Quantifying the Utility of Causal Models for Decision-Making

Elena Korshakova (ekorshak@stevens.edu)
Department of Computer Science, 1 Castle Point on Hudson
Hoboken, NJ 07030 USA

Jessecae K. Marsh (jessecae.marsh@lehigh.edu)
Department of Psychology, 17 Memorial Drive East
Bethlehem, PA 18015 USA

Samantha Kleinberg (samantha.kleinberg@stevens.edu)
Department of Computer Science, 1 Castle Point on Hudson
Hoboken, NJ 07030 USA

Abstract

Many methods exist to learn causal models from data, as causal relationships form the basis for successful actions. These methods are frequently evaluated based on the completeness of the models they can infer. Yet, there is a gap between the highly complete and potentially complex models algorithms can learn and the types of information people can use successfully to make decisions. To address this we conduct two experiments to understand how the size and features of causal models influence how well they can be used for decision-making. In Experiment 1 we systematically vary model size for a series of topics, finding that there is a negative and linear relationship between causal model size and decision accuracy. In Experiment 2 we examine how model structure influences decisions, varying whether the models include feedback loops, again finding that smaller models lead to better choices, and that feedback loops are also beneficial.

Keywords: causal models; decision-making; evaluating models

Introduction

Over the last 40 years, methods for learning causal models from data have proliferated. From Bayesian networks (BNs) (Pearl, 2000) to additive noise models (ANMs) (Hoyer, Janzing, Mooij, Peters, & Schölkopf, 2008), numerous methods exist to discover causal relationships from observational data (Assaad, Devijver, & Gaussier, 2022). Experiments are often challenging, expensive, or unethical in domains such as healthcare, making it critical to find ways to use observational data to identify causal relationships. A core argument for needing causal models is that they enable action that correlations alone cannot, including interventions, explanations, and more reliable predictions (Prosperi et al., 2020). For example, Google Flu used a correlation between user search keywords and flu cases to successfully predict flu incidence, but this correlation suddenly failed and the algorithm began predicting twice as many cases as there actually were (Lazer, Kennedy, King, & Vespignani, 2014). In another case, studies showed an association between high HDL cholesterol and a reduced risk of heart attacks, leading companies to develop drugs designed to raise HDL. Yet more recent studies raise questions about this causal link, showing genetic factors raising HDL do not lead to the expected risk reduction (Voight et al., 2012). Methods that can learn causal relationships may help avoid such errors, leading to more precise intervention targets and more reliable predictions.

Despite people being the ultimate users of causal models – as developers of policies, implementers of interventions, and followers of causal guidance – methods for learning such models from data have not been evaluated based on how well people can use them. Instead, they are often evaluated based on their completeness and accuracy. This raises questions about the translation from inferred models to actual use. Benchmark datasets have been developed for comparing algorithms, with most evaluations focusing on quantitative assessments of the causal structures identified or using down-stream tasks such as prediction from the inferred model (Cheng et al., 2022). Recent work proposed new metrics such as the distance between ground truth and inferred models (Peyrard & West, 2020) and new evaluations using the data generated by the model (Parikh, Varjao, Xu, & Tchetgen, 2022). These approaches raised concerns about the use of simulated data benchmarks (Reisach, Seiler, & Weichwald, 2021). However, these metrics still all focus on the relationship between inference and ground truth rather than the relationship between inference and actual human use.

To assess current practices in evaluation of causal inference methods, we surveyed papers from five major machine learning and artificial intelligence conferences from 2000 to 2022: Association for the Advancement of Artificial Intelligence (AAAI), International Conference on Artificial Intelligence and Statistics (AISTATS), Conference on Neural Information Processing Systems (NeurIPS), Conference on Uncertainty in Artificial Intelligence (UAI), and International Joint Conference on Artificial Intelligence (IJCAI). We collected all papers ($N = 234$) published in these venues during this time frame that describe methods for extracting causal structures (also called causal discovery) and annotated the evaluation metrics used in each paper.

The majority of papers we surveyed (58%) used only accuracy to evaluate causal inference output, such as computing the false positive or false negative rate, or recall of causal relationships. This relies on knowing what the ground truth is and prioritizes finding a larger total fraction of the existing relationships, while finding fewer incorrect ones. Significantly less work evaluated the efficiency of the algorithm (e.g., running time, data requirements) (15%), used a combination of accuracy and efficiency (8%), or used qualitative evaluation in comparison to prior work (2%). Lastly, 17% of papers were theoretical and did not perform any evaluation. We did not identify any papers evaluating how well a person could use the output to do a task, such as whether it helped them identify the correct intervention target or helped them reach a decision faster than they would on their own. Further, all quantitative evaluations we identified considered each relationship equally important to identify. That is, inferences were not scored differently depending on whether they identified a modifiable factor versus one that cannot be intervened on, nor did they consider the timing or strength of the causes. Thus, it is an open question as to whether the models that score the best on these evaluations (those that are most complete) will be the most beneficial for users. While some of the work discusses the inspectability of machine learning models to ensure that users can understand them (Khan & Vice, 2022; Zerilli, 2022), it's important to note that an inspectable model does not necessarily guarantee successful use for decision-making.

