It Is Like Finding a Polar Bear in the Savannah! Concept-level AI Explanations with Analogical Inference from Commonsense Knowledge

He, G.; Balayn, A.M.A.; Buijsman, S.N.R.; Yang, J.; Gadiraju, Ujwal

DOI
10.1609/hcomp.v10i1.21990

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing

It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge

Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, Ujwal Gadiraju

Delft University of Technology, Netherlands
{g.he, a.m.a.balayn, s.n.r.buijsman, j.yang-3, u.k.gadiraju}@tudelft.nl

Abstract

With recent advances in explainable artificial intelligence (XAI), researchers have started to pay attention to concept-level explanations, which explain model predictions with a high level of abstraction. However, such explanations may be difficult to digest for laypeople due to the potential knowledge gap and the concomitant cognitive load. Inspired by recent work, we argue that analogy-based explanations composed of commonsense knowledge may be a potential solution to tackle this issue. In this paper, we propose analogical inference as a bridge to help end-users leverage their commonsense knowledge to better understand the concept-level explanations. Specifically, we design an effective analogy-based explanation generation method and collect 600 analogy-based explanations from 100 crowd workers. Furthermore, we propose a set of structured dimensions for the qualitative assessment of analogy-based explanations and conduct an empirical evaluation of the generated analogies with experts. Our findings reveal significant positive correlations between the qualitative dimensions of analogies and the perceived helpfulness of analogy-based explanations. These insights can inform the design of future methods for the generation of effective analogy-based explanations. We also find that the understanding of commonsense explanations varies with the experience of the recipient user, which points out the need for further work on personalization when leveraging commonsense 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-stake 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 and Kim 2017; Ghorbani and al 2019; Ribeiro, Singh, and Guestrin 2016a).

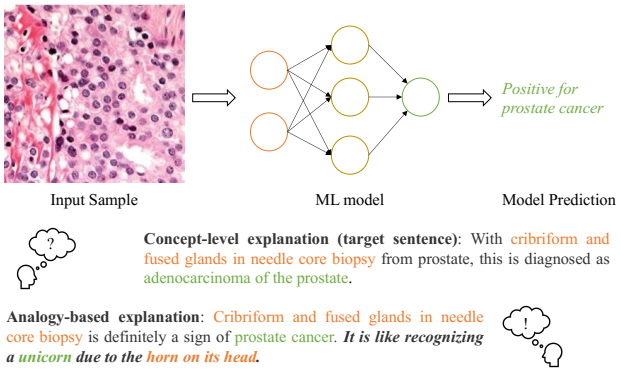

Figure 1: Example of analogy-based explanation in prostate cancer detection. The medical image and the concept-level explanation are sourced from (Verhoef et al. 2019).

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, Singh, and Guestrin 2016b; Lundberg and 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. (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 it unnecessary for users or stakeholders (*e.g.,* patients or loan applicants taking medical or finan-

[1]https://www.autoweek.com/news/green-cars/a37114603/tesla-fsd-mistakes-moon-for-traffic-light/

cial 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. For example, when identified risky for a disease or denied a loan, patients or loan applicants only need explanations that can offer actionable recourse (Vaughan and Wallach 2020). This is similar to how airplane passengers do not need to fully understand either the four forces of flight – *lift*, *drag*, *weight*, and *thrust* – or the inner-workings of an internal-combustion engine to inform their decision to fly. Providing such explanations to passengers nonetheless, would increase their cognitive load without necessarily informing their decision making.

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 for a wide range of systems. Likewise, according to GDPR, the users of AI systems should have the right to access meaningful explanations of model predictions (Selbst and 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.,* cabriform 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 and 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 and 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 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 "*calibriform and fused glands in needle core biopsy*" are also a strong pattern which indicates the model prediction "*positive for prostate cancer*". Users can therefore 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 the cognitive load.

Despite the intuitive promise and potential of analogy-based explanations, two main challenges and corresponding research questions remain to be addressed:
(**RQ1**) *How can we systematically assess the quality of analogy-based explanations?*
(**RQ2**) *How can we generate high-quality analogy-based explanations using non-experts*?

