Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making

GAOLE HE, Web Information Systems, Delft University of Technology, The Netherlands

UJWAL GADIRAJU, Web Information Systems, Delft University of Technology, The Netherlands

Although AI systems have proved to be powerful in supporting decision making in critical domains, the underlying complexity and their poor explainability pose great challenges for humans to understand their working mechanisms. When humans make the final decisions supported by AI assistance, optimal team performance can be achieved if humans and AI can complement each other. Without appropriate interpretation about AI systems, users tend to overestimate or underestimate the effectiveness of AI systems when making decisions with their assistance. This consequently hinders users from achieving an optimal complementary team performance. To promote appropriate trust and reliance on AI systems, researchers have proposed to generate explanations that can open-up the opaque box of AI decision making. Existing methods (*e.g.,* feature attribution) have achieved partial success in improving human understanding of AI decision making. However, only a few works have reported an improvement with respect to appropriate trust or reliance on AI systems. Analogies, which can borrow commonsense knowledge and mental models from our daily experience, help people quickly comprehend new information and build mental models to deal with new situations. This paper argues that analogies can help explain statistical concepts pertaining to AI systems and the complex reasons behind AI advice, by using commonsense knowledge mastered by people. We argue that this can be particularly useful for laypeople or users who may lack subject expertise or an affinity with technology interaction. In such contexts, generating analogy-based explanations for AI systems has the potential to be an effective instrument in promoting appropriate reliance in human-AI decision making processes.

Additional Key Words and Phrases: human-AI decision making, analogy, trust, appropriate reliance, human-AI interaction

ACM Reference Format:
Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *CHI 2022: ACM Conference on Human Factors in Computing Systems, April 14–15, 2022, New Orleans, LA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, AI decision support systems have been widely adopted in critical domains [13] — including medical diagnosis [14, 18], employee recruitment [10] and risk control in finance [6]. While AI systems have been proved to be powerful in supporting human decision making, it is still unclear how humans can appropriately rely on AI systems to achieve complementary team performance [3, 12]. Here, we refer to *appropriate reliance* as the notion of humans relying on AI systems when they are accurate (or more accurate than humans) and not relying on them when the systems are inaccurate (or less accurate than humans). Users in the real world, however, seldom know when their decision is an inaccurate one, nor do they easily determine when they need to depend on an AI system to inform their decision [3]. To facilitate fruitful human-AI collaboration, researchers have explored how to generate explanations for laypeople [8].

Their results show that both interactive explanations and "whitebox" explanations (*i.e.,* that show the inner workings of an algorithm) can improve users' comprehension.

In practice, it is common that laypeople do not have enough expertise in both AI systems and the application domain. To bridge such a knowledge gap, explanations can be provided to explain AI systems and domain-specific patterns. Designed under different principles, explanations can be tailored to specific audiences. To promote appropriate reliance on AI systems for laypeople, we need explanations which are characterized by both comprehensibility and a reasonable cognitive load. On the one hand, comprehension forms the basis for users to construct mental models about model behavior and performance. A mental model is "a hypothesis about the explained event's history, specifically in a way which can generalize to other events" [19]. In general, laypeople can seldom understand statistical concepts and working mechanisms of AI systems, which are typically far beyond their knowledge. Thus, explanations should be simple enough to comprehend (*i.e.,* only requiring commonsense knowledge to fully understand). On the other hand, if the explanations impose a large cognitive load on users, users may give up considering the appropriateness of relying on AI systems. As a result, users tend to perform more randomly and show inappropriate patterns of reliance (over-reliance and under-reliance) with a higher probability [5, 28].

An analogy can be interpreted as a structural mapping of a target domain that is to be clarified, onto a source domain which the recipient of the analogy is more familiar with [16, 17]. The aim is to facilitate inferences about the target domain by allowing the recipient of the analogy to make them in the source domain, and then map them back onto the target domain. This clarification of the target domain only works if the mapping between the two is sufficiently accurate regarding the features of the target domain that are meant to be elucidated by the analogy. As a simple example, one might elucidate the urgency of filling a job vacancy by saying '*it is as urgent as ants in a hot pan.*' As the recipient is likely to know that ants would feel an elevated urgency in a hot pan, such an inference on the target domain can result in an understanding that filling the vacancy is extremely urgent. Similarly, one might elucidate how dangerous a job is by saying '*it is like treading on thin ice.*' As the recipient is likely to know that treading on thin ice can be dangerous, an inference on the target domain can be drawn such that the relevant job can also be dangerous. With analogical inference, users can interpret new information with familiar concepts. Furthermore, such an analogy can also borrow the mental models of users from their experience of dealing with familiar concepts to reduce their cognitive load. Thus, we argue that analogies can satisfy the comprehensibility needs of laypeople without adversely increasing their cognitive load.

