Journal of Artificial Intelligence Research 81 (2024) 117-162 Submitted 06/2023; published 09/2024

Opening the Analogical Portal to Explainability: Can Analogies Help Laypeople in AI-assisted Decision Making?

Gaole He G.HE@TUDELFT.NL
Agathe Balayn A.M.A.BALAYN@TUDELFT.NL
Stefan Buijsman S.N.R.BUIJSMAN@TUDELFT.NL
Jie Yang J.YANG-3@TUDELFT.NL
Ujwal Gadiraju U.K.GADIRAJU@TUDELFT.NL
Delft University of Technology,
Mekelweg 5, 2628 CD Delft

Abstract

Concepts are an important construct in semantics, based on which humans understand the world with various levels of abstraction. With the recent advances in explainable artificial intelligence (XAI), concept-level explanations are receiving an increasing amount of attention from the broad research community. However, laypeople may find such explanations difficult to digest due to the potential knowledge gap and the concomitant cognitive load. Inspired by prior work that has explored analogies and sensemaking, we argue that augmenting concept-level explanations with analogical inference information from commonsense knowledge can be a potential solution to tackle this issue. To investigate the validity of our proposition, we first designed an effective analogy-based explanation generation method and collected 600 analogy-based explanations from 100 crowd workers. Next, we proposed a set of structured dimensions for the qualitative assessment of such explanations, and conducted an empirical evaluation of the generated analogies with experts. Our findings revealed significant positive correlations between the qualitative dimensions of analogies and the perceived helpfulness of analogy-based explanations, suggesting the effectiveness of the dimensions. To understand the practical utility and the effectiveness of analogy-based explanations in assisting human decision-making, we conducted a follow-up empirical study ($N = 280$) on a skin cancer detection task with non-expert humans and an imperfect AI system. Thus, we designed a between-subjects study spanning five different experimental conditions with varying types of explanations. The results of our study confirmed that a knowledge gap can prevent participants from understanding concept-level explanations. Consequently, when only the target domain of our designed analogy-based explanation was provided (in a specific experimental condition), participants demonstrated relatively more appropriate reliance on the AI system. In contrast to our expectations, we found that analogies were not effective in fostering appropriate reliance. We carried out a qualitative analysis of the open-ended responses from participants in the study regarding their perceived usefulness of explanations and analogies. Our findings suggest that human intuition and the perceived plausibility of analogies may have played a role in affecting user reliance on the AI system. We also found that the understanding of commonsense explanations varied with the varying experience of the recipient user, which points out the need for further work on personalization when leveraging commonsense explanations. In summary, although we did not find quantitative support for our hypotheses around the benefits of using analogies, we found considerable qualitative evidence suggesting the potential of high-quality analogies in aiding non-expert users in their decision making with AI-assistance. These insights can inform the design of future methods for the generation and use of effective analogy-based explanations.

1. Introduction

In recent years, we have witnessed the rise of machine learning (ML) methods for various applications (*e.g.,* machine translation and object detection). Despite their high accuracy, more and more researchers recognize the necessity to obtain meaningful explanations of these ML methods for real-world scenarios, especially in high-stakes scenarios like medical diagnosis. Machine learning models may provide unreliable predictions based on spurious patterns (*e.g.,* Tesla's self-driving system mistook the moon for a yellow traffic light[1]), which may cause catastrophic consequences (Kelly et al., 2019). With meaningful explanations, humans can better understand the internal working mechanisms and exercise control over powerful machine learning models. With this perspective, a growing number of explainable artificial intelligence (XAI) methods are being proposed to provide explanations for ML model behaviors (Doshi-Velez & Kim, 2017; Ghorbani et al., 2019; Ribeiro et al., 2016).

Identifying and communicating the salient parts of the input (*e.g.,* through pixels in image, or highlighted tokens in text) as explanations is a typical and model-agnostic XAI method (Ribeiro et al., 2016; Lundberg & Lee, 2017; Balayn et al., 2022b), called feature attribution. While such salient parts of the input may be helpful for AI practitioners who have the relevant knowledge, it is still challenging for laypeople to interpret them. To provide more human-friendly explanations, Kim *et al.* (Kim et al., 2018) proposed to derive high-level concepts to describe the internal state of models. Compared with low-level salient features, high-level concepts have been shown to be more understandable for laypeople. However, in many real-world tasks, these high-level concepts (*e.g.,* chemicals, cells in medical diagnosis) are still not comprehensible for laypeople due to the gap of domain knowledge and expertise. At the same time, it is unnecessary for users or stakeholders (*e.g.,* patients or loan applicants taking medical or financial advice) to fully understand the explanation technically. Their information need is often satisfied by understanding explanations adequately enough to achieve better decision making for their own benefit.

The challenge, therefore, is to provide the right kind of explanations. Transparency about systems, and the provision of explanations, is likely to be a requirement in the AI Act (Sovrano et al., 2022) for a wide range of systems. Likewise, according to General Data Protection Regulation (GDPR),[2] the users of AI systems should have the right to access meaningful explanations of model predictions (Selbst & Powles, 2018). This implies that intelligible explanations which can facilitate such an understanding for laypeople are required. We argue that analogy-based explanations can be a potential solution to fill in this gap in understanding. We illustrate our motivation through an example in Figure 1. Given a concept-based explanation extracted from an ML model, laypeople may still have difficulties connecting the concepts (*i.e.,* cribriform and fused glands in needle core biopsy) with specific model predictions (*i.e.,* positive for prostate cancer). Such explanations can be difficult to understand due to the lack of domain knowledge and expertise, and they can be a heavy burden when figuring out the causality or relevance of observing these concepts to make the prediction (Abdul et al., 2020; He & Gadiraju, 2022; Ehrmann et al., 2022).

An analogy can be interpreted as a structural mapping from a target domain to be clarified, onto a source domain which the recipient of the analogy is more familiar with (Gentner, 1983; Hofstadter & Sander, 2013). For example, in Figure 1, the target domain, *medical diagnosis*, is clarified based on a source domain: *fantasy*. Through everyday experiences, laypeople master

1. https://www.autoweek.com/news/green-cars/a37114603/tesla-fsd-mistakes-moon-for-traffic-light/
2. https://gdpr-info.eu/

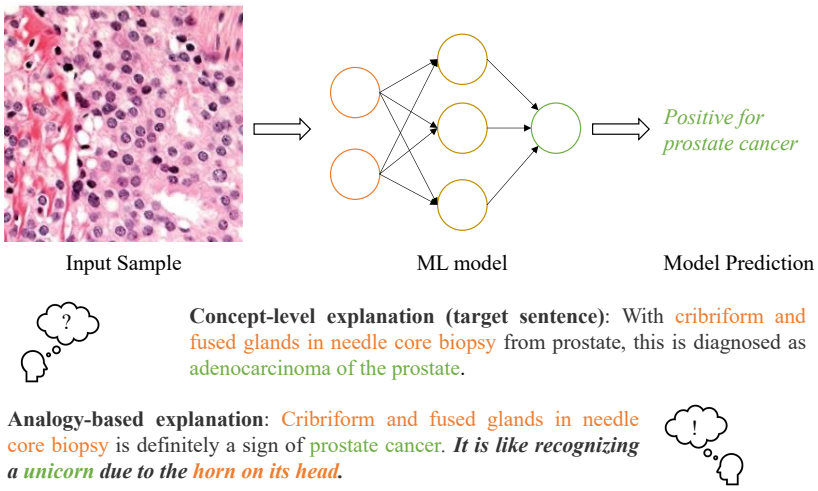


Figure 1: Example of analogy-based explanation in prostate cancer detection. The medical image and the concept-level explanation are sourced from (Verhoef et al., 2019).

commonsense knowledge of the world and build up sophisticated mental models to deal with regular tasks; *e.g.,* a single horn on the head of a beast is an important pattern for recognizing a unicorn. With analogy-based explanations, high-level concepts and model predictions can be translated into everyday concepts that laypeople are familiar with, by triggering their capabilities of analogical inference. From this standpoint, we argue that laypeople can leverage the sophisticated mental models of their worldly experiences to interpret the behavior of ML models and generate meaningful analogy-based explanations. Thus, users can understand that the complex concepts in "*cribriform and fused glands in needle core biopsy*" are also a strong pattern which indicates the model prediction "*positive for prostate cancer*." Laypeople (or non-expert users) can thereby use the explanation adequately enough to inform their decisions, without having to understand the concepts from a technical standpoint, addressing the knowledge gap while reducing their cognitive load.

Despite the intuitive promise and potential of analogy-based explanations, how to generate such analogy-based explanations remains an open question. In addition, we also lack a framework to qualitatively characterize and evaluate the generated analogies. Hence, in this work, we first address the following research questions:

(**RQ1**) *How can we generate high-quality analogy-based explanations using non-experts*?
(**RQ2**) *How can we systematically assess the quality of analogy-based explanations?*

To the best of our knowledge, no work has yet investigated whether conceptually high-quality, analogy-based, explanations can be helpful for human-AI collaborative decision making. Inspired by recent literature on human-centered explainable AI (Ehsan & Riedl, 2020; Liao & Varshney, 2021), a human-grounded evaluation (Doshi-Velez & Kim, 2017) can further our understanding of the impact of analogy-based explanations in decision support. Hence, as a second step of our work beyond generating analogy-based explanations and evaluating their *conceptual quality*, it is also

important to validate their effectiveness in assisting human decision making *in practice*. To this end, we aim to address the following questions:

(**RQ3**): *How do analogies for concept-level explanations shape the understanding of an AI system among non-expert users*?
(**RQ4**): *How do analogy-based explanations affect user reliance on AI systems*?

To answer **RQ1**, we designed a novel analogy generation method that leverages templates and crowd computing to obtain high-quality analogy-based explanations. To answer **RQ2**, we first defined a structured set of dimensions through which one can conceptually assess the quality of analogy-based explanations. Then we recruited 100 crowd workers as non-experts to generate analogy-based explanations using our method. After that, we carried out an expert evaluation of the quality of the collected explanations across the different dimensions. To answer **RQ3** and **RQ4**, we formulated four hypotheses about the effect of the analogy-based explanations on user understanding, appropriate reliance, cognitive load, and decision making efficiency. We tested these hypotheses in an empirical study with crowd workers (N = 280), asked to perform a skin cancer detection task, in four different human-AI collaborative decision making settings.

In our empirical study, we found that the mere presence of the target domain information within the analogy-based explanations was most effective in mitigating under-reliance but also gave rise to over-reliance. However, we did not find an improved understanding of the AI system or a statistically significant increase in appropriate reliance when all the information contained in the analogy-based explanations was presented. This was particularly the case when analogies were provided on demand. Surprisingly, such analogy-based explanations could even have some negative impact on the appropriate reliance. Analyzing the participants' qualitative feedback about the analogy-based explanations helped us understand the unexpected reliance patterns (*i.e.,* over-reliance and under-reliance) and the potential role of human intuition and plausibility in shaping our findings. Introducing analogies did not pose a significantly higher cognitive load on users, or cause a significant delay in decision making efficiency. Collectively, our findings suggest that although analogies may not be universally effective in fostering appropriate reliance in the context of human-AI decision making, there is some potential for analogy-based explanations in assisting laypeople for efficient decision making if they can be personalized. Our main contributions can be summarized as follows:

- A novel analogy-based explanation generation method with non-expert crowds and a dataset of analogies generated using this method.
- An elaborate set of qualitative dimensions to assess the quality of analogy-based explanations.
- An extensive evaluation of the quality of the analogy-based explanations collected from two distinct AI tasks.
- A rigorous empirical study in the context of human-AI decision making to understand the effectiveness of analogy-based explanations in a skin cancer detection task.
- Guidelines for the generation of effective analogy-based explanations and for the appropriate use of such analogy-based explanations.

Note that this manuscript is an extended version of the paper (He et al., 2022), extended in the following ways: To validate the effectiveness of analogy-based explanations, (1) we proposed new research questions and hypotheses about the impact of analogy-based explanations on a user's understanding of an AI system and their appropriate reliance on the system; and (2) we conducted an

empirical study of human-AI decision making on a skin cancer detection task to test these hypotheses; (3) based on the results from our empirical study, we synthesized guidelines for future work on the generation and use of analogy-based explanations in the context of human-AI decision making.

If not used appropriately, analogy-based explanations may not work as expected to improve human-AI collaborative decision making. To the best of our knowledge, this is the first work that combines analogy-based explanations with commonsense knowledge in the context of human-centered explainable AI. Based on the results from our empirical study, we synthesize promising future directions for further XAI research.

2. Background and Related Work

We position our work in the following realms of related literature: *commonsense knowledge*, *analogy-based explanation*, *human-AI decision making* and the context of *human-centered explainable AI*.

2.1 Commonsense Knowledge

Commonsense knowledge is "information that humans typically have that helps them make sense of everyday situations" (Ilievski et al., 2021). It has been proved to be highly useful in various AI applications, like question answering (Lin et al., 2019), dialogue systems (Young et al., 2018) and visual reasoning (Zellers et al., 2019). However, due to the intrinsic implicitness, commonsense knowledge is usually omitted in oral or written communication (Ilievski et al., 2021). To collect such implicit knowledge, researchers have proposed to make use of the wisdom of crowds, through text mining of corpora (Singh et al., 2002; Speer et al., 2017), and via games with a purpose (von Ahn et al., 2006; Balayn et al., 2022a).

In recent years, commonsense knowledge has been used to also improve the explainability of AI models. In commonsense reasoning tasks, explanations from humans which contain rich commonsense knowledge, have been shown to be highly useful both to boost performance and to aid understanding (Rajani et al., 2019). In addition to generating commonsense explanations with humans, some studies have also demonstrated that commonsense knowledge can help build connections between multiple statements (Ji et al., 2020) and enhance natural language explanation generation with extractive rationales (Majumder et al., 2021).

To facilitate the understanding of concept-level explanations, we propose to generate commonsense explanations for laypeople. The commonsense knowledge contained within such explanations forms the source domain over which laypeople can exercise their analogical reasoning, to improve their understanding of the concept-level explanations.

2.2 Analogy-based Explanations

Analogy-based explanations have been extensively studied in many research domains such as logic, linguistics, and philosophy. "An analogy is created when some aspects of an unknown target are compared with those of a source about which more is known" (Gilbert & Justi, 2016). Due to such intrinsic property for elucidating new knowledge with existing knowledge, analogies have been adopted as explanation in education, and supported by multiple research work (Nashon, 2004; Geelan, 2012; Mozzer & Justi, 2012).