To address these RQs, we first defined a structured set of dimensions through which one can assess the quality of analogy-based explanations. Then we designed a novel analogy generation method to obtain high-quality analogy-based explanations. Next, we recruited crowd workers as non-experts to generate such explanations using our method. Finally, we carried out an expert evaluation of the quality of the collected explanations across the different dimensions. 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.[2]
- 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.

Our results demonstrate that our method is highly efficient in obtaining high-quality analogy-based explanations which can be used for explaining ML model behaviors to laypeople. All Likert-based qualitative dimensions were significantly positively correlated with the perceived helpfulness of explanations, demonstrating their appropriateness. Meanwhile, our results also highlight the subjective nature of the qualitative dimensions that characterize analogies. 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 qualitative evaluation, 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*, and the broader context of *human-centered explainable AI*.

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, Chin, and Havasi 2017), and via games with a purpose (von Ahn, Kedia, and Blum 2006; Balayn et al. 2022a).

[2]Data and code can be found at https://github.com/delftcrowd/HCOMP2022_ARCHIE

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.

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 and 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 and Justi 2012).

In the context of artificial intelligence, the importance of analogies has been recognized by multiple AI applications such as representation learning (Liu, Wu, and Yang 2017), preference learning (Bounhas et al. 2019), and image processing (Law, Thome, and Cord 2017). Readers can refer to (Prade and 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 and 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, besides human-based methods like for teaching purpose (Duit et al. 2001; Cosgrove 1995), some research also explored the automatic generation of analogies. Veale (2005) explored how lexical resource HowNet (Dong and 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, Poupart, and DiMarco (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.

Human-centered Explainable AI

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 and Varshney 2021). Consequently, more and more researchers start to work with human-centered explainable artificial intelligence (HCXAI) (Ehsan and Riedl 2020; Wang et al. 2019; Liao and Varshney 2021; Ehsan et al. 2022), putting the human at the center of technology design (Ehsan and Riedl 2020).

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 and 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.

Effective Analogies

Properties of analogical argument. Analogies have been widely used as explanations for educational and learning purposes (Nashon 2004; Mozzer and 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, Hansen, and Riefer 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 2013):

- 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 and 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 between 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 and Thagard 1989), people use analogies guided by a series of constraints that favour coherence in analogical reasoning (Mozzer and 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 (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 and 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 (2011) also argued that the source domain of effective analogies should be "easily interpretable and understandable".

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 and Thagard 1989) and "structural alignment" (Gentner and Markman 1997). "Helpfulness" corresponds to the "purpose" mentioned in Holyoak and Thagard's multiconstraint theory (Holyoak and 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

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]**.

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

empirical lens, and verify the effectiveness of our proposed set of dimensions in determining the quality of the analogy-based explanations.

Generation Tasks. 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, Chin, and Havasi 2017). Hence, we select two image classification tasks: calorie level classification (CLC) and scene classification (SC).

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

Figure 2: Example of tasks used to generate analogies.

For the calorie level classification task, we used the dataset provided by Buçinca 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.

Templates. 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). 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).

Hint Domains. 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.

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. Meanwhile, all participants are forbidden

to fill in concepts belonging to the task domain (like places and furnitures in the Places task). An example of the analogy generation interface is shown in Figure 3.

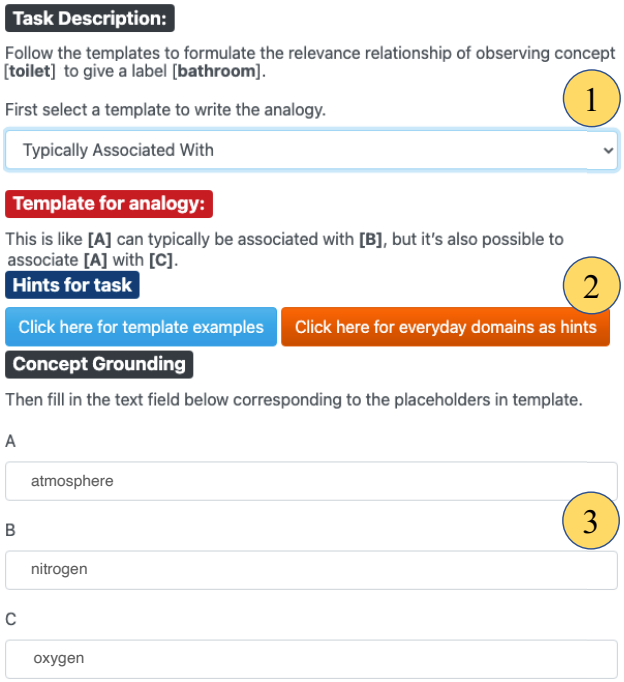

Figure 3: Analogy generation main interface and workflow. (1) Participants select a template to describe the relevance level; (2) refer to examples and everyday domains as hints; and (3) fill in concepts in placeholders to generate analogy.