In this work, we make a case for using analogies to explain AI systems and help laypeople collaborate with AI decision support systems more effectively. We argue that analogy-based explanations promise to be an effective avenue to promote appropriate reliance on AI systems, and synthesize promising directions for further research.

2 USAGE AND RATIONALE

Analogies can be used to elucidate statistical concepts pertaining to AI systems and explain the causality behind their decision making process. Prior works have partially supported such usage of analogies. With analogy-based explanations, humans can draw inferences from their existing knowledge and experience to aid their understanding of AI systems [7]. Such inference reduces the cognitive load of users, and also helps build appropriate mental models for reliance on AI systems [25]. Figure 1 illustrates how analogy-based explanations help interpret the statistical concepts and causally explain AI system behavior for users.

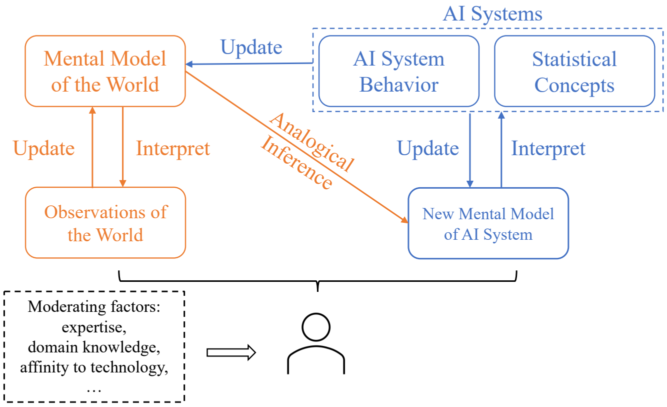
AI Systems
Mental Model of the World
Update
AI System Behavior
Statistical Concepts
Analogical Inference
Update
Interpret
Update
Interpret
Observations of the World
New Mental Model of AI System
Moderating factors: expertise, domain knowledge, affinity to technology, …

Fig. 1. This illustration presents our understanding of the working mechanism of analogy-based explanations on statistical concepts (*e.g.,* global accuracy) and the overall behavior of AI systems. The orange colours represent elements in a user's everyday observations and their constantly evolving mental model of the world, based on their experiences of the world. The blue colours represent elements pertaining to interactions and experiences with AI systems. In general, people build a powerful mental model of the world to interpret observations in daily situations. Along with interpreting more observations of the world, such mental models get updated. With the aid of analogical inference, people may build up a new mental model of an AI system based on their existing mental model of the world. They can interpret AI systems and update both mental models after that.

2.1 Elucidating Statistical Concepts

When users make decisions under the support of AI systems, there are many statistical concepts to elucidate (*e.g.,* system performance and potential impact).

Local confidence scores and global performance measures are widely used to assess system performance. Most machine learning models are trained, tuned and evaluated with training, development and test set respectively. To assess the global performance, evaluation metrics (*e.g.,* accuracy, NDCG [20]) on the test set are always reported along with AI systems. While such evaluation metrics are widely used by researchers and developers, laypeople may have difficulty in understanding them to build mental models [2] of system performance. Without a proper mental model about AI systems, laypeople may feel confused about AI systems and even be misled to make mistakes on cases that they are capable of dealing with. As a result, they may overestimate or underestimate AI systems' effectiveness [2], which in turn can hinder them from relying on AI systems appropriately when making decisions. For example, laypeople may find it difficult to interpret how often they should rely on AI systems when presented with system accuracy or other performance metrics [26]. However, most of laypeople should be familiar with weather forecasts. When leveraging such an analogy to explain the system accuracy, for instance, "*the AI system is as accurate as a 5-day weather forecast*", laypeople can leverage their mental model of weather forecasts to inform their reliance on the AI system. Similar to global performance measures, users rely on local confidence scores to assess AI system trustworthiness locally (*i.e.,* at the level of each decision). In some high-stake decision scenarios (*e.g.,* recruitment, loan), it is very important to ensure human decision makers are aware of the risks behind each decision. In prior works, analogies have been found to be effective tools to improve risk perception [4] and help patients understand medical statistical concepts (*e.g.,* effectiveness of preventive medical treatments) [15].