In the context of artificial intelligence, the importance of analogies has been recognized by multiple AI applications such as representation learning (Liu et al., 2017), preference learning (Bounhas

et al., 2019), and image processing (Law et al., 2017). Readers can refer to (Prade & Richard, 2021) for a more comprehensive survey of analogical inference in the context of AI, which is beyond the scope of this paper. However, only a few works (Hüllermeier, 2020; He & Gadiraju, 2022) explored the potential of analogy-based explanations in the context of XAI. While such works show and argue that analogy-based explanations have great potential in XAI, it is still unclear how we can measure the quality of analogy-based explanations and how we can efficiently generate such analogy-based explanations for machine learning applications.

As for analogy generation, in addition to previous methods that relied on human intelligence for drawing out analogies in instructional, teaching and educational contexts (Duit et al., 2001; Cosgrove, 1995), some research has also explored the automatic generation of analogies. Veale *et al.* (Veale, 2005) explored how lexical resource HowNet (Dong & Dong, 2003) can support analogy generation with two approaches: (1) abstraction via a taxonomic backbone, (2) selective projection via structure-mapping on propositional content. Chiu *et al.* (Chiu et al., 2007) propose to generate lexical analogies with the help of dependency relations from unstructured text data. However, such methods do not incorporate commonsense knowledge, making it inappropriate for explaining to laypeople the complex concept-level explanations. That is why we adopt a crowd computing-based method to generate analogy-based explanations.

In this paper, we propose structured dimensions for the qualitative assessment of analogy-based explanations. We also design a crowd computing method to generate such explanations, and empirically evaluate its effectiveness.

2.3 Human-Centered XAI and the Human-AI Decision Making

Explainability is a concern for AI systems, especially for black box deep learning models. To provide meaningful explanations for AI predictions, a wide range of explainable artificial intelligence (XAI) tools have been proposed (Arrieta et al., 2020). However, due to the inherent human-centric property of explainability (*i.e.,* explanations are only successful if they match the specific needs of the person receiving them), there is no one-size-fits-all solution in the growing collection of XAI techniques (Liao & Varshney, 2021). Consequently, researchers have increasingly begun to explore the area of human-centered explainable artificial intelligence (HCXAI) (Ehsan & Riedl, 2020; Wang et al., 2019; Liao & Varshney, 2021; Ehsan et al., 2022), by putting the human at the center of technology design (Ehsan & Riedl, 2020).

Human-AI decision making has emerged as an important paradigm to augment human capabilities with the computational prowess of AI systems, leading to complementary teamwork and effective decision making (Lai et al., 2021). In the collaborative decision making process, human factors (*e.g.,* AI literacy (Chiang & Yin, 2022) and cognitive bias (Bertrand et al., 2022)) and interaction with AI systems (tutorial intervention (Lai et al., 2020; He et al., 2023) and performance feedback (Lu & Yin, 2021)) are observed to affect subjective trust and reliance behaviors greatly. In recent works with human-AI decision making, researchers have shown great interest in achieving complementary team performance with appropriate reliance on the AI system by exploring a multitude of factors including human and task factors (Schemmer et al., 2022; He et al., 2023; Erlei et al., 2024; Salimzadeh et al., 2024).

To help users better understand AI advice and inform the trustworthiness, XAI methods are widely analyzed in human-AI decision making. Based on a comprehensive literature review, Wang *et al.* (Wang & Yin, 2021) summarized three desiderata of AI explanations to facilitate comple-

mentary teamwork: (1) Explanations of an AI should improve people's understanding of it, (2) Explanations of an AI should help people recognize the uncertainty underlying the AI, and rely on the high-confidence predictions when model confidence is calibrated, (3) Explanations of an AI should empower people to trust the AI appropriately. However, most XAI methods are rarely found helpful in achieving a complementary performance in human-AI decision making (Bansal et al., 2021; Liu et al., 2021; Fok & Weld, 2023). Sometimes, XAI methods can even make users suffer from automation bias (Vered et al., 2023), which will cause over-reliance on the AI system.

AI systems have become ubiquitous in intelligent applications around our daily life, and involve nearly everyone as stakeholder rather than experts only. Different communities of stakeholders (Preece et al., 2018) have different goals and explainability needs. For example, system developers require explainability to debug the system, while system users may place more emphasis on the explainability of outputs in order to aid their own decision making (Preece et al., 2018; Langer et al., 2021). As a result, explanations should be tailored to different stakeholders.

Inspired by previous studies about analogy-based explanations (Hüllermeier, 2020; He & Gadiraju, 2022), we focus on explainability for laypeople using such explanations:

- Laypeople lack technical expertise and domain knowledge to interpret AI systems. Analogy-based explanations fill in such knowledge gap with concepts they are familiar with.
- Analogy-based explanations provide familiar information for laypeople, which reduces the cognitive load for comprehension compared to concept-level explanations which contain uncommon terminologies.

3. Quality of Analogy-based Explanations

We first conducted a systematic review of existing works in the area of analogy-based explanations, in order to understand how the quality of analogy-based explanations has been empirically investigated in prior literature.

3.1 Effective Analogies

Properties of analogical argument. Analogies have been widely used as explanations for educational and learning purposes (Nashon, 2004; Mozzer & Justi, 2012). With analogical inference, humans can compare one new topic that is being introduced with another topic they are already familiar with, which leads to a better understanding of the new topic by relating back to previous knowledge (Halpern et al., 1990). However, to make the analogy-based explanations work as an aid to understand new knowledge or events, several properties need to be satisfied by the analogical arguments. Aristotle's theory provides us with four important and influential criteria for the evaluation of analogical arguments (Bartha, 2022):

- The strength of an analogy depends upon the number of similarities.
- Similarity reduces to identical properties and relations.
- Good analogies derive from underlying common causes or general laws.
- A good analogical argument need not pre-suppose acquaintance with the underlying universal (generalization).

In previous studies, researchers also emphasized the importance of the quality of structural mapping. According to (Gilbert & Justi, 2016; Gentner, 1983), an analogy needs to fullfill certain constraints to work as expected – (i) there should only be a single one-to-one correspondence be-

tween each pair of elements; (ii) it must involve common relationships across the source domain and target domain (iii) an analogy must describe systems of connected relations, which permits the generation of inferences. According to the multiconstraint theory (Holyoak & Thagard, 1989), people use analogies guided by a series of constraints that favour coherence in analogical reasoning (Mozzer & Justi, 2012). The constraints are semantic similarity, structural correspondence, and purpose. Specifically, the similarity in concept level contributes to analogical reasoning, while the structural constraint helps to establish an isomorphism between source domain and target domain. Furthermore, the analogical reasoning is guided by the purpose. In addition to ensuring the analogical properties of the structural mapping, Thalheim *et al.* (Thalheim, 2011) further considered the "degree of structural adjustment" (*i.e.,* the extent to which the structure is considered independent on the later use). This dimension evaluates the *transferability* of the generated source artifact.

Factors shaping the effectiveness of analogies. Apart from the properties of analogical argument, there are other factors which affect the effectiveness of analogy-based explanations. To guarantee the usefulness of analogy-based explanations, explanation consumers should be familiar with the source domain (*e.g.,* the generated commonsense explanations in our case). According to Galesic *et al.* (Galesic & Garcia-Retamero, 2013), the most helpful analogies boast a high relational similarity between the source and target domain and a high familiarity with the source domain. Thalheim *et al.* (Thalheim, 2011) also argued that the source domain of effective analogies should be "easily interpretable and understandable".

3.2 Synthesizing a Structured Set of Dimensions

Analogical Properties. According to the above, the quality of generated analogy-based explanations is largely reflected by the quality of the analogical properties, that rely on comparing the source domain (*i.e.,* generated commonsense explanation) to the target sentence. In this paper, we base the quality of analogical properties on four aspects: (1) **structural correspondence** between the target domain (*i.e.,* observed concepts and model prediction) and source domain (*i.e.,* concepts used in the explanation), (2) **relational similarity** between the target domain (*i.e.,* relation between observed concepts and model prediction) and source domain (*i.e.,* relation between concepts in explanation), (3) **transferability**, *i.e.,* the extent to which the structure is considered independent of its later use, and (4) **helpfulness**, *i.e.,* the extent to which the generated commonsense explanation is considered helpful to understand the target sentence.

Among these dimensions, "relational similarity" and "structural correspondence" have been highlighted by existing works with phrases like "semantic similarity" (Holyoak & Thagard, 1989) and "structural alignment" (Gentner & Markman, 1997). "Helpfulness" corresponds to the "purpose" mentioned in Holyoak and Thagard's multiconstraint theory (Holyoak & Thagard, 1989), while "transferability" corresponds to the "degree of structural adjustment" (Thalheim, 2011). To assess the "helpfulness" of explanations, we need to ground them within specific tasks. In this paper, we conduct human-based evaluation to assess the extent to which the analogy-based explanations can be helpful to explain the original concept-level explanations. In practice, the generated analogy-based explanation may also be fit to explain other concept-level explanations which show similar information. To serve that purpose, one can argue that high-quality analogy-based explanations should be capable of generalizing to more tasks. Thus, we also consider the "transferability" of generated analogy-based explanations.

As mentioned above, the generated analogy-based explanations can be used to explain other tasks than the one used for generation. In such cases, it is also necessary to evaluate the quality of the explanations. All the dimensions we propose can be used to assess such quality for these new tasks.

Utility. In addition to the above dimensions, we identified dimensions specifically related to the generated commonsense explanations. These dimensions are independent of the target sentence, but may also affect the effectiveness of analogy-based explanations.

Some dimensions are identified from the factors shaping the effectiveness of analogies mentioned previously. They are: (5) explainee's **familiarity** with the concepts mentioned in generated explanation; (6) **simplicity** of the analogy-based explanation, which describes how easily laypeople can interpret and understand the explanation would be (Thalheim, 2011). We also identify other dimensions based on intuitively desirable expectations from effective explanations. Reducing the scope for misunderstanding can aid the overall comprehension of analogy-based explanations. Thus, we also consider the dimension of (7) **misunderstanding**, which occurs when different interpretations exist for a single analogy-based explanations. For example, the phrase "*subway definitely contains seats*" can be interpreted as referring to *e.g.,* either the restaurant, "Subway", or an underground railway. To ensure the utility of generated explanations, it is vital to ensure that they are (8) **syntactically correct**, and (9) **factually correct**. That means the explanations are comprehensible according to syntactic grammar, and describe the truth about the world. Further details including our annotation of these dimensions are provided in section 5.

4. Analogy Generation

We propose a crowd computing method to generate analogy-based explanations using image classification tasks as an empirical lens, and verify the effectiveness of our proposed set of dimensions in determining the quality of the analogy-based explanations.

(a) Calorie dataset. (b) Places dataset.

Figure 2: Example of tasks used to generate analogies.

Tasks for Analogy Generation. To collect useful analogy-based explanations from crowd workers, we need to adopt task contexts which non-experts are capable of interpreting and explaining. We also consider the relationship explicitness in the task domain. In some domains, it is difficult to elucidate relationships between concepts and labels other than ascribing correlation (*e.g.,* food to

calorie level). In others (such as furniture to places), most concepts and the labels have a clear indication of relationships like "PartOf", "SignOf", and "FoundAt", which also appear in commonsense knowledge bases like ConceptNet (Speer et al., 2017). Hence, we select two image classification tasks: calorie level classification (CLC) and scene classification (SC).

For the calorie level classification task, we used the dataset provided by Buçinca *et al.* (Buccinca et al., 2020), where two possible labels are attached to images: (1) *high calorie level*, fat more than 30%, (2) *low calorie level*, otherwise. In this task, participants are given an image (see Figure 2(a)) along with concepts highlighted with bounding boxes (*i.e.,* chocolate and ice cream) and the predicted calorie level. For the scene classification task, we used a subset of the Places dataset (Zhou et al., 2018), which covers six place labels: *living room*, *bathroom*, *hospital room*, *conference room*, *bedroom*, *dining room* (Figure 2(b) is an example of a *conference room*). In both tasks, we ask participants to describe the relevance of given concept(s) and labels, *e.g.,* the relevance of food concept(s) and calorie levels, with explanations constructed using everyday concepts and given templates.

Table 1: Templates used in *analogy generation* with placeholders presented to the users (bold text in square brackets).

Relevance	Template	Example
Positive Evidence	Definite Sign Of	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	Chocolate is typically associated with high calorie food, while rarely associated with low calorie food. This is like **[printers]** can typically be associated with **[offices]**, but it's also possible to associate **[printers]** with **[homes]**.
Inconclusive Evidence	Insufficient	Bread is not sufficient to indicate high calorie, as both high calorie food and low calorie food 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]**.
	Irrelevant	A plate is irrelevant to indicate high calorie food. This is similar to to how **[an arbitrary stone]** is irrelevant for **[recognising a continent]**.
Negative Evidence	Seldom Found At	Carrots are seldom found in high calorie food. This is like **[cats]** can seldom be found in **[water]**.
	Contradict With	A vegetable salad contradicts with high calorie food. This is similar to how one cannot find **[water]** in **[electrical appliances]**.

Templates for Analogy-based Explanations. To help crowd workers associate the concepts with model predictions, we provide templates for generating analogy-based explanations. Machine learning models may learn both useful concepts and spurious concepts to make predictions (Kim et al., 2018). Some of the useful concepts can directly lead to the correct conclusion, while others are highly relevant and helpful to predict the label but not definite. In comparison, the spurious concepts are irrelevant or insufficient (like predicting a *dog* in image by focusing on *grass field*) to make the prediction, and sometimes even contradict with our commonsense knowledge, leading to an incorrect prediction. Hence, we decide to use six templates based on three different relevance levels (*i.e.,* positive evidence, inconclusive evidence, and negative evidence). For each relevance level, we have one template to indicate the type of relationship and another one to indicate relevance. The templates along with examples can be found in Table 1.

Task Selection. To balance the generated analogies in each relevance category, we manually selected two tasks for each category according to the authors' interpretation of their relevance levels. Thus, we use 12 tasks for analogy generation: 6 for calorie level (CLC) and 6 for scene classification (SC).