5 Study Design and Experimental Setup

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

Analogy Generation

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.

[3]https://www.prolific.co/

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.

Analogy Evaluation

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. previous section), 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 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 2013), degree of structural analogy (Thalheim 2011), semantic similarity (Holyoak and 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

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	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}
	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 causes lead to more than single interpretation?	{Yes, No}

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

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 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.

[4]https://github.com/delftcrowd/HCOMP2022_ARCHIE

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

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.*"

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*.

6 Results and Analysis

Descriptive Statistics

Among the 294 generated analogy-based explanations, 255 (nearly 87%) were recognized as valid by all five experts (*i.e.,* crowd workers generate explanations with concepts in a different domain from the target sentence). 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).

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.

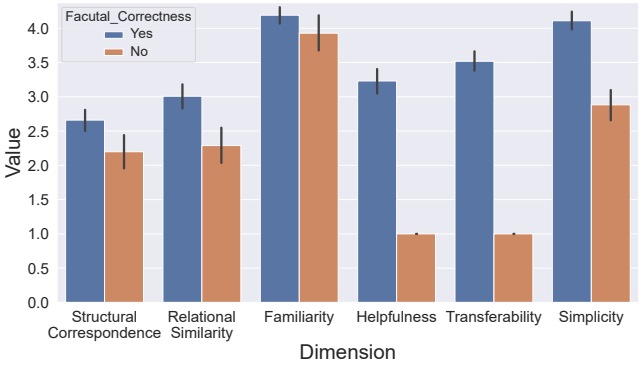

Figure 4: Bar plot illustrating the difference across the qualitative dimensions based on Factual Correctness. All dimensions were measured on a 5-point Likert scale.

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

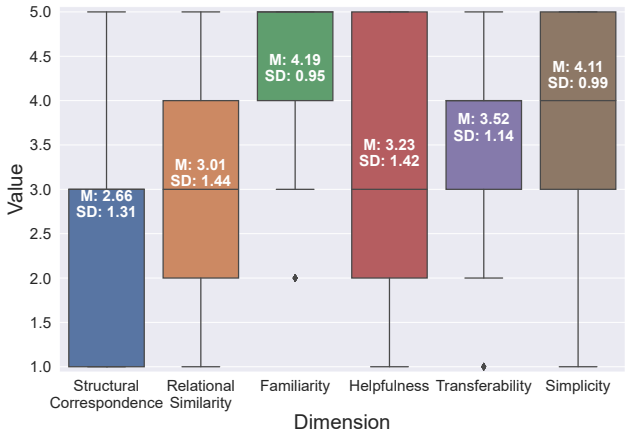

Figure 5: Box plot illustrating the distribution of the different dimensions considered in our study. All dimensions were measured on a 5-point Likert scale. For all dimensions, 1 indicates a poor quality while 5 indicates a good quality. M and SD represent mean and standard deviation respectively.

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.

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 gener-

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.

ated 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 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 relationship. However, such good analogical 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.

7 Discussion

Key Findings and Implications

Subjectivity of Analogies. Our results especially highlight the subjective nature of the qualitative dimensions that characterize analogies. According to the 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 and 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 dimension, 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 to the need for further studies about the impact of user factors (*e.g.,* experience, belief) and qualitative dimensions on 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 and Flach 2020; Liao and 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.

Analogy Generation. In our study, 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, Poupart, and DiMarco 2007). Knowledge bases store real world facts in a pre-defined format, typically a triplet ⟨ subject, predicate, object ⟩. 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.

Automatic Analogy Evaluation. Our results highlight that most qualitative dimensions show significant positive correlation to the perceived helpfulness. Yet, it would be expensive to always obtain 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[5]) 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.

