, accountability, and transparency, pp. 90–99.

41. Green, B., & Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–24.

42. Halpern, D. F., Hansen, C., & Riefer, D. (1990). Analogies as an aid to understanding and memory.. Journal of educational psychology, 82(2), 298.

43. He, G., Balayn, A., Buijsman, S., Yang, J., & Gadiraju, U. (2022). It is like finding a polar bear in the savannah! concept-level ai explanations with analogical inference from commonsense knowledge. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 10, pp. 89–101.

44. He, G., Buijsman, S., & Gadiraju, U. (2023). How stated accuracy of an ai system and analogies to explain accuracy affect human reliance on the system. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), 1–29.

45. He, G., & Gadiraju, U. (2022). Walking on eggshells: Using analogies to promote appropriate reliance in human-ai decision making. In Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22).

46. He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on ai systems. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–18.

47. Hofstadter, D. R., & Sander, E. (2013). Surfaces and essences: Analogy as the fuel and fire of thinking. Basic Books.

48. Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cogn. Sci., 13(3), 295–355.

49. Hube, C., Fetahu, B., & Gadiraju, U. (2019). Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12.

50. H¨ullermeier, E. (2020). Towards analogy-based explanations in machine learning. In Torra, V., Narukawa, Y., Nin, J., & Agell, N. (Eds.), Modeling Decisions for Artificial Intelligence - 17th International Conference, MDAI 2020, Sant Cugat, Spain, September 2-4, 2020, Proceedings, Vol. 12256 of Lecture Notes in Computer Science, pp. 205–217. Springer.

51. Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D. L., & Szekely, P. A. (2021). Dimensions of commonsense knowledge. Knowl. Based Syst., 229, 107347.

52. Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., Ploeg, J. v. d., Romaszko, L., Aroyo, L., & Sips, R.-J. (2014). Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In International semantic web conference, pp. 486–504. Springer.

53. Ji, H., Ke, P., Huang, S., Wei, F., & Huang, M. (2020). Generating commonsense explanation by extracting bridge concepts from reasoning paths. In Wong, K., Knight, K., & Wu, H. (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020, pp. 248–257. Association for Computational Linguistics.

54. Jin, W., Li, X., & Hamarneh, G. (2023). Rethinking ai explainability and plausibility. arXiv preprint arXiv:2303.17707.

55. Kawahara, J., Daneshvar, S., Argenziano, G., & Hamarneh, G. (2018). Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics, 23(2), 538–546.

56. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 1–9.

57. Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Vi´egas, F. B., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Dy, J. G., & Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsm¨assan, Stockholm, Sweden, July 10-15, 2018, Vol. 80 of Proceedings of Machine Learning Research, pp. 2673–2682. PMLR.

58. K˙orber, M. (2018). Theoretical considerations and

development of a questionnaire to measure trust in automation. In Congress of the International Ergonomics Association, pp. 13–30. Springer.

59. Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.. Journal of personality and social psychology, 77(6), 1121.

60. Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-ai decision making: a survey of empirical studies. arXiv preprint arXiv:2112.11471.

61. Lai, V., Liu, H., & Tan, C. (2020). "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In Bernhaupt, R., Mueller, F. F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjøn, P., Zhao, S., Samson, B. P., & Kocielnik, R. (Eds.), CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020, pp. 1–13. ACM.

62. Langer, M., Oster, D., Speith, T., Hermanns, H., K¨astner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. Artificial Intelligence, 296, 103473.

63. Law, M. T., Thome, N., & Cord, M. (2017). Learning a distance metric from relative comparisons between quadruplets of images. Int. J. Comput. Vis., 121(1), 65–94.

64. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50–80.

65. Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. arXiv preprint arXiv:2110.10790.

66. Lin, B. Y., Chen, X., Chen, J., & Ren, X. (2019). Kagnet: Knowledge-aware graph networks for commonsense reasoning. In Inui, K., Jiang, J., Ng, V., & Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pp. 2829–2839. Association for Computational Linguistics.

67. Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1–45.

68. Liu, H., Wu, Y., & Yang, Y. (2017). Analogical inference for multi-relational embeddings. In Precup, D., & Teh, Y. W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, Vol. 70 of Proceedings of Machine Learning Research, pp. 2168–2178. PMLR.

69. Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., & Drucker, S. M. (Eds.), CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021, pp. 78:1–78:16. ACM.

70. Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., Dengel, A., & Ahmed, S. (2022). Exaid: A multimodal explanation framework for computer-aided diagnosis of skin lesions. Computer Methods and Programs in Biomedicine, 215, 106620.

71. Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., & Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 4765–4774.

72. Majumder, B. P., Camburu, O., Lukasiewicz, T., & McAuley, J. J. (2021). Rationale-inspired natural language explanations with commonsense. CoRR, abs/2106.13876.

73. Marshall, C. C., & Shipman, F. M. (2013). Experiences surveying the crowd: Reflections on methods, participation, and reliability. In Proceedings of the 5th Annual ACM Web Science Conference, pp. 234–243.

74. Mozzer, N. B., & Justi, R. (2012). Students' pre-and post-teaching analogical reasoning when they draw their analogies. International Journal of Science Education, 34(3), 429–458.

75. Nashon, S. M. (2004). The nature of analogical explanations: High school physics teachers use in kenya. Research in Science Education, 34(4), 475–502.

76. Nourani, M., Honeycutt, D. R., Block, J. E., Roy, C., Rahman, T., Ragan, E. D., & Gogate, V. (2020a). Investigating the importance of first impressions and explainable ai with interactive video analysis. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–8.