Hints for Analogy Generation. Through a pilot study, we learned that although non-expert crowd workers can generate analogies based on their own experience, it becomes challenging to generate new analogies after a handful of tasks. To help crowd workers in generating high-quality analogies, we provide a list of hint domains with a clickable button in the interface. The list contains: weather, animals and plants, place, transportation, food, art, education, sports, finance, clothes, electronics, games and toys, health.

Analogy Generation Procedure. To generate high-quality analogies, we provide the six templates shown in Table 1 to each participant. Participants are first asked to select one template, comprising one sentence with placeholders for concepts. They can then refer to our example analogies and everyday domains provided as hints. Next, based on the template, they are asked to fill in one word or phrase (up to five words) as a concept in each placeholder. All participants are forbidden to fill in concepts belonging to the task domain (such as places and furniture in the Places task). An example of the analogy generation interface is shown in Figure 3.

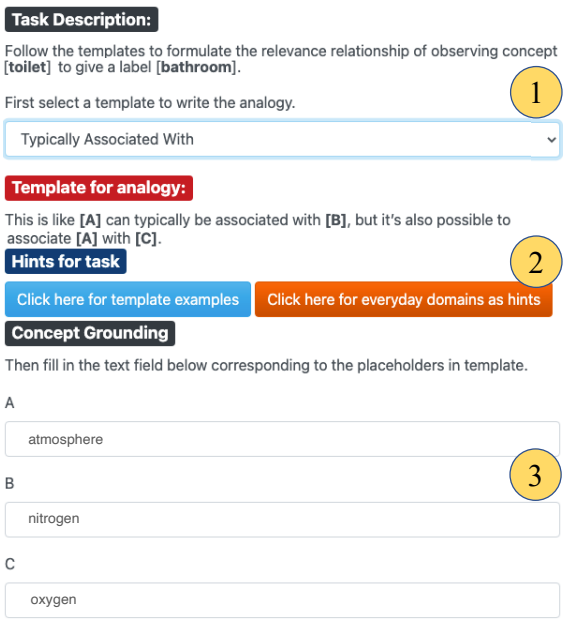


Figure 3: Analogy generation main interface and workflow. (1) Participants select a template to describe the relevance level; (2) refer to examples and everyday domains as hints; and (3) fill in concepts in placeholders to generate analogy.

5. Study I: Analogy Generation and Evaluation

In the first study, our experiment mainly consists of two stages: (1) analogy generation with crowd workers, (2) evaluation of generated analogies with third-party experts.

5.1 Analogy Generation Based on Non-experts

Pilot Study. We conducted a pilot study with 7 participants hired from Prolific[3] crowdsourcing platform. All participants were asked to complete 12 tasks (6 for CLC, 6 for SC). Through the pilot study, we gained the following insights:

- After generating several analogies, participants found it difficult to generate new analogies (*i.e.,* required more time for analogy generation and repeated concepts used). To help with this issue, we provided a list of daily domains as hints. As a consequence, we also reduced the number of tasks that each participant was required to complete in the analogy generation phase of the main study.
- Some participants used the examples or concepts shown in one task (*e.g.,* calorie) as answers for another one (*e.g.,* places). To counter such behavior, we decided to limit each participant to a single generation task.

Informed by these observations, we asked each participant in the main study to work on 6 analogy generation tasks from one task domain (either CLC or SC).

Participants. In the main study, we recruited 50 crowd workers for the calorie task, and 50 crowd workers for the places task. In total, 600 analogy-based explanations were generated. We compensated each worker with £1.35 (*i.e.,* 9 min $\times$ hourly salary £9). All participants were proficient English-speakers above the age of 18 and they had an approval rate of at least 90% on the Prolific platform.

Quality Control. To discourage unreliable behavior (*e.g.,* copy-pasting concepts from the task description and examples provided), we enforce all concepts mentioned in the task description and possible labels in each task as taboo phrases (words). We also prevent participants from generating the same analogy-based explanations twice.

5.2 Analogy Evaluation with Experts

Experts. To ensure a fair evaluation of the quality of generated analogies, we recruited 5 external experts from the department of the authors' institute using a purposeful sampling strategy (Stratton, 2021). All experts had at least a basic knowledge of machine learning and explainable AI.

For the purpose of this evaluation, we considered a subset of the analogies generated from 23 participants in the calorie task and 26 participants in the place task (we randomly sampled around half of the participants in our study). In total, we consider 294 analogy-based explanations for evaluation. We ensured a 10% (*i.e.,* 29 analogy-based explanations) overlap across experts. Thus, each expert evaluated 82 different analogy-based explanations. On average, each expert spent 2.5 hours on this qualitative evaluation.

Qualitative Assessment. Based on our synthesis of the dimensions for quality of analogies (cf. Section 3.2), the quality of analogy-based explanations was mainly assessed across two categories: (1) analogical properties and (2) utility. We followed an iterative coding process (Strauss, 1987) to

3. https://www.prolific.com/

Table 2: Structured dimensions used in qualitative assessment of analogy-based explanations.

Category	Dimension	Questionnaire	Scale
Analogical Properties	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
	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
	Transferability	How well can the explanation be used in other contexts?	5-point Likert
	Helpfulness	How helpful is this explanation for you to understand the target sentence?	5-point Likert
Utility	Familiarity	How familiar are you with the concepts in the explanation?	5-point Likert
	Simplicity	Do you think the explanation is simple enough for others to understand?	5-point Likert
	Misunderstanding	Do you think this explanation lead to more than single interpretation?	{Yes, No}
	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? (switch concept A and concept B in template)	{Yes w/o switch, Yes & switch, No}

characterize the quality of the analogy-based explanations across dimensions informed by our synthesis from literature. While different terminologies (*e.g.,* degree of structural parallelism (Bartha, 2022), degree of structural analogy (Thalheim, 2011), semantic similarity (Holyoak & Thagard, 1989)) were adopted to assess the quality of analogies and their quality as explanations, we aimed to address the redundant definitions and integrate a structured set of dimensions for the qualitative assessment (see dimension and questionnaire in Table 2).

Annotation Rubrics. Through iterative coding interspersed with discussions, the authors finally constructed the following annotation rules to guide the qualitative assessment:

- If the concepts of commonsense explanation are of the same domain as the target sentence (regarded as invalid due to non-compliance with analogy generation instruction), annotators can skip that annotation.
- For *Factual Correctness*, take the generated explanation "*The pink feather is definitely a sign of flamingo*" as an example. This explanation can be factually correct after we switch the order of "pink feather" and "flamingo".
- When *Misunderstanding* exists, we consider one analogy as factually correct when a single interpretation can be true. For example, "subway is definitely a sign of seat". When interpreting the "subway" as the one in transportation, we can consider it as being factually correct.
- For *Transferability* and *Helpfulness*, assign '1' when *Factual Correctness* = No
- We devised additional, concrete rubrics for each of the other dimensions. While we do not present them here for space consideration, they can be found online.[4]

Procedure. In the beginning, we provided an annotation manual for each expert. They spent around 10 minutes on reading the annotation manual which contains both dimensions and annotation rules we mentioned above. In this process, we also answered their questions to clarify any issues related

4. https://github.com/delftcrowd/HCOMP2022_ARCHIE/blob/main/annotation_manual/annotation_manual.pdf

Table 3: Evaluation of the following analogy by 5 experts illustrating disagreement – "*Lemon is seldom found in high calorie food. This is similar to how having hair is irrelevant for recognising a human.*"

Dimension	E_1	E_2	E_3	E_4	E_5
Structural Correspondence	4	3	5	1	2
Relational Similarity	1	1	5	1	3
Familiarity	4	5	5	5	2
Helpfulness	1	5	5	1	2
Transferability	4	5	5	1	2
Simplicity	3	5	5	2	3

to quality evaluation. After that, each expert independently worked on the 82 samples provided according to the rubric we provided.

Annotation Agreement. We calculated the annotation agreement based on 29 samples (overlap for experts) in evaluation experiment. As 7 analogy-based explanations are recognized as invalid (crowd workers generate the explanation with concepts via the same domain as target sentence), we calculated the Krippendorff's α scores based on the valid 22 analogy-based explanations. Due to the subjectivity in evaluating the dimensions in the 5-point Likert scales, we merge the 5 items into three levels of attitude (*i.e.,* Negative=$\{1, 2\}$; Neutral=$\{3\}$; Positive=$\{4,5\}$) when calculating the Krippendorff's α scores. The results are respectively 0.15 for *Structural Correspondence*, 0.17 for *Relational Similarity*, 0.22 for *Factual Correctness*, 0.64 for *Syntactic Correctness*, 0.35 for *Misunderstanding*, 0.03 for *Familiarity*, 0.14 for *Helpfulness*, 0.11 for *Transferability*, and 0.14 for *Simplicity*. Naturally, the experts show relatively higher agreement on *Factual Correctness*, *Syntactic Correctness*, and *Misunderstanding*, which are more objective than the other dimensions. The disagreement on other dimensions is due to the subjectivity of the task (Checco et al., 2017): knowledge and the quality of an analogy-based explanation vary depending on one's own experience of the world.

For further illustrative analysis, let us consider an example analogy-based explanation which received disagreement among experts on most dimensions — "*Lemon is seldom found in high calorie food. This is similar to how having hair is irrelevant for recognising a human*". All experts see this analogy-based explanation as factually correct and syntactically correct without any misunderstanding. As the experts assessment reveals in Table 3, the experts diverge on most dimensions of the Likert scale.

For further insights in the disagreement, we ask the experts to explain their scoring. We find multiple user factors can lead to disagreement. For instance, we observed that: (i) The overall negative attitude of E_4 ("*I just gave it a low number because I didn't really understand what it was trying to tell me*") towards this explanation, and the severity of E_5 make them rate most dimensions lower. (ii) As the relationship between "lemon" and "high calorie" is not explicit, experts seem to have different interpretation of the relationship, leading to disagreement on *Relational Similarity*. While E_1, E_2, E_5 would rate it low, E_3 judge it high, because *"calorie is a common property of food, which is not unique to Lemon. having hair is also a common (mostly) property of humans, which is not unique to a specific person"*. (iii) Some experts have more abstract thinking on the properties and relations, again causing disagreement. E_1 gives a 4 to *Structural Correspondence* because they

think "human" and "high calorie" have some connections. And E_2 would rate *Relational Similarity* as 1 because "*people have hair, lemon are not high calorie food*". Besides, we also notice that both E_1 and E_5 take this explanation as unhelpful due to poor *Relational Similarity*.

5.3 Results and Analysis

In this subsection, we present the quality assessment results for the generated analogies with the proposed approach.

Dimension	Label	Example
Structural Correspondence	1	Chocolate and cream contradict with low calorie food. This is similar to how one cannot find tsumanis in uk.
	3	Nuts is insufficient to indicate high calorie. This is similar to how we can find hairdryer in both hotel and hairdresser, you can't determine where it is if you see hairdryer.
	5	A medical monitor is a definite sign of hospital room. This is like an echocardiogram is definitely a sign of pulse oximeter.
Relational Similarity	1	Nuts are seldom found in high calorie food. This is similar to how one cannot find fire hydrants in boats.
	3	Fireplace is not sufficient to indicate bedroom. This is similar to how we can find wig in both pantomime and courtroom, you can't determine where it is if you see wig.
	5	A medical monitor is a definite sign of hospital room. This is like doctor is definitely a sign of surgery.
Transferability	1	A fireplace is a definite sign of bedroom. This is like art is definitely a sign of human expression.
	3	Beet and apple contradict with high calorie food. This is similar to how one cannot find toys in a clothes store.
	5	Chocolate and ice cream is a definite sign of being high-calorie. This is like keyboard is definitely a sign of having a computer.
Helpfulness	1	Toothbrush and towel are insufficient to recognize a bathroom. This is similar to how we can find reading in both education and hobby.
	3	Chocolate and cream are definitely a sign of high calorie food. This is like udders are definitely a sign of cow.
	5	A fireplace can seldom be found in a bedroom. This is like dogs can seldom be found in a fishtank.
Familiarity	1	Chocolate and cream contradict with low calorie food. This is similar to how one cannot find bargains in harrods.
	3	Chocolate and cream are seldom found in low calorie food. This is like roar can seldom be found in big animal.
	5	Nuts is not sufficient to indicate high calorie food. This is similar to how we can find books in both libraries and schools, you can't determine where it is if you see books.
Simplicity	1	Carrot is not sufficient to indicate high calorie. This is like diets can typically be associated with field of hay, but it's also possible to associate diets with gemstones in a gold mine.
	3	Table and chair is insufficient to indicate a conference room. This is like atmosphere can typically be associated with nitrogen, but it's also possible to associate atmosphere with oxygen.
	5	Chocolate and ice-cream are a definite sign of high-calorie. This is like duvet is definitely a sign of bed.

Table 4: Examples of analogies generated for the different scale items of each dimension of the qualitative analysis.

5.3.1 Descriptive Statistics

In the analogy generation experiment, crowd workers are asked to generate explanations with concepts in a different domain from the target sentence. The generated analogies that violate this requirement are then regarded as being invalid. Among the 294 generated analogy-based explanations, 255 (nearly 87%) were recognized as valid by all five experts. As the annotation rubric described, experts only provide qualitative evaluation for valid analogy-based explanations. Finally, we gathered 358 valid evaluation results for 410 samples (82×5, with 29 samples overlap for each).

When generating the analogy-based explanations, crowd workers used everyday concepts in domains "Animals", "Scene/Place", and "Weather" most frequently, which are also in the hint list we provide. For the identified relationship between concepts in generated analogy, crowd workers prefer to use "FoundAt" (175 times), "SignOf" (158 times), and "PartOf" (24 times).

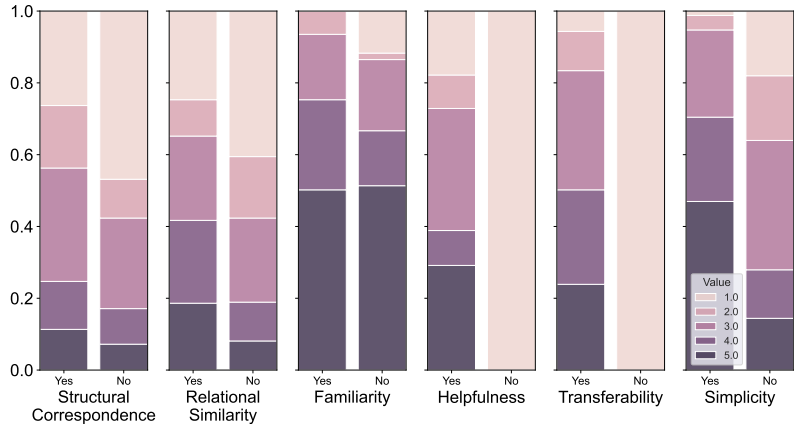


Figure 4: Stacked histogram illustrating the difference across the qualitative dimensions based on Factual Correctness. All dimensions were measured on a 5-point Likert scale.