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, Fetahu, and Gadiraju 2019; Draws et al. 2021). These templates show an initial bias to relationships 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.

Generalization Issue. 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.

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.

8 Conclusions and Future Work

In this paper, we propose to elucidate concept-level AI explanations with analogical inference from commonsense knowledge. We construct a structured set of dimensions to assess the quality of analogy-based explanations. To verify the effectiveness of this approach and the assimilated qualitative dimensions, we carried out an empirical study with non-experts who generated analogies and followed it up with expert-based evaluation of the generated analogies. We designed a template-based method and recruited crowd workers to generate analogy-based explanations using two image classification tasks – calorie level classification and scene classification. Results show that our method can generate high-quality analogy-based explanations with non-experts.

In this work, we focused on generating analogy-based explanations using crowd workers. In the imminent future, we plan to further explore scenarios in which experts can power and generate analogy-based explanations. It is now evident that analogy generation is a challenging and time-consuming task for humans. We will therefore consider including machine learning algorithms and knowledge bases as means to automate and achieve better scalability and efficiency in analogy-based explanation generation. While analogy-based commonsense explanations show great potential for aiding laypeople in understand AI systems, such explanation may be limited by the cascading effects from the concept-level explanations used as reference. In the future, we will delve into generating faithful concept-level explanations, which are fit for further analogy-based interpretation.

[5]https://www.grammarly.com/

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, 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-1806. We finally thank all participants from Prolific and experts from our department.

References

Abdul, A.; von der Weth, C.; Kankanhalli, M.; and 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*, 1–14.

Aroyo, L.; and 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.; and 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.; and Médini, L., eds., *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 1709–1719. ACM.

Balayn, A.; Rikalo, N.; Lofi, C.; Yang, J.; and 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*, 1–16.

Bartha, P. 2013. *Analogy and analogical reasoning*. Stanford Encyclopedia of Philosophy.

Bounhas, M.; Pirlot, M.; Prade, H.; and Sobrie, O. 2019. Comparison of Analogy-Based Methods for Predicting Preferences. In Amor, N. B.; Quost, B.; and Theobald, M., eds., *Scalable Uncertainty Management - 13th International Conference, SUM 2019, Compiègne, France, December 16-18, 2019, Proceedings*, volume 11940 of *Lecture Notes in Computer Science*, 339–354. Springer.

Buçinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In Paternò, F.; Oliver, N.; Conati, C.; Spano, L. D.; and Tintarev, N., eds., *IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020*, 454–464. ACM.

Checco, A.; Roitero, K.; Maddalena, E.; Mizzaro, S.; and Demartini, G. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

Chiu, A.; Poupart, P.; and 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*, 561–570. ACL.

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.

Dong, Z.; and Dong, Q. 2003. HowNet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003*, 820–824. IEEE.

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

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Duit, R.; Roth, W.-M.; Komorek, M.; and 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.; and Mazwi, M. L. 2022. Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. *Nature Medicine*, 1–2.

Ehsan, U.; and Riedl, M. O. 2020. Human-centered explainable ai: towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*, 449–466. Springer.

Ehsan, U.; Wintersberger, P.; Liao, Q. V.; Watkins, E. A.; Manger, C.; Daumé III, H.; Riener, A.; and 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*, 1–7.

Galesic, M.; and 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.; and Markman, A. B. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1): 45.

Ghorbani, A.; and al. 2019. Towards automatic concept-based explanations. In *NeurIPS*.

Gilbert, J. K.; and Justi, R. 2016. Analogies in modelling-based teaching and learning. In *Modelling-based teaching in science education*, 149–169. Springer.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J. T.; Hashimoto, T.; and 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.; and Yatani, K., eds., *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, 115:1–115:19. ACM.

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

He, G.; and Gadiraju, U. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making.

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

Holyoak, K. J.; and Thagard, P. 1989. Analogical Mapping by Constraint Satisfaction. *Cogn. Sci.*, 13(3): 295–355.

Hube, C.; Fetahu, B.; and 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*, 1–12.