Analogies can also be used to illustrate the potential impact of AI systems, such as fairness, strengths and weaknesses. The accurate description of such abstract and complex concepts, for example statistical data distribution shifts [9], goes beyond the general know-how of laypeople. Although laypeople may find it difficult to easily understand such niche concepts, they master commonsense knowledge and build mental models to deal with everyday concepts that surround us. If some everyday concepts (*e.g.,* the dynamics of visitor numbers in different seasons, or the menu at a vegetarian restaurant) share similar properties with target concepts of AI systems (*i.e.,* data distribution shifts), laypeople can understand that AI systems can not deal with out-of-distribution data samples just like a restaurant may not be able to provide a dish out of the menu. In this sense, analogies can provide a vivid understanding drawn from commonsense knowledge and mental models of dealing with daily concepts.

2.2 Explaining Causality of AI Behaviors

To illustrate the causality behind AI decision behaviors, the explainable artificial intelligence (XAI) community has put great efforts into model-agnostic approaches, like SHAP [21] and LIME [24], which show salient parts of input as explanations. Prior works have leveraged such approaches to compare and contrast AI system behavior with human understanding of different tasks [22, 27]. However, it still requires some domain expertise to connect such salient parts of input and understand AI decision making processes. Analogies can provide concepts and relations that laypeople are familiar with to bridge such expertise gaps and aid understanding. In this sense, analogies can be a potentially powerful instrument to help laypeople understand AI decision making processes and rely on AI systems appropriately. For example, in loan approval tasks, consider a case where an explanation for a decision highlights the following features corresponding to a loan applicant – [`age: 48`, `salary: 2000` USD per month, owner of a `large house` in an urban area, and having `no credit history`]. However, laypeople may still feel confused to connect such highlighted features to a decision of rejection. In such a context, an analogy-based explanation that reads as follows can be effective — "It is strange that the applicant of age 48 has a large house in an urban area. According to their salary, they cannot afford such a house without any credit history. *It is like finding one who never received any formal education solving an extremely complex and world famous puzzle in mathematics.*"

Following the theory of mind literature, Jacovi *et al.* [19] establish a framework to describe the concrete information that humans comprehend from explanations. They argued that effective explanations should be coherent and complete to establish a coherent mental model. While it is hard to build a new coherent mental model, analogies may be a practical and effective way to borrow mental models in dealing with similar contexts from experience.

3 DISCUSSION

To our knowledge, analogies can play an important role for laypeople to understand complex AI systems. For statistical concepts (*e.g.,* global accuracy, local confidence score, impact of AI systems), analogies can improve risk perception and aid understanding. For local decision behaviors, analogies are able to translate complex causality behind AI advice into commonsense interplay mastered by laypeople through their real-world experiences. Besides improving understanding of statistical concepts or complex causality, analogies also reduce users' cognitive load [25], which creates a friendly environment for humans to collaborate with AI systems. Furthermore, it contributes to building up coherent and complete mental models to deal with AI systems efficiently from experience [1]. Such functionalities make analogy a potentially powerful method to promote appropriate reliance for laypeople or users lacking domain expertise, while interacting with AI systems.