Analogy quality. Among 358 valid evaluation results, 310 cases were found to be syntactically correct, 198 cases were factually correct without switching placeholder A and B, 49 cases are factually correct with switching (in total, 79.7% of explanations could be generated as factually correct). Meanwhile, only 53 cases were found to potentially lead to multiple interpretations. We compare the quality of analogy-based explanations based on the category of *Factual Correctness*. As shown in Figure 4, the factually correct analogy-based explanations show better quality in nearly all dimensions in 5 point Likert scale than factually incorrect counterparts. As factually incorrect analogies would not be taken as effective explanations for humans, we only report qualitative results on the factually correct ones in the following analysis.

The distribution of dimensions in 5-point Likert scale can be visualized with the boxplots in Figure 5. Overall, the generated analogies show good quality in most qualitative dimensions except *Structural Correspondence* and *Relational Similarity*. The experts consider that the analogies are easy to understand and involve familiar everyday concepts, which indicates these explanations are of relatively low cognitive load. To be concrete about how the explanations differ in quality, we show examples of scoring 1, 3, 5 for dimensions in 5 point Likert scale in Table 4. Note that we

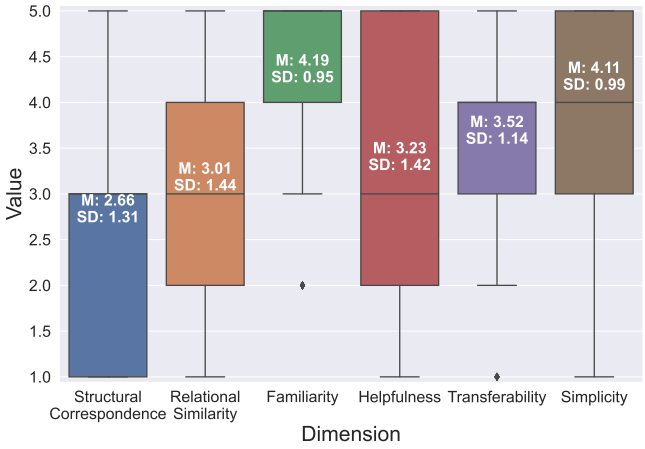


Figure 5: Box plot illustrating the distribution of the different dimensions considered in our study. All dimensions were measured on a 5-point Likert scale. For all dimensions, 1 indicates a poor quality while 5 indicates a good quality. M and SD represent mean and standard deviation respectively.

do not expand on examples for *Factual Correctness*, *Syntactic Correctness*, and *Misunderstanding*, which are trivial.

To further investigate how qualitative dimensions affect the perceived helpfulness of analogy-based explanations, we calculated Spearman rank-order correlation coefficients between *Helpfulness* and the other Likert-based dimensions. We found a significant positive correlation between all dimensions and *Helpfulness*: *Structural Correspondence*, $r(247) = 0.191$, $p = 0.003$; *Relational Similarity*, $r(247) = 0.374$, $p = 0.000$; *Familiarity*, $r(247) = 0.312$, $p = 0.000$; *Transferability*, $r(247) = 0.445$, $p = 0.000$; *Simplicity*, $r(247) = 0.467$, $p = 0.000$. This confirms that our qualitative dimensions are substantially indicative of their perceived helpfulness. Our findings suggest that if we ensure the generated explanations are of high quality across these dimensions, they have a higher likelihood of being helpful in understanding the target sentence.

5.3.2 Comparison between Different Tasks

Among 410 annotations, 174 cases are generated from calorie level classification (CLC) task, while 236 cases are generated from scene classification (SC) task. According to the results, 109 and 138 cases are identified as both valid and factually correct for CLC and SC tasks, respectively. We compared the difference between the quality of analogies generated with the calorie task and places task. We found a significant difference ($\alpha = 0.05$) on the assessed *Relational Similarity* ($H(1) = 7.54$, $p = 0.006$) with a Kruskal-Wallis H-test. Post-hoc Mann-Whitney tests further show that the *Relational Similarity* of analogy-based explanations generated from SC task is significantly better than the counterparts from CLC task. However, no significant difference exists in the other qualitative dimensions.

The reason for such a phenomenon may be that the relationship between "concept" and "label" in the SC task is more explicit than in the CLC task. This may make it easier for participants to generate analogy-based explanations while keeping similar relationships. However, such good ana-

logical properties do not translate to higher perceived *Helpfulness*. This indicates that the interplay between qualitative dimensions and perceived helpfulness may be complex. Better quality on a single dimension (*Relational Similarity* here) may not necessarily lead to a better understanding.

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

Our first study showed that our proposed method can generate conceptually high-quality analogy-based explanations when non-expert workers are involved in the collection process. Besides evaluating analogy-based explanations with qualitative dimensions, it is also important to check how effective they are when assisting users in decision making in practice. Thus, we conducted an empirical study of human-AI decision making in medical analysis. In this section, we first present our hypotheses and experimental setup, which had all been preregistered before any data collection.[5] Then, we show the experimental results. Finally, we discuss the findings and implications of this study. This study was approved by the human research ethics committee of our institution.

6.1 Hypotheses

It is still unknown how analogies will affect user understanding of concept-level explanations and how analogy-based explanations affect user reliance on AI systems. Based on our findings from Study I and findings from existing work (Nashon, 2004; Geelan, 2012; Mozzer & Justi, 2012), analogies have proven effective in aiding users in understanding new knowledge. Little has been done to build an empirical understanding of the effectiveness of analogies in real-world decision-making tasks where concept-level explanations are employed (He et al., 2023). Addressing this research gap, we hypothesize that analogies can help users better understand AI systems, and that such an improved understanding will further help users rely on AI systems more appropriately.

H1: Using analogy-based explanations can help users better understand AI systems, compared to conventional concept-based explanations.

H2: Using analogy-based explanations can facilitate appropriate reliance on AI systems, compared to conventional concept-based explanations.

Analogies have proven to be effective in helping humans understand new knowledge and reduce the cognitive load for learning new knowledge (Richland & Hansen, 2013). While analogies can help improve users' understanding, the additional analogical inference requires more effort, which may be time-consuming. Therefore, we hypothesize that users can maintain a similar team performance and be more efficient in their decision making when engaging with analogy-based explanations when they deem it to be necessary (*i.e.,* on demand).

H3: Analogy-based explanations can reduce the perceived cognitive load of users in their decision making process.

H4: Providing analogy-based explanations on demand can improve users' efficiency in their decision making process.

5. https://osf.io/jm3ap

6.2 Task

In our study, we selected a real-world medical diagnosis scenario — skin cancer detection based on skin lesions as a test bed to verify the effectiveness of analogy-based explanations in human-AI decision making. All task data are selected from the HAM10000 (Tschandl et al., 2018) dataset. In this task, given an image of a pigmented skin lesion, users are asked to decide whether the shown image depicts a '**malignant**' or '**benign**' skin lesion. The rationale for selecting the skin cancer detection task is three-fold: (1) This is a realistic scenario for human-AI collaboration, where humans are designated to make final decisions due to accountability concerns. (2) Medical concepts in this task are relatively challenging for laypeople to digest, which fits our motivation of providing analogy-based commonsense explanations that can be leveraged and used to communicate the explanations to laypeople. (3) There is a substantial need for AI assistance to help doctors and medical experts check increasingly large volumes of images. Thus, the setting we chose is realistic and aligned with real-world needs.

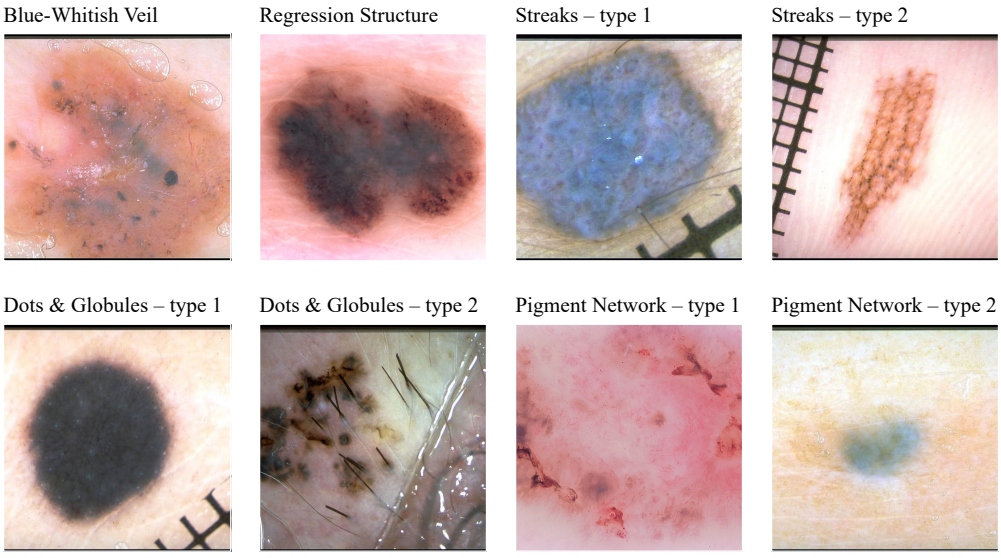


Figure 6: The overview of medical concepts shown to participants in Study II.

Medical Concepts. In our study, we followed Yuksekgonul *et al.* to adopt eight medical concepts to help users diagnose skin cancer based on their assessment of **malignant** versus **benign** skin lesions (Yuksekgonul et al., 2023). The eight concepts are: Blue-Whitish Veil, Regular Dots & Globules, Irregular Dots & Globules, Regression Structures, Irregular Streaks, Regular Streaks, Atypical Pigment Network, and Typical Pigment Network. Note that these concept names contain words like "Irregular" and "Atypical", which can clearly indicate their correlation to the model's prediction (*i.e.,* benign and malignant) — simplifying an otherwise complex decision making task. To test the learning effect potentially stemming from concept-based explanations, we replaced such hints with the abstractions of "type 1" and "type 2". In our study, we provided participants with an overview figure illustrating the eight different medical concepts to aid their decision making (shown in Fig 6). For each concept, we provided an image of an example skin lesion to swiftly illustrate the concept and help user understanding. To help participants remember and rely on these concepts

along with concept-level explanations in their decision making, we provided a button (cf. Figure 9) below the concept-level explanations, that triggers a pop-up window containing the overview of medical concepts.

Selection of Tasks. To ensure diversity in the selected tasks and to cover the use of different medical concepts, we selected 14 tasks based on seven fine-grained categories in the HAM10000 dataset. To faithfully reflect the performance of the AI system used, we selected tasks based on performance of the post-hoc concept bottleneck model (Yuksekgonul et al., 2023) on the HAM10000 dataset.

Table 5: Descriptive statistics of the HAM10000 dataset and AI performance across the seven categories in the dataset.

Category	Label	#Tasks	Error Rate	Selected Task
Benign keratosis-like lesions	benign	220	9.1%	1 correct, 1 wrong
Dermatofibroma	benign	23	4.3%	2 correct
Melanoma	malignant	223	35.4%	1 correct, 1 wrong
Vascular lesions	benign	28	10.7%	2 correct
Basal cell carcinoma	malignant	103	28.2%	1 correct, 1 wrong
Melanocytic nevi	benign	1,341	2.9%	1 correct, 1 wrong
Actinic keratoses	benign	65	10.8%	2 correct

First, we generate model predictions on the validation set of the HAM10000 dataset (same split as (Yuksekgonul et al., 2023)). Then, based on the performance of each category (shown in Table 5) and the sample size of each category, we selected 14 tasks (10 with correct predictions, 4 with wrong predictions). In our study, the accuracy of the AI system is 71.4% (10 / 14).

Pilot Study. To understand how capable non-expert crowd workers are in this task, we recruited 20 participants from Prolific. The Prolific platform has been shown to be a reliable source for participant recruitment in similar XAI studies over the last few years (Chromik et al., 2021; Robbemond et al., 2022; He et al., 2023) and has a growing reputation as a suitable platform for human subjects research across different scientific domains (Adams et al., 2020; Douglas et al., 2023). Each participant in our study received 2 GBP (8 GBP per hour [6]) for working on the 14 trial tasks independently. We filtered out three outliers who spent less than 5 mins on the tasks. On average, the remaining 17 participants achieved an accuracy of 59.2% on 14 tasks, which is worse than the AI performance (71.4 %). Thus, the introduction of the AI system in the decision making process within our study can be beneficial to achieve better team performance.

6.3 Experimental Setup

6.3.1 Experimental Conditions

To answer the above research questions, we designed a between-subjects study consisting of four experimental conditions. Example explanations in different conditions are shown in Table 6. Participants in all these conditions saw the systems' advice, but the five conditions differed in the inclusion of additional explanations.

6. This was rated as a 'good' hourly rate by the platform at the time of running the study.

- *Control*: no additional explanation.
- *Concept*: concept-based explanation from post-hoc Concept Bottleneck Models (Yuksekgonul et al., 2023), similar to ExAID (Lucieri et al., 2022) (see Table 6).
- *Concept-Imp*: we provide more details about how important each concept is, which is the target domain in our proposed analogy-based explanations (see Table 6).
- *Analogy*: analogy-based explanation for each concept (see Table 6).
- *Analogy-OD*: We show the same explanations as the *Concept-Imp* condition. When users require further clarification and indicate this by clicking the Clarify button, we provide an analogy on demand.

Table 6: Example of explanations in different conditions. In condition *Analogy-OD*, when the "clarify" button is clicked, the analogy is shown on another line for the sake of clarity.