Hüllermeier, E. 2020. Towards Analogy-Based Explanations in Machine Learning. In Torra, V.; Narukawa, Y.; Nin, J.; and Agell, N., eds., *Modeling Decisions for Artificial Intelligence - 17th International Conference, MDAI 2020, Sant Cugat, Spain, September 2-4, 2020, Proceedings*, volume 12256 of *Lecture Notes in Computer Science*, 205–217. Springer.

Ilievski, F.; Oltramari, A.; Ma, K.; Zhang, B.; McGuinness, D. L.; and 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.; and Sips, R.-J. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *International semantic web conference*, 486–504. Springer.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; and Huang, M. 2020. Generating Commonsense Explanation by Extracting Bridge Concepts from Reasoning Paths. In Wong, K.; Knight, K.; and 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*, 248–257. Association for Computational Linguistics.

Kelly, C. J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; and 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.; and Sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Dy, J. G.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 2673–2682. PMLR.

Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; and 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.; and Cord, M. 2017. Learning a Distance Metric from Relative Comparisons between Quadruplets of Images. *Int. J. Comput. Vis.*, 121(1): 65–94.

Liao, Q. V.; and 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.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In Inui, K.; Jiang, J.; Ng, V.; and 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*, 2829–2839. Association for Computational Linguistics.

Liu, H.; Wu, Y.; and Yang, Y. 2017. Analogical Inference for Multi-relational Embeddings. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 2168–2178. PMLR.

Lundberg, S. M.; and 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.; and 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*, 4765–4774.

Majumder, B. P.; Camburu, O.; Lukasiewicz, T.; and McAuley, J. J. 2021. Rationale-Inspired Natural Language Explanations with Commonsense. *CoRR*, abs/2106.13876.

Mozzer, N. B.; and 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.

Prade, H.; and 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*, 4568–4576. ijcai.org.

Preece, A.; Harborne, D.; Braines, D.; Tomsett, R.; and Chakraborty, S. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184*.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In Korhonen, A.; Traum, D. R.; and 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*, 4932–4942. Association for Computational Linguistics.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016a. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD interna-*

tional conference on knowledge discovery and data mining, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016b. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Krishnapuram, B.; Shah, M.; Smola, A. J.; Aggarwal, C. C.; Shen, D.; and Rastogi, R., eds., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144. ACM.

Schaekermann, M.; Beaton, G.; Habib, M.; Lim, A.; Larson, K.; and 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.

Selbst, A.; and Powles, J. 2018. "Meaningful Information" and the Right to Explanation. In *Conference on Fairness, Accountability and Transparency*, 48–48. PMLR.

Singh, P.; Lin, T.; Mueller, E. T.; Lim, G.; Perkins, T.; and Zhu, W. L. 2002. Open Mind Common Sense: Knowledge Acquisition from the General Public. In Meersman, R.; and 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*, volume 2519 of *Lecture Notes in Computer Science*, 1223–1237. Springer.

Sokol, K.; and Flach, P. A. 2020. One Explanation Does Not Fit All. *Künstliche Intell.*, 34(2): 235–250.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In Singh, S.; and Markovitch, S., eds., *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 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.; and Thalheim, B., eds., *Handbook of Conceptual Modeling - Theory, Practice, and Research Challenges*, 543–577. Springer.

Vaughan, J. W.; and Wallach, H. 2020. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence*.

Veale, T. 2005. Analogy Generation with HowNet. In Kaelbling, L. P.; and Saffiotti, A., eds., *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, 1148–1153. Professional Book Center.

Verhoef, E. I.; van Cappellen, W. A.; Slotman, J. A.; Kremers, G.-J.; Ewing-Graham, P. C.; Houtsmuller, A. B.; van Royen, M. E.; and 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.; and Blum, M. 2006. Verbosity: a game for collecting common-sense facts. In Grinter, R. E.; Rodden, T.; Aoki, P. M.; Cutrell, E.; Jeffries, R.; and Olson, G. M., eds., *Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*, 75–78. ACM.

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

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge. In McIlraith, S. A.; and 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*, 4970–4977. AAAI Press.

Zellers, R.; Bisk, Y.; Farhadi, A.; and 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*, 6720–6731. Computer Vision Foundation / IEEE.

Zhou, B.; Lapedriza, À.; Khosla, A.; Oliva, A.; and Torralba, A. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1452–1464.