It is important to also consider the limitations of using analogies to improve appropriate reliance. To ensure that analogies are beneficial to both enhancing understanding and reliance behaviors, there are two prerequisites: (1) the source domain of analogy-based explanation should be able to characterize the target domain (*e.g.,* statistical concepts) [16], and (2) users should be familiar with the source domain [11] and not be biased with respect to it [23]. If the source (analogy) domain cannot faithfully capture the characteristics of the target domain, analogies provide misleading information which hinders appropriate reliance. Suppose users were provided with an analogy-based explanations which goes against their belief or preference, it is inevitable to generate biased decision behaviors and negative feedback. In the use cases we considered (cf. Section 2), users are not good at understanding statistical concepts of AI systems and the causality that drives the behavior of AI systems. If the task is easy enough to comprehend for laypeople, using analogies may not work as expected – as opposed to enhancing users' understanding and reducing their cognitive load, they may only serve to increase the cognitive load without effecting users' understanding. Despite these limitations that can constrain the usage of analogies, utilizing analogies to promote appropriate reliance is supported by theories and observations from existing works, and we argue that this line of research deserves further exploration.

4 OPEN RESEARCH QUESTIONS

We should consider how user factors (*e.g.,* preference, familiarity, experience) and task factors (*e.g.,* complexity, expertise requirement) can affect how humans perceive the role of analogies in interpreting AI systems. There are several promising research directions for the imminent future, that can help us understand and explore the extent to which analogies can be used as an instrument to promote appropriate reliance on AI systems in the context of human-AI decision making. We aim to tackle these research gaps through empirical work, and present a brief synthesis below:

- How can user factors and task factors affect the perception of analogy-based explanations? Do these lead to biased decision making processes?
- How can we find appropriate commonsense source domains (*i.e.,* everyday concepts that users are familiar with) to explain the target domain (*i.e.,* statistical concepts or causality)?
- How can we efficiently generate analogy-based explanations? Can we use human-in-the-loop approaches to improve the quality of analogies thus generated?
- How can we leverage analogy-based explanations to achieve personalization and promote appropriate reliance in human-AI interaction?

REFERENCES

[1] Michael Stuart Arnold. 1996. Teaching a scientific mental model, a case study: Using analogy to construct a model of thermal processes. (1996).

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*. 1–16.

[4] Elisa Barilli, Lucia Savadori, Stefania Pighin, Sara Bonalumi, Augusto Ferrari, Maurizio Ferrari, and Laura Cremonesi. 2010. From chance to choice: The use of a verbal analogy in the communication of risk. *Health, Risk & Society* 12, 6 (2010), 546–559.

[5] Tad T Brunyé, Shaina B Martis, and Holly A Taylor. 2018. Cognitive load during route selection increases reliance on spatial heuristics. *Quarterly Journal of Experimental Psychology* 71, 5 (2018), 1045–1056.

[6] Longbing Cao. 2022. AI in Finance: Challenges, Techniques, and Opportunities. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–38.

[7] Jaime G Carbonell. 1983. Learning by analogy: Formulating and generalizing plans from past experience. In *Machine learning*. Springer, 137–161.

[8] Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, 559.

[9] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.

[10] Sam Desiere, Kristine Langenbucher, and Ludo Struyven. 2019. Statistical profiling in public employment services: An international comparison. (2019).

[11] Reinders Duit. 1991. On the role of analogies and metaphors in learning science. *Science education* 75, 6 (1991), 649–672.

[12] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For What It's Worth: Humans Overwrite Their Economic Self-Interest to Avoid Bargaining With AI Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

[13] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 43–52.

[14] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

[15] Mirta Galesic and Rocio Garcia-Retamero. 2013. Using analogies to communicate information about health risks. *Applied Cognitive Psychology* 27, 1 (2013), 33–42.

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

[17] Douglas R Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books.

[18] Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci. 2017. Risk stratification of lung nodules using 3D CNN-based multi-task learning. In *International conference on information processing in medical imaging*. Springer, 249–260.

[19] Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. 2022. Diagnosing AI Explanation Methods with Folk Concepts of Behavior. *CoRR* abs/2201.11239 (2022).

[20] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[21] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[22] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2022. What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. In *Proceedings of the ACM Web Conference (2022)*.

[23] Nataliia Reva. 2019. The Analogy in Decision-Making and the Implicit Association Bias Effect. *Studia Humana* 8, 2 (2019), 25–31.

[24] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 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, San Francisco, CA, USA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1135–1144.

[25] Lindsey E Richland and Janice Hansen. 2013. Reducing cognitive load in learning by analogy. *International Journal of Psychological Studies* 5, 4 (2013), 69.

[26] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 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*. 1–12.

[27] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[28] Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. 2017. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP conference on human-computer interaction*. Springer, 23–39.