Condition	**Explanation Type**
Concept	absence of Streaks - type 1: strong evidence
Concept-Imp	Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign.
Analogy	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.
Analogy-OD	Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign. Clarify
Concept	observation of Dots & Globules - type 1: moderate evidence
Concept-Imp	Dots & Globules - type 1 can typically be associated with benign.
Analogy	Dots & Globules - type 1 can typically be associated with benign. This is like fish can typically be associated with oceans, but it's also possible to associate fish with rivers.
Analogy-OD	Dots & Globules - type 1 can typically be associated with benign. Clarify

Explanation Generation. The AI system in our study is based on a post-hoc concept bottleneck model (Yuksekgonul et al., 2023). We trained the post-hoc concept bottleneck model following its official implementation.[7] As tested by Yuksekgonul et al., it can provide concept-based explanations aligned with medical knowledge. The post-hoc concept bottleneck model first learned concept activation vector for skin lesions based on concept banks from the Derm7pt (Kawahara et al., 2018) dataset. Then a linear classifier is trained to make binary predictions. Based on the linear layer weight $\mathbf{w} \in \mathbb{R}^k$ and concept activation vector $\mathbf{c} \in \mathbb{R}^k$ for each image, we generate concept-level explanations based on the contribution of each concept. For concept $c_i, i \in [1, k]$, the contribution to final prediction is $s_i = w_i * c_i$. To generate simple heuristics-based concept-level explanations (*Concept* condition), we use two thresholds to identify the importance of each concept:

$$evidence\ strength = \begin{cases} strong, & |s_i| >= \epsilon_1 \\ moderate, & \epsilon_2 <= |s_i| < \epsilon_1 \\ ignore, & otherwise. \end{cases} \quad (1)$$

7. https://github.com/mertyg/post-hoc-cbm

In our study, we set $\epsilon_1 = 0.5, \epsilon_2 = 0.1$. A positive value for contribution s_i indicates that the absence/presence of concept c_i helps predict that the lesion is **`malignant`**, while a negative value indicates the tendency to predict **`benign`**. Following the templates used in Table 1 for *Concept-Imp* condition, we generate the target domain of analogy-based explanations. To account for errors caused by the absence of concepts, we further clarify the target domain with relation to the alternative class prediction. Instead of claiming "absence of [`concept`] is definitely a sign of [model prediction]", we use "[`concept`] is definitely a sign of [alternative option]. Thus, absence of `concept` helps make prediction of [model prediction]." For example, *Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign.* To increase clarity and reduce scope for misinterpretations caused by using double negative expressions (*e.g., Absence of [`concept`] seldom found at benign*), we do not provide any explanations in the form of double negative expressions.

To provide high-quality analogies, we generate analogy-based explanation with two stages. In the first stage, based on the evaluation results of Section 5.3, we only consider analogies which are syntactically correct, factually correct, and easy to understand (Simplicity > 3). In the second stage, we manually curated and selected the analogies reserved, which resulted in 37 valid analogies: "Definite Sign Of" (11), "Typically Associated With" (9), "Seldom Found At" (9), "Contradict With" (8). Based on the contribution of each concept s_i and the sign of predictions, we map each concept to a template. Then we generate the analogies by randomly sampling valid candidates in each template.

6.3.2 Measures and Variables

All variables analyzed in this work are summarized in Table 7.

Dependent Variables. To assess the learning effect for participants (**H1**), we calculated the F1 measures with respect to benign and malignant cases, respectively. In the post-task questionnaire, we asked participants to select the concepts positively associated with benign and malignant labels. To analyze the impact of analogy-based explanations on user reliance, we adopted the **Switch Fraction** metrics as reliance measures (Yin et al., 2019; Zhang et al., 2020). To assess the appropriate reliance (**H2**), we followed Max *et al.* (Schemmer et al., 2022) to adopt *Relative positive AI reliance* (**RAIR**) and *Relative positive self-reliance* (**RSR**) metrics. The two measures assessed users' appropriate reliance from two dimensions (*i.e.,* appropriate adoption of AI advice and insistence on their own decision), which can help analyze the dynamics of reliance. To provide an overview of participants' performance under initial disagreement, we considered **Accuracy-wid** (*i.e.,* accuracy with initial disagreement). To analyze the impact of analogy-based explanations on cognitive load (**H3**), we adopted NASA-TLX questionnaire (Colligan et al., 2015). For the analysis of decision making efficiency (**H4**), we measured the average time spent on each decision task, which is measured in seconds.

Covariates and Trust. As pointed out by prior studies (Zhang et al., 2022), user domain expertise also affects their trust and reliance on the AI system. Thus, we assessed participants' general medical expertise by gathering responses on a 5-point Likert-scale ranging from *1:* to *5:* ("*To what extent are you knowledgeable about medical diagnosis?*"), and specific expertise on skin cancer detection task ("Do you have any experience or knowledge about skin cancer?") on a 5-point Likert-scale ranging from *1:* to *5:*. We accounted for the effect of participants' affinity with technology through the Affinity for Technology Interaction Scale (ATI) (Franke et al., 2019). To assess partic-

Table 7: The different variables considered in our experimental study. "DV" refers to the dependent variable. **RAIR**, **RSR**, and **Accuracy-wid** are indicators of appropriate reliance.

Variable Type	Variable Name	Value Type	Value Scale
Learning Effect (DV)	F1 of malignant concepts	Continuous	[0.0, 1.0]
	F1 of benign concepts	Continuous	[0.0, 1.0]
Performance (DV)	Accuracy	Continuous, Interval	[0.0, 1.0]
	Accuracy-wid	Continuous	[0.0, 1.0]
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous	[0.0, 1.0]
	RAIR	Continuous	[0.0, 1.0]
	RSR	Continuous	[0.0, 1.0]
Trust (DV)	TiA-Reliability/Competence	Likert	5-point, 1: poor, 5: very good
	TiA-Understanding/Predictability	Likert	5-point, 1: poor, 5: very good
	TiA-Intention of Developers	Likert	5-point, 1: poor, 5: very good
	TiA-Trust in Automation	Likert	5-point, 1:strong distrust, 5: strong trust
Cognitive Load (DV)	Mental Demand	Likert	-7: very low, 7: very high
	Physical Demand	Likert	-7: very low, 7: very high
	Temporal Demand	Likert	-7: very low, 7: very high
	Performance	Likert	-7: Perfect, 7: Failure
	Effort	Likert	-7: very low, 7: very high
	Frustration	Likert	-7: very low, 7: very high
Efficiency (DV)	Time of decision making	Continuous	[0.0, $+\infty$] (s)
Covariates	ATI	Likert	6-point, 1: low, 6: high
	TiA-Propensity to Trust	Likert	5-point, 1: tend to distrust, 5: tend to trust
	TiA-familiarity	Likert	5-point, 1: unfamiliar, 5: familar
	Medical diagnosis expertise	Likert	5-point, 1: no expertise, 5: extensive expertise
	Skin cancer expertise	Likert	5-point, 1: no expertise, 5: extensive expertise
Other	Helpfulness of Explanation	Likert	5-point, 1: unhelpful, 5: helpful
	Helpfulness of Analogy	Likert	5-point, 1: unhelpful, 5: helpful
	Experience	Category	{Yes, No}
	Confidence	Likert	5-point, -2: unconfident, 2: confident

ipants' subjective trust in the AI system, we adapted the Trust in Automation (TiA) questionnaire (Körber, 2018) to the context of the "AI system". We included six subscales from the TiA questionnaire: Reliability/Competence (TiA-R/C), Understanding/Predictability (TiA-U/P), Propensity to Trust (TiA-PtT), Familiarity (TiA-Familiarity), Intention of Developers (TiA-IoD), and Trust in Automation (TiA-Trust).

Other Variables. Meanwhile, for conditions with explanations (analogies), we also assessed the helpfulness of explanations (analogies) with the question, "*To what extent did you find the explanations (analogies) helpful to make decisions?*" Responses were gathered on a 5-point Likert scale from *1* to *5* corresponding to the labels *unhelpful*, *somewhat unhelpful*, *neutral*, *somewhat helpful*, *helpful*. We further collected the reasons (open text) for perceived helpfulness with "*Why did you find the explanation (analogies) to be helpful or not helpful?*" For participants in *Analogy* and *Analogy-OD* conditions, we collected their comments and feedback (open text) to the analogies with: "*Please share any comments, remarks or suggestions regarding the use of analogies to explain the medical concepts.*" For a deeper analysis of our results, we collected responses from participants regarding their perceived user experience ("*Have you ever had this or seen it on others?*") and confidence ("*How confident are you with your decision?*") on 5-point Likert-scales along with each trial task.

6.3.3 Participants

Sample Size Estimation. Before recruiting participants, we computed the required sample size in a power analysis for a between-subjects study using G*Power (Faul et al., 2009). We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.0125$ (*i.e.,* $\frac{0.05}{4}$, due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and the consideration of 5 different experimental conditions. This resulted in a required sample size of 265 participants. We thereby recruited 486 participants from the crowdsourcing platform Prolific[8], in order to accommodate potential exclusion.

Compensation. All participants were rewarded with £2, amounting to an hourly wage of £8 (estimated completion time was 15 minutes). In addition to this, we rewarded participants with extra bonuses of £0.1 for every correct decision in the 14 trial cases. Such monetary bonuses have been shown to motivate and encourage participants to exert genuine effort in decision making tasks, which is also a contextual requirement to encourage appropriate system reliance (Lee & See, 2004).

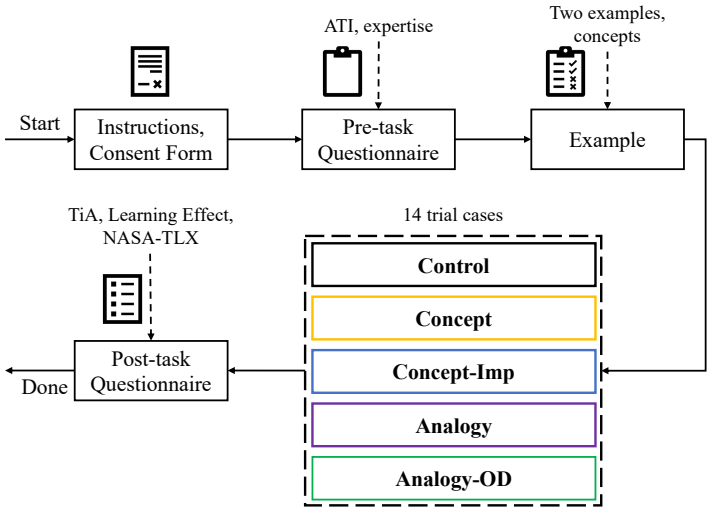


Figure 7: Illustration for the decision making setup.

Filter Criteria. All participants were proficient English speakers above the age of 18, and they had finished more than 40 tasks and maintained an approval rate of at least 90% on the Prolific platform. We excluded participants from our analysis if they failed at least one attention check or any missing response. The resulting sample of 280 participants had an average age of 37 ($SD = 13.0$) and a gender distribution (51.4% female, 48.6% male).

6.3.4 Procedure

The entire procedure of our study is illustrated in Figure 7. All participants first read the same basic instructions and consent forms. Next, participants were asked to complete a pre-task questionnaire to measure their affinity for technology interaction and expertise in medical diagnosis and skin cancer. To onboard participants on the skin cancer detection task, and help them understand the

8. https://www.prolific.co

labels **`malignant`** and **`benign`**, we provided them with two examples of benign and malignant skin lesions before they began working with the tasks. After the examples, all participants excluding the *Control* condition obtain an overview of the medical concepts relevant to our study (cf. Figure 6).

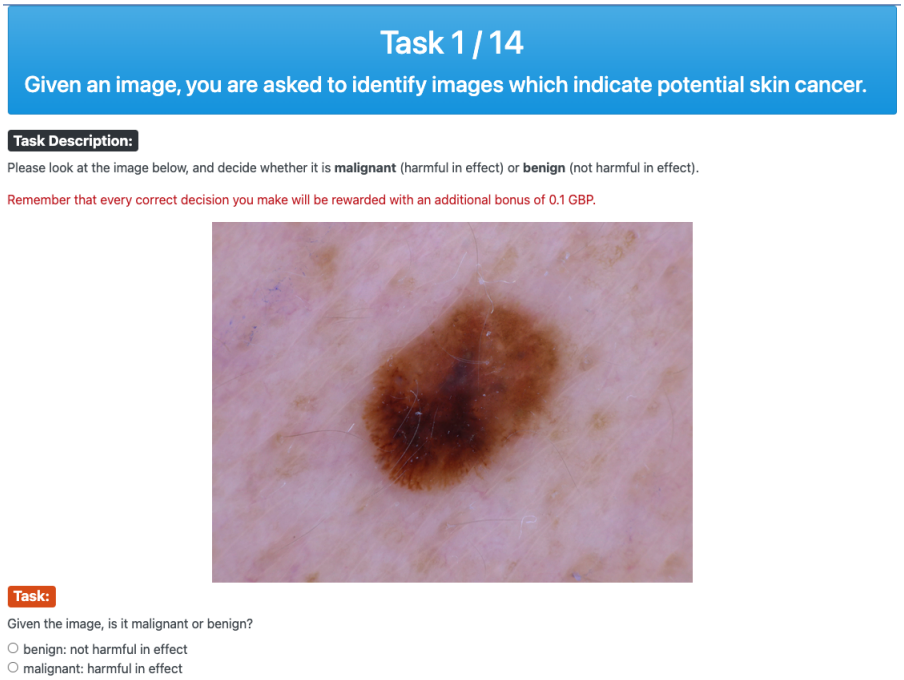


Figure 8: Screenshot of the task interface in the first stage of decision making.

Next, participants across all conditions worked on 14 trial tasks. In each trial task, we followed a two-stage decision making process (Green & Chen, 2019b, 2019a; Dietvorst et al., 2018). In the first stage, participants worked on the task without any extra information (one example shown in Figure 8). In the second stage, AI advice and explanations were provided, and participants had a chance to alter their decision (one example shown in Figure 9). After the task phase, post-task questionnaires were adopted to assess their cognitive load, their trust in the AI system, and criteria of making final decisions (open text). For all participants excluding the *Control* condition, we assessed their learning effect through a specific question ("*Please select the concepts positively associated with malignant/benign skin lesions.*"), their perceived helpfulness of explanations, and open text reasons for the perceived helpfulness. Participants in condition *Analogy* and *Analogy-OD* were additionally asked to report their perceived helpfulness of the analogies and to provide rationales/feedback in open text fields.

Attention Checks. To ensure the reliability of participants' responses, three attention check questions were placed at the pre-task questionnaire (ATI), task phase, and post-task questionnaire (Trust in automation). Each attention check asked participants to select a specific option (Marshall & Shipman, 2013; Gadiraju et al., 2015).

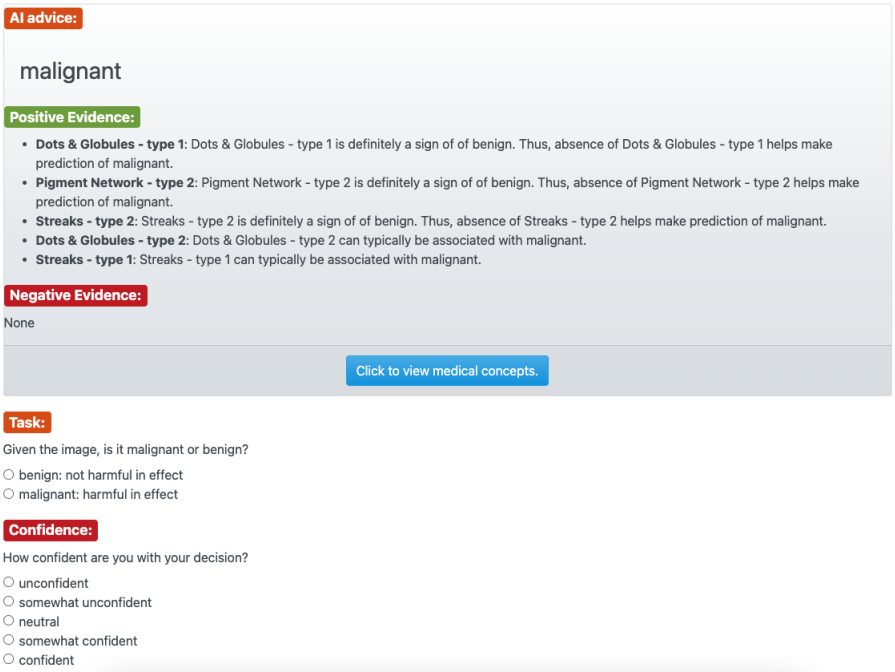


Figure 9: Screenshot of the task interface in the second stage of decision making for the *Concept-Imp* condition.

6.4 Experimental Results

6.4.1 Descriptive Statistics

In our analysis, we only consider participants who passed all attention checks. Participants were distributed in a balanced fashion over the four experimental conditions as follows: 55 (*Control*), 55 (*Concept*), 55 (*Concept-Imp*), 53 (*Analogy*), 62 (*Analogy-OD*).

Distribution of Covariates. The covariates' distribution is as follows: *ATI* ($M = 3.87$, $SD = 0.87$, 6-point Likert scale, and *1: low, 6: high*), *Medical Diagnosis Expertise* ($M = 1.47$, $SD = 0.81$, 5-point Likert scale, and *1: no expertise, 5: extensive expertise*), *Skin Cancer Expertise* ($M = 1.59$, $SD = 0.81$, 5-point Likert scale, and *1: no expertise, 5: extensive expertise*), *TiA-Propensity to Trust* ($M = 2.76$, $SD = 0.57$, 5-point Likert scale, *1: tend to distrust*, *5: tend to trust*), *TiA-Familiarity* ($M = 2.31$, $SD = 1.05$, 5-point Likert scale, *1: unfamiliar*, *5: familiar*).

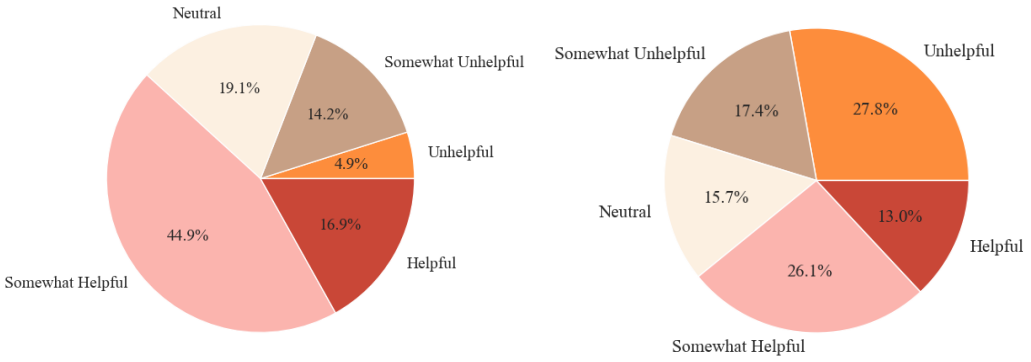
Performance Overview. On average across all conditions, participants achieved an accuracy of 63.3% ($SD = 0.11$), which is worse than the AI accuracy (71.4%). The agreement fraction was found to be 0.79 ($SD = 0.16$) while the switch fraction was 0.57 ($SD = 0.30$). With these measures, we confirm that in the face of disagreement with AI advice, participants in our study did not always switch to AI advice or blindly rely on the AI system. As all dependent variables are not normally distributed, we used non-parametric statistical tests to verify our hypotheses.

Performance Per Task. Considering the 14 tasks in our study, we calculated the accuracy and confidence based on all valid participants. The results are shown in Table 8. Generally, the accuracy

Table 8: Accuracy, experience, and confidence for the 14 tasks used in our study. "Acc" and "Con" refer to accuracy and confidence. The subscript $_i$ and $_f$ refer to the initial and final decisions, respectively. "Experience ratio" refers to the ratio of participants who reported seeing similar skin lesions in their life.

Task ID	Acc_i	Acc_f	Con_i	Con_f	Experience ratio	Ground Truth	AI correctness
ISIC-0033051	0.864	0.954	0.52	1.07	0.05	malignant	✓
ISIC-0032013	0.857	0.950	0.21	0.91	0.14	benign	✓
ISIC-0027107	0.657	0.889	0.00	0.60	0.07	benign	✓
ISIC-0028763	0.632	0.864	-0.01	0.57	0.09	benign	✓
ISIC-0034271	0.557	0.832	0.01	0.57	0.09	benign	✓
ISIC-0027665	0.554	0.818	-0.06	0.34	0.10	benign	✓
ISIC-0034155	0.443	0.793	-0.04	0.48	0.04	malignant	✓
ISIC-0033790	0.539	0.771	0.00	0.29	0.05	benign	✓
ISIC-0028076	0.457	0.750	-0.06	0.24	0.05	benign	✓
ISIC-0032557	0.043	0.368	0.93	0.30	0.05	benign	✓
ISIC-0029323	0.525	0.304	-0.05	0.29	0.05	malignant	×
ISIC-0032269	0.386	0.282	0.00	0.38	0.06	malignant	×
ISIC-0024924	0.379	0.186	0.26	0.61	0.14	benign	×
ISIC-0029260	0.311	0.100	-0.03	0.71	0.04	benign	×

of participants increased after being exposed to correct AI advice and decreased after being exposed to wrong AI advice. Overall, participants showed higher confidence after being exposed to AI advice. The only exception is task ISIC-0032557, where participants showed less confidence in their final decision. Among all tasks, most participants indicated that they never saw the skin lesion image on themselves or on someone they know. This is illustrated by the low experience ratios observed across all tasks (cf. Table 8).



(a) Helpfulness of explanations.

(b) Helpfulness of analogies.

Figure 10: Distribution of perceived helpfulness of explanations and analogies.

Helpfulness of Explanations and Analogies. In the post-task questionnaire, participants were asked to report their perceived helpfulness of explanations (for conditions with explanations) and

perceived helpfulness of analogies (for condition *Analogy* and *Analogy-OD*). The distributions of perceived helpfulness are shown in Figure 10. Overall, 61.8% participants reported positive attitudes towards the provided concept-based explanations. Meanwhile, 39.1% participants in condition *Analogy* and *Analogy-OD* found that the provided analogies were helpful to some extent.

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

To analyze **H1**, we compared the F1 of learned concepts for the benign and malignant skin lesions. Considering that five concepts are positively correlated with label "malignant" and three concepts are positively correlated with label "benign", we adopted the weighted average F1 measures ($\text{F1}_{avg} = \frac{5}{8}\text{F1}_{malignant} + \frac{3}{8}\text{F1}_{benign}$) to assess user understanding of the AI system. The Kruskal-Wallis H-test results are: $H(279) = 1.79, p = 0.616$. The mean and std are: $M \pm SD$(*Concept*) $= 0.55 \pm 0.20$; $M \pm SD$(*Concept-Imp*) $= 0.58 \pm 0.19$; $M \pm SD$(*Analogy*) $= 0.56 \pm 0.21$; $M \pm SD$(*Analogy-OD*) $= 0.52 \pm 0.22$. No significant difference was found to suggest a learning effect. Thus, we did not find empirical support for **H1** in our study.

Table 9: Kruskal-Wallis H-test results for performance-based and reliance-based dependent variables across five conditions. † and †† indicate the effect of variable is significant at the level of 0.05 and 0.0125, respectively.

Dependent Variables	Accuracy	Agreement Fraction	Switch Fraction	Accuracy-wid	RAIR	RSR
H	2.18	11.03	8.42	15.81	12.77	6.16
p	.703	**.026**†	.078	**.003**††	**.012**††	.187
M(*Control*)	0.63	0.83	0.55	0.50	0.53	0.32
M(*Concept*)	0.64	0.76	0.51	0.49	0.49	**0.48**
M(*Concept-Imp*)	**0.65**	**0.83**	**0.67**	**0.65**	**0.70**	0.35
M(*Analogy*)	0.62	0.77	0.55	0.51	0.54	0.39
M(*Analogy-OD*)	0.63	0.78	0.58	0.55	0.58	0.43

To verify **H2**, we used Kruskal-Wallis H-tests to compare participants' performance across all conditions. The results are shown in Table 9. Among the dependent variables we analyzed across the conditions, we found that participants exhibited significant differences in their appropriate reliance. Through post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125, we found that: (1) participants in condition *Concept-Imp* showed significantly higher **Accuracy-wid** than participants in conditions *Control*, *Concept*, *Analogy*; (2) participants in condition *Concept-Imp* showed a significantly higher **RAIR** than participants in conditions *Control*, *Concept*, *Analogy*. The results indicate that the target domain of our analogy-based explanation can help users appropriately rely on AI systems, which is mainly by addressing the under-reliance. However, this may also trigger over-reliance on the AI system, which is reflected by the relatively low **RSR** in comparison with other conditions. At the same time, we found that the analogies did not have the expected effect in facilitating appropriate reliance. However, our results suggest that providing analogies on demand can have a better impact on appropriate reliance (non-significant). Thus, we did not find empirical support for **H2** in our study.

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

To analyze **H3** for the impact of experimental conditions on cognitive load, we conducted a one-way ANOVA. Our findings are shown in Table 10. Overall, participants who received explanations

Table 10: ANOVA test results for user cognitive load across five conditions. "Avg" refers to the average cognitive load among six dimensions. † and †† indicate the effect of variable is significant at the level of 0.05 and 0.0125, respectively.

Cognitive Load	Avg	Mental	Physical	Temporal	Performance	Effort	Frustration
F	5.81	7.01	0.65	0.67	1.08	3.03	1.98
p	**.000**††	**.000**††	.625	.616	.368	**.018**†	.098
M(*Control*)	-2.25	-0.02	-5.25	-4.35	-1.04	0.87	-3.69
M(*Concept*)	-1.42	2.05	-4.45	-4.33	-1.11	1.85	-2.53
M(*Concept-Imp*)	-0.88	2.62	-4.89	-3.85	-0.13	2.82	-1.85
M(*Analogy*)	-1.07	2.13	-4.53	-3.85	-0.32	2.04	-1.91
M(*Analogy-OD*)	-0.98	2.95	-4.68	-4.32	0.11	2.35	-2.31

reported a higher perceived cognitive load. Through post-hoc Turkey HSD tests using a Bonferroni-adjusted alpha level of 0.0125, we found a significant difference: For both average cognitive load and mental demand, *Control*< *Concept*, *Analogy*, *Concept-Imp*, *Analogy-OD*. Thus, we did not find support for **H3**.

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

To analyze **H4**, we compared participants' task completion time (units: seconds) in the 14 tasks with Kruskal-Wallis H-test. The results show a significant difference: $H(279) = 23.73, p = .000$. Post-hoc Mann-Whitney test results showed that participants who received explanations spent significantly more time making decisions: *Control* < *Concept*, *Analogy*, *Concept-Imp*, *Analogy-OD*. $M \pm SD$(*Control*) $= 462 \pm 309$; $M \pm SD$(*Concept*) $= 548 \pm 210$; $M \pm SD$(*Concept-Imp*) $= 575 \pm 209$; $M \pm SD$(*Analogy*) $= 574 \pm 242$; $M \pm SD$(*Analogy-OD*) $= 658 \pm 341$.

For a more fine-grained analysis, we calculated the average time spent on each correct/wrong decision per person for each user group (shown in Table 11). With Kruskal-Wallis H-test, we compared the average time per correct/wrong decision. The post-hoc Mann-Whitney test results are still consistent with the overall decision making efficiency: participants who received explanations spent significantly more time making decisions. Thus **H4** is not supported by our experimental results.

Table 11: Time per decision (in seconds). The "Decision-level" is calculated by average on all decisions in each condition. The "Human-level" is calculated by the average of all humans in each condition.

Granularity	Decision-level		Human-level	
Correctness	$M \pm SD$ (Correct)	$M \pm SD$ (Wrong)	$M \pm SD$ (Correct)	$M \pm SD$ (Wrong)
Control	34.19 ± 39.85	30.92 ± 31.06	34.17 ± 23.71	30.16 ± 21.31
Concept	37.90 ± 29.74	41.35 ± 32.17	37.83 ± 14.82	41.67 ± 19.00
Concept-Imp	39.86 ± 32.34	43.25 ± 39.69	40.78 ± 16.18	41.84 ± 20.46
Analogy	40.71 ± 35.37	41.51 ± 31.49	40.84 ± 19.79	42.78 ± 19.92
Analogy-OD	47.21 ± 59.91	46.66 ± 51.51	46.99 ± 27.13	46.63 ± 32.16

6.5 Exploratory Analysis

6.5.1 The Impact of First Impression

Prior work has demonstrated the significant impact of first impressions of AI systems in shaping user trust and reliance (Nourani et al., 2020a, 2020b; Tolmeijer et al., 2021). We thereby analyzed the potential impact of task ordering and the accuracy of AI advice. To this end, we grouped participants according to the AI accuracy within the first five tasks. Participants who either never encountered wrong AI advice or did so only once are grouped within "*Good First Impression*", and others are grouped within "*Bad First Impression*." We compared participants' performance and reliance on AI systems with Kruskal-Wallis H-test. We found no significant difference, suggesting that first impressions of the AI system did not have an effect within our study.

6.5.2 Analysis of Trust and Covariates

An ANCOVA analysis across the experimental conditions revealed no significant difference in the perceived trust of the participants in the AI system (TiA). For all covariates, we conducted Spearman rank-order tests with dependent variables.

The impact of propensity to trust. As shown in Table 12, TiA-Propensity to Trust significantly affected user trust in the AI system. With Spearman rank-order test, we found that TiA-Propensity to Trust positively correlated with all trust measures: TiA-R/C, $r(278) = .650$, $p = .000$; TiA-U/P, $r(278) = .344$, $p = .000$; TiA-IoD, $r(278) = .283$, $p = .000$; TiA-Trust, $r(278) = .677$, $p = .000$. Meanwhile, TiA-Propensity to Trust also showed significant positive correlation with performance and appropriate reliance measures: Agreement Fraction, $r(278) = .227$, $p = .000$; Switch Fraction, $r(278) = .220$, $p = .000$; RAIR, $r(278) = .183$, $p = .002$; RSR, $r(278) = -.216$, $p = .000$. It is worth noting that the general propensity to trust positively correlated with all trust dimensions, and **Agreement Fraction**, **Switch Fraction**, **RAIR**, but negatively correlated with **RSR**. Thus, participants with a higher propensity to trust tend to rely more on the AI system after the XAI is provided. However, this addresses under-reliance to some extent but also causes over-reliance.

Table 12: ANCOVA test results corresponding to user trust across experimental conditions. † and †† indicate the effect of variable is significant at the level of 0.05 and 0.0125, respectively.

Dependent Variables	**TiA-R/C**			**TiA-U/P**			**TiA-IoD**			**TiA-Trust**		
Variables	F	p	η^2	F	p	η^2	F	p	η^2	F	p	η^2
Experimental Condition	1.02	.397	0.01	0.45	.769	0.01	4.47	**.002**††	0.05	3.06	**.017**†	0.02
Medical Expertise	0.47	.493	0.00	3.05	.082	0.01	0.03	.868	0.00	0.22	.639	0.00
Skin Cancer Expertise	2.97	.086	0.01	0.09	.766	0.00	1.64	.201	0.01	0.01	.927	0.00
ATI	0.58	.448	0.00	2.10	.149	0.01	3.68	.056	0.01	2.03	.155	0.00
TiA-Propensity to Trust	182.14	**.000**††	0.39	31.72	**.000**††	0.10	35.53	**.000**††	0.11	223.51	**.000**††	0.44
TiA-Familiarity	1.58	.210	0.00	0.22	.641	0.00	0.52	.471	0.00	3.35	.068	0.01

Other covariates. For TiA-Familiarity, we found a strong positive correlation with some trust measures: TiA-R/C, $r(278) = .232$, $p = .000$; TiA-Trust, $r(278) = .286$, $p = .000$. For ATI, we found a strong positive correction with TiA-Trust, $r(278) = .149$, $p = .012$. However, according to the results of ANCOVA analysis of trust (Table 12), the impact of ATI and TiA-Familiarity is insignificant. No strong correlation was found for the covariates of expertise in medical diagnosis expertise. Meanwhile, We found a strong negative correlation with the skin cancer expertise and

Switch Fraction: $r(278) = -.175$, $p = .003$. Among 280 participants, 166 reported zero skin cancer experience or expertise, which confirms that most participants are laypeople.

6.5.3 Impact of User Opinions towards Explanations And Analogies

Opinion towards explanations. To understand how users' perceived helpfulness of explanations affects user trust and reliance on the AI system, we conducted the Spearman rank-order test for participants in the condition *Concept*, *Concept-Imp*, *Analogy*, and *Analogy-OD*. The results show that, the perceived helpfulness of explanations is positively correlated with user trust: **TiA-R/C**, $r(223) = .400, p = .000$; **TiA-U/P**, $r(223) = .397, p = .000$; **TiA-IoD**, $r(223) = .249, p = .000$; **TiA-Trust**, $r(223) = .407, p = .000$. However, there is no significant correlation between the perceived helpfulness of explanations and reliance-based dependent variables.

Opinion towards analogies. Similarly, to understand how users' perceived helpfulness of analogies affects user trust and reliance on the AI system, we conducted the Spearman rank-order test for participants in condition *Analogy* and *Analogy-OD*. The results show that the perceived helpfulness of analogies is positively correlated with user trust: **TiA-R/C**, $r(113) = .303, p = .001$; **TiA-U/P**, $r(113) = .290, p = .002$; **TiA-IoD**, $r(113) = .368, p = .000$; **TiA-Trust**, $r(113) = .297, p = .001$. Meanwhile, there is no significant correlation between the perceived helpfulness of analogies and reliance-based dependent variables.

Table 13: Resulting main themes from the thematic analysis of participants' responses to the open questions pertaining to the decision criteria.

Topic	Frequency	Participant Feedback
Picture	91	(1) I looked at the pictures and tried to match them with the descriptions for either malignant or benign. - *Analogy-OD* (2) based on the image content and my understanding of malignant features. - *Control* (3) by judging the photos. - *Analogy*
Examples	77	(1) Based on the examples shared and severity of the colours and depth of the shape. - *Analogy* (2) I looked at the image and referred back to the malignant and benign images and tried to think which it resembled. - *Analogy*
Explanations	77	Started off by remembering the concepts and applying them to the initial image. Then refining that based on the AI. Generally trusted the AI's decisions more than my own. I weighed up the Positive and Negative evidence. - *Analogy*
Intuition	68	(1) I went entirely on instinct. If the image made me feel uncomfortable I labelled it malignant. Funnily enough most of the time my instincts were in agreement with the AI. - *Control* (2) how i thought it maybe should look if it was something bad. - *Analogy*
AI advice	62	(1) Applied the knowledge that I previously had and the information taught in this task; used AI to help if I was a bit confused and it was labeling the image. - *Analogy-OD* (2) Based on my intuition and recommendations from the AI system. - *Analogy*

24.5% participants in the *Analogy* condition found the analogies to be helpful (perceived helpfulness > 0), while 51.6% participants in the *Analogy-OD* condition thought the analogies are helpful. This may also help explain why participants in the *Analogy* condition showed slightly lower **Switch Fraction**, **Accuracy-wid**, **RAIR** and **RSR** in comparison with the *Analogy-OD* condition. Combined with the strong positive correlation between perceived helpfulness and user trust in the AI system, we can infer that participants in the *Analogy* condition showed less trust and reliance on the AI system (*i.e.,* they exhibited under-reliance on the AI system). Meanwhile, participants in the *Concept-Imp* condition showed very low **RSR**, which indicates over-reliance on the AI system.

6.5.4 Qualitative Analysis of Feedback

We asked all participants in our study for their rationales in their decision making using an open-ended question ("*Please describe how you made your decisions in these tasks.*"). Using the thematic analysis software, ATLAS.ti,[9] we conducted a thematic analysis and selected the top-5 topics mentioned by users (shown in Table 13).

For participants who received explanations along with the AI advice, we asked for their feedback regarding the usefulness of explanations. They showed diverse opinions regarding the helpfulness of the explanations and analogies. To illustrate the main reasons, we listed the top reasons in Table 14.

Table 14: Main reasons for perceived helpfulness of explanations.

Opinion	**Reason**
explanations are "helpful" or "somewhat helpful"	(1) explanations enrich the context of decision making or help make decision - 32.4%; (2) explanations help improve the understanding of the AI system - 18.7% (3) explanations help confirm or validate their decision - 7.2%
explanations are "unhelpful" or "somewhat unhelpful"	(1) participants lack knowledge or expertise to interpret explanations - 41.9%; (2) participants failed to understand the explanations - 16.3%; (3) explanations are difficult to apply - 11.6%
Analogies are unhelpful	(1) participants failed to connect the source domain with the target domain - 22.9%; (2) participants think the analogies do not make sense - 18.6%; (3) participants think the concepts are not relevant - 14.3%; (4) participants fail to understand the analogies - 12.9%; (5) participants think the analogies are not necessary - 10%.

We asked participants in conditions *Analogy* and *Analogy-OD* for their feedback and comments on the provided analogies. Overall, we found conflicting attitudes toward the provided analogies. While some users found merit in their use, others found them to be distracting. This is reflected in the sample quotes from two participants below.

> "*It's definitely useful and helpful for getting the point across to laymen like myself*".

> "*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 the effectiveness of analogy-based explanations. Based on the feedback from participants in the relevant experimental conditions in our study, we summarized the following potential directions to further improve the effectiveness of analogy-based explanations:

9. https://atlasti.com

- *Enhancing the relation between the target domain and the source domain (analogies).* Among participants who found analogies to be "unhelpful," many of them claimed that they failed to understand the analogies or make immediate connections or associations with the target domain.

- *Providing analogies in a more relevant domain.* Some participants complained that they failed to connect the concepts used in the analogies with the context of medical analysis. Analogies in a relevant domain can potentially help improve the plausibility and trustworthiness of analogy-based explanations.

- *Providing analogies selectively or on demand.* When the original explanation is clear enough, some participants would take the analogies as unnecessary or even distracting. Some others reported feeling annoyed: "*However, when the concept is straightforward or otherwise readily met in normal daily life, the use of an analogy can easily be perceived as condescending or even irritating and thus antagonize, rather than assist, the person concerced.*" However, if a lot of analogies are used, users may feel overwhelmed, which may hurt their trust and satisfaction with the analogy-based explanations.

7. Discussion

In summary of the experimental results, Table 15 provides an overview of the findings. Based on the findings in Study I and Study II, we elaborately discussed the potential effect of analogy properties. We also identify and synthesize the limitations of our studies.

Table 15: Summary of key findings in two studies.

Study	Findings
Study I	The proposed qualitative dimensions were found to positively correlate with the perceived helpfulness of analogy-based explanations.
	The expert evaluation results show that experts do not always agree on some qualitative dimensions (*e.g., Structural Correspondence*).
Study II	The analogy-based explanations fail to bring improved user understanding, which is assessed by the learning effect of the concepts.
	Participants showed similar levels of performance across all conditions, participants showed better appropriate reliance in condition *Concept*.
	Participants who received explanations indicated higher cognitive load.
	Participants who received explanations spent significantly more time making decisions.

7.1 Key Findings and Implications

Subjectivity of Analogies. The results of the study I especially highlight the subjective nature of the qualitative dimensions that characterize analogies. According to Krippendorff's α, we find that experts show clear disagreement on most qualitative dimensions. This is possibly because of the different experiences of the world each expert has, leading to different interpretations and familiarity of the commonsense facts in the analogies. Prior work on inter-rater disagreement suggested that disagreement is not always noise but can also be a signal (Aroyo & Welty, 2015). With disagreement from multiple explainees, we can address the ambiguity and vagueness of analogy-based explanations and seek further improvement (Inel et al., 2014; Schaekermann et al., 2019). When

evaluators find that one commonsense explanation falls short in specific dimensions, we can involve another crowd worker to improve it according to the feedback.

The comparison between the quality of explanations generated from the two tasks shows that better quality on a single dimension (like *Relational Similarity*) does not necessarily translate to better helpfulness in understanding the target sentence. However, if an explainee (*e.g.,* E_1 and E_5) thinks the explanation is of poor *Relational Similarity*, they may tend to judge it unhelpful. Meanwhile other user factors (like abstract thinking, personal interpretation, and general attitude in disagreement analysis) may also affect the perceived helpfulness and other qualitative dimensions. This points out the need for further studies about the impact of user factors (*e.g.,* experience, belief) and qualitative dimensions on the helpfulness of analogy-based explanations.

Contradicting with the assumption that commonsense knowledge should be accepted and understood by all humans (Ilievski et al., 2021), the disagreement from experts also reveals that commonsense explanations are not one-size-fits-all solutions for laypeople. This is in line with findings for explainable AI (Sokol & Flach, 2020; Liao & Varshney, 2021). In the future, one should adjust the commonsense explanations according to the explainee's belief about the world to ensure the effectiveness of such analogical inference from commonsense knowledge. This also suggests that the role of personalization should be carefully considered when generating commonsense explanations.

Automatic Analogy Generation and Evaluation. In study I, we observed that around one-third of generated analogies are not factually correct, and that it can be difficult for workers to generate analogies that demonstrate a high *Structural Correspondence* and *Relational Similarity*. This highlights the need for strategies to support workers in generating effective analogies. Especially, we envision the development of machine-in-the-loop crowdsourcing tasks, *e.g.,* by using relational knowledge bases and machine learning methods as an auxiliary toolkit to facilitate automation (Veale, 2005; Chiu et al., 2007). Knowledge bases store real world facts in a pre-defined format, typically a triplet $\langle$ subject, predicate, object $\rangle$. Hence, once the relationship between the concept and label in a target sentence is identified, it would be straightforward to find correct everyday facts sharing the same relationship along with high *Structural Correspondence*. This would provide high-quality candidate concepts to the crowd workers, reducing their work load.

Our results of study I highlight that most qualitative dimensions show a significant positive correlation to perceived helpfulness. Yet, it would be expensive to always obtain a human evaluation for quality control. Future work should hence investigate the (semi-)automatic assessment of the different quality dimensions (or at least of *helpfulness*). For *Syntactic Correctness*, one could involve automation toolkits (like syntactic error detection provided by Grammarly[10]) to provide suggestions for fixing syntactic errors when participants generate analogies on the fly. For *Simplicity* and *Misunderstanding*, one could maintain a list of everyday concepts and a list of concepts with multiple interpretations for ease of automatic check. Recent work on jury learning (Gordon et al., 2022) proposed a method to conduct automatic pseudo-human value judgement with machine learning models, which can be an alternative to expert-based quality evaluation, while accounting for the subjectivity of each dimension.

The Role of Human Intuition. In study II, many participants reported that they relied on their intuition to make their final decisions. This indicates that human intuitions play a critical role in shaping user understanding and reliance behaviors. Our findings suggest that human intuition can be a potential factor to achieve the goal of appropriate reliance on AI systems. This is in line with

10. https://www.grammarly.com/

prior findings about human intuition in the human-AI decision making context (Chen et al., 2023a, 2023b).

On the one hand, human intuition may facilitate complementary collaboration with the AI system. On the other hand, human intuition can also cause bias when making decisions. In our study, we found that the **Agreement Fraction** is relatively high (on average, around 0.80 across all conditions), while **RSR** is low for most conditions. In other words, when AI advice is wrong and users disagree, they tend to rely on AI advice instead of their initial decision (which is correct). This indicates a clear over-reliance on the AI system. This is also found in prior studies about the pitfalls of XAI interventions (Bansal et al., 2021; Wang & Yin, 2021). Such over-reliance can be associated with confirmation bias and the illusion of explanatory depth (Bertrand et al., 2022). Meanwhile, participants also showed clear under-reliance in condition *Control*, *Concept*, *Analogy* (significantly worse than condition *Concept-Imp*). A potential cause for such under-reliance can be the Dunning-Kruger effect (Kruger & Dunning, 1999). As reported by He *et al.* (He et al., 2023), "users who overestimated their capability on the task tend to exhibit under-reliance." In our study, several participants reported that they did not find explanations and analogies helpful. However, we found a strong positive correlation between the perceived helpfulness of explanations (analogies) and the subjective trust in the AI system. We can infer that participants' trust was negatively affected by the perceived unhelpfulness of analogies, which may have further impacted user reliance on the AI system. In the broader context of human-AI decision making, it would be arguably impossible for most laypeople to comprehensively understand complex AI systems. According to Lee *et al.* (Lee & See, 2004), "trust guides reliance when complexity and unanticipated situations make a complete understanding of the automation impractical." Thus, participants in our study may have exhibited under-reliance due to uncalibrated trust.

The Role of Plausibility. Through the results of the empirical study, we found that many participants thought (1) the target domain of proposed analogy-based explanations was clear enough; and (2) extra analogies are not always helpful, especially when participants fail to connect them with the target domain. Such findings can be partially explained by the plausibility of explanations. Participants implicitly hold the belief that "plausible explanations typically imply correct decisions, and vice versa" (Jin et al., 2023). Those participants who may have found the analogies to be implausible may have perceived certain AI advice as untrustworthy and thereby relied less on the AI system. Such under-reliance could result in sub-optimal team performance. This may help explain the finding that participants in the *Analogy* condition showed worse **RAIR** than participants in the *Concept-Imp* condition. Compared to the *Analogy* condition, more participants in *Analogy-OD* took the analogies as plausible (perceived helpfulness > 0). Meanwhile, participants in *Analogy-OD* condition showed higher **Switch Fraction**, **Accuracy-wid**, and **RAIR** and **RSR**. This indicates that providing analogies on demand may be a good design to facilitate human-AI collaboration. When analogies are not used appropriately, both under-reliance and over-reliance can be triggered due to implausibility.

7.2 Caveats and Limitations

Bias in Templates. We used 6 pre-defined templates to help participants generate analogy-based explanations. While crowd workers can generate syntactically correct explanations to elucidate the relevance level in concept-based explanations, these templates may lead to biases in the analogy generation (Hube et al., 2019; Draws et al., 2021). These templates show an initial bias to rela-

tionships which may limit the participants' creativity in generating useful analogies. However, as we found through our study, participants benefit from domain cues that can help them anchor their creativity and generate high-quality analogies.

Restricted Usage. Meanwhile, analogy-based explanations may not be the ideal solution for all application scenarios. According to results from our study, we summarize several scenarios inappropriate to adopt analogy-based explanations. First, when the original task is simple enough and only involves everyday concepts, analogy-based explanations may not work as expected. In such scenarios, analogy-based explanations turn out to pose more cognitive load and make it confusing to users. Second, when no explicit properties and relationship are associated with the task domain (like CLC in our study), analogy-based explanations may not be as effective for laypeople. In these tasks, it would be very hard to generate effective analogies due to a lack of explicit structural correspondence and relational similarity.

As the analogy-based explanations are generated based on concept-level explanations, cascading effects are also a limitation for analogy-based explanations. If the concept-level explanations do not faithfully reflect the internal state of AI systems, there is no chance for analogy-based explanations to do so. Furthermore, as analogy-based explanations are more familiar to most users, they have the potential to be more persuasive than original concept-based explanations. In other words, when the concept-level explanations mislead AI system users, effective analogy-based explanations generated from them may amplify such impact.

Potential Human Biases. Draws *et al.* have demonstrated that cognitive biases introduced by task design and workflow can negatively impact crowdsourcing experiments (Draws et al., 2021). Using the Cognitive Biases Checklist (Draws et al., 2021), we analyzed the potential biases in our study and reported our findings here. On the task ISIC-0032557 most participants thought that they made correct decisions and reported a high confidence in their decisions. However, that may have been a result of an illusion of their competence on the task. They achieved only 4.3% accuracy on this task. This suggests that **Overconfidence or Optimism Bias** bias (*i.e.,* Dunning-Kruger effect (Kruger & Dunning, 1999; He et al., 2023)) may have played a role in shaping these outcomes. Meanwhile, some participants also reported that the explanations helped confirm and validate their initial decision, suggesting a potential role of **Confirmation Bias** in shaping our findings. In our study, we provide 4-7 concept-level explanations / analogy-based explanations along with each task. From the open text feedback, two participants reported an information overload. This may have some negative impact on user trust and reliance. Due to the **Self-interest Bias**, crowd workers may not have thoroughly checked explanations in each task.

Threats to generalizability. In study I, we generated and evaluated analogy-based explanations on two relatively simple and low-stake tasks. The perceived quality of analogy-based explanations should be further evaluated with more realistic decision scenarios which require AI support. Although the generated analogy-based explanations are thought to be highly transferable, it is unknown how our findings and insights can generalize to complex and high-stake tasks. If the generated analogies are not always transferable, it would be valuable to investigate how to generate effective analogy-based explanations for specific high-stake tasks, *e.g.,* with experts.

Since human intuition may have heavily affected decision making in this task, some findings in study II may not generalize to tasks where human intuition does not have a dominant role. In our studies, only the relevance level between concepts and model predictions is highlighted and explained with analogies. However, analogies can be used to express more complex structural

corresponding and relationally similar events in real-world problems. Our findings may not carry forward to more complex concept-level explanations (*e.g.,* in case of a greater number of concepts or more complex relational structures between concepts).

8. Conclusions and Future Work

In this paper, we propose to elucidate concept-level AI explanations with analogical inference from commonsense knowledge in order to facilitate meaningful collaborations between an AI system and non-expert humans receiving advice from the AI system. To this end, we first designed a template-based analogy generation method, and we instantiated our method by recruiting crowd workers to generate analogy-based explanations using two image classification tasks – calorie level classification and scene classification (**RQ1**). To assess the quality of the generated explanations, we then synthesized a structured set of quality dimensions and applied it to our explanations (**RQ2**). An expert-led evaluation showed that our proposed method can generate high-quality analogy-based explanations with non-expert workers.

To comprehensively explore how analogy-based explanations affect user understanding of and reliance on the AI system, we then conducted a follow-up empirical study on a skin cancer detection task (**RQ3** and **RQ4**). Results from this second study showed that (1) the lack of domain expertise hinders user understanding of concept-level explanations; (2) compared to traditional concept-level explanations, the improved concept-level explanations (*i.e.,* target domain of our analogy-based explanations) can promote appropriate reliance on the AI system by mitigating under-reliance, but may also trigger over-reliance; (3) providing analogies on demand can be a good design for adoption of analogy-based explanations; (4) yet analogy-based explanations should be carefully designed and used in order to effectively elucidate concept-level explanations. Experimental results provide limited support that analogy-based explanations can facilitate user understanding of the AI system or appropriate reliance on the AI system. However, we cannot deny the potential of analogy-based explanation in assisting laypeople for effective decision making. Compared to concept-level explanations, the additional analogies do not cause a significant delay in decision making or pose a significantly higher cognitive load. Our findings suggest that the key challenge is in generating high-quality analogies and the potential for personalization. Based on the qualitative analysis of participants' feedback and user reliance patterns, we summarized guidelines for future work about generating effective analogy-based explanations and on the appropriate usage of analogy-based explanations.

In this work, we focused on generating high-quality analogy-based explanations using non-expert crowd workers, and evaluating their effectiveness. With the results from the first study ($N = 100$), it is evident that both generation and evaluation of analogy-based explanations are challenging and time-consuming. In the imminent future, we will consider including machine learning algorithms and leverage knowledge bases to automate this task while achieving scalability and efficiency. In our second study ($N = 280$), we found that analogy-based explanations do not work as expected in facilitating appropriate reliance. However, we found enough evidence that highlights their potential for aiding laypeople in understanding AI systems. Hence, further research about the generation of effective analogy-based explanations and their appropriate use is required. Particularly, we also found that the understanding of commonsense explanations varies with the experience of the recipient user, which points out the need for further work on the personalization of commonsense explanations.

Acknowledgments

We thank Lorenzo Corti for the helpful discussions. This work was partially supported by the Delft Design@Scale AI Lab, the 4TU.CEE UNCAGE project, the Convergence Flagship "ProtectMe" project, and the HyperEdge Sensing project funded by Cognizant. We made use of the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-3888 and EINF-5571. We finally thank all participants from Prolific and experts from our department.

References

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.

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.

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, *36*(1), 15–24.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, 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.

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., & Médini, L. (Eds.), *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pp. 1709–1719. ACM.

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.

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.

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.

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.

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, Compiègne, France, December 16-18, 2019, Proceedings*, Vol. 11940 of *Lecture Notes in Computer Science*, pp. 339–354. Springer.

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.

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*.

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.

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.

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.

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.

Chromik, M., Eiband, M., Buchner, F., Krüger, 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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé 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.

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)*.

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.

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

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.

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.

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

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

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

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

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

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.

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.

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.

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

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.

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.

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)*.

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.

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

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

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.

Hüllermeier, 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.

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

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.

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.

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

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.

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.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, 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, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 2673–2682. PMLR.

Körber, 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.

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.

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*.

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.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, 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.

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.

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

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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

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.

Nourani, M., King, J., & Ragan, E. (2020b). 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*, Vol. 8, pp. 112–121.

Prade, H., & Richard, G. (2021). Analogical proportions: Why they are useful in AI. In Zhou, Z. (Ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 4568–4576. ijcai.org.

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*.

Rajani, N. F., McCann, B., Xiong, C., & Socher, R. (2019). Explain yourself! leveraging language models for commonsense reasoning. In Korhonen, A., Traum, D. R., & Màrquez, L. (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4932–4942. Association for Computational Linguistics.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Richland, L. E., & Hansen, J. (2013). Reducing cognitive load in learning by analogy. *International Journal of Psychological Studies*, *5*(4), 69.

Robbemond, V., Inel, O., & Gadiraju, U. (2022). Understanding the role of explanation modality in ai-assisted decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 223–233.

Salimzadeh, S., He, G., & Gadiraju, U. (2024). Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-ai decision-making. In *In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*.

Schaekermann, M., Beaton, G., Habib, M., Lim, A., Larson, K., & Law, E. (2019). Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making. In *ACM Conference on Human Factors in Computing Systems (CHI'22), Workshop on Trust and Reliance in AI-Human Teams (trAIt)*.

Selbst, A., & Powles, J. (2018). "meaningful information" and the right to explanation. In *Conference on Fairness, Accountability and Transparency*, pp. 48–48. PMLR.

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. In Meersman, R., & Tari, Z. (Eds.), *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002 Irvine, California, USA, October 30 - November 1, 2002, Proceedings*, Vol. 2519 of *Lecture Notes in Computer Science*, pp. 1223–1237. Springer.

Sokol, K., & Flach, P. A. (2020). One explanation does not fit all. *Künstliche Intell.*, *34*(2), 235–250.

Sovrano, F., Sapienza, S., Palmirani, M., & Vitali, F. (2022). Metrics, explainability and the european ai act proposal. *J*, *5*(1), 126–138.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In Singh, S., & Markovitch, S. (Eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 4444–4451. AAAI Press.

Stratton, S. J. (2021). Population research: convenience sampling strategies. *Prehospital and disaster Medicine*, *36*(4), 373–374.

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge university press.

Thalheim, B. (2011). The theory of conceptual models, the theory of conceptual modelling and foundations of conceptual modelling. In Embley, D. W., & Thalheim, B. (Eds.), *Handbook of Conceptual Modeling - Theory, Practice, and Research Challenges*, pp. 543–577. Springer.

Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., & Bernstein, A. (2021). Second chance for a first impression? trust development in intelligent system interaction. In Masthoff, J., Herder, E., Tintarev, N., & Tkalcic, M. (Eds.), *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, pp. 77–87. ACM.

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, *5*(1), 1–9.

Veale, T. (2005). Analogy generation with hownet. In Kaelbling, L. P., & Saffiotti, A. (Eds.), *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, pp. 1148–1153. Professional Book Center.

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence*, 103952.

Verhoef, E. I., van Cappellen, W. A., Slotman, J. A., Kremers, G.-J., Ewing-Graham, P. C., Houtsmuller, A. B., van Royen, M. E., & van Leenders, G. J. (2019). Three-dimensional analysis reveals two major architectural subgroups of prostate cancer growth patterns. *Modern Pathology*, *32*(7), 1032–1041.

von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity: a game for collecting common-sense facts. In Grinter, R. E., Rodden, T., Aoki, P. M., Cutrell, E., Jeffries, R., & Olson, G. M. (Eds.), *Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*, pp. 75–78. ACM.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15.

Wang, X., & Yin, M. (2021). Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pp. 318–328.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12.

Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., & Huang, M. (2018). Augmenting end-to-end dialogue systems with commonsense knowledge. In McIlraith, S. A., & Weinberger, K. Q. (Eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 4970–4977. AAAI Press.

Yuksekgonul, M., Wang, M., & Zou, J. (2023). Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6720–6731. Computer Vision Foundation / IEEE.

Zhang, Q., Lee, M. L., & Carter, S. (2022). You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–28.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., & Zanfir-Fortuna, G. (Eds.), *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pp. 295–305. ACM.

Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, *40*(6), 1452–1464.