Prior work suggests that the current metrics used to evaluate causal models may not be strongly correlated to user decision-making accuracy. Korman and Khemlani (2020) found that systems presented in a single large model are perceived as more complete, but did not examine whether people can use these more complex models successfully. However, Kleinberg and Marsh (2021) showed that when giving users complex causal models, decision accuracy did not differ from when participants received no information (answering based only on their existing knowledge). On the other hand, that same research found that simple diagrams targeted to specific decisions did improve accuracy. This suggests that the models perceived as complete, and which would be scored better by current evaluation metrics, may not be the models that are most useful for decision-making.

Thus there is a tension between the types of causal models computational methods aim to find (e.g., prioritizing complex models) and the models people can use best (e.g., simple models). However, prior work has not explored in depth what in the structure of causal models makes them more complex and influences how well people can use them to make everyday choices, nor has this work aimed to identify at what point a model becomes unwieldy. A complex model can be comprised of a large number of nodes, could involve feedback loops, and many other structures in between. To address this, we conduct two experiments to examine how causal model complexity influences decision-making accuracy in everyday domains. In Experiment 1, we examine how diagram size influences decision accuracy across four decision-making domains. In Experiment 2, we test the effect of feedback loops on decision-making accuracy in the same set of domains. Together our experiments show a strong need to consider model structure in evaluations of utility, and suggest future directions for the development of metrics aimed at optimizing the use of causal diagrams.

Table 1: Decision-making domains used in the experiments.

Domain	Questions
Life decisions	Time management Career change
Health	Mental health Alcohol addiction
Societal issues	Legacy building Fundraising
Personal finances	Car Purchase Investment management

Experiment Overview

We first describe the recruitment procedure, materials, procedure, and analyses that were common across both experiments, before describing each experiment in detail.

Participants

For each experiment, we recruited participants through Prolific who were U.S. residents aged 18-64. We excluded any participants who completed the study but did not provide viable demographic information (e.g., entering random symbols or numbers in response to country of birth or in free text “Other” options). Participants were compensated $3 based on an expected study duration of 20 minutes.

Materials

We developed decision-making questions for two topics within each of four domains spanning life decisions, health, societal issues, and personal finances. The specific topics are shown in Table 1. The questions were created such that each had four answer options, with one option designated as the target for that scenario. Target options accounted for the context of the given question and should help the individual described in the scenario to achieve the goal stated in the question in the fastest, most effective way. For health topics, we used information from the Centers for Disease Control and Prevention website to additionally determine the target.

Every question had a diagram designed to assist participants. Diagrams varied in complexity and features across the experiments, including more or less detail as needed so that we could test the effect of different model features on decision-making. Overall, the diagrams were designed to be simpler than typical AI models, with the goal of ensuring that they are easy for non-experts to understand and use. All the diagrams featured only positive relationships between variables. Below is the question about time management, and Figure 1 shows all diagram variations created for this question. For example, in the five-node diagram on effective time management (Figure 1(c)), participants were expected to identify the target node ”effective time management” and explore the remaining four nodes that contribute to achieving the goal. The node ”delegation” was intended to assist participants in selecting the best answer option, which was to hire more employees and delegate technical tasks. It should be noted that the best option remained consistent across all diagrams, regardless of their size.

Jay runs a profitable machine learning startup. He manages the sales, marketing, business development, and backend technical aspects of his business. Jay works more than 60 hours each week. His startup is evolving successfully but Jay started having problems with his family. His wife told him that whenever he's at home, he is barely present and that he never spends time with their kids.

What is the BEST way for Jay to manage his time?

(a) *Explain to his wife that his business is important and requires family sacrifice.*

(b) *Hire more employees and delegate technical tasks.*

(c) *Convert his start-up into an online business and work from home.*

(d) *Go to couples therapy to work out his family problems.*

Procedure

After consenting to the study, participants were given instructions on how to use causal diagrams. After that, participants answered eight multiple-choice decision-making questions (one per topic listed in Table 2). The order of topics was randomized by participant. Each experiment had multiple diagrams for each question that varied on the given dimension of interest, but an individual participant only saw one. At the end of the survey, we collected participants' demographic information and asked for their comments, and whether they experienced any technical difficulties.

Data Analysis

We used a generalized estimating equation (GEE) analysis with a binary logistic link function and independent correlation structure to statistically model participants' answers to the decision-making questions as a repeated measures dependent variable. We defined the "correct" response as choosing our target option for a question. Accuracy was coded by assigning a score of 1 for every correct answer and 0 for every incorrect answer given by the participants in response to each question. As such, mean accuracy reflects the proportion of times the target option was selected across topics. We examined the main effects of the variable of interest (e.g., model size, presence of feedback loops) and interactions between variables when there was more than one (e.g., interaction between the number of nodes and the presence of feedback loops). The variables used were: number of nodes (Experiment 1) and number of nodes and presence of feedback loops (Experiment 2). Data analysis was conducted using SPSS Statistics v. 29.

Table 2: Decision-making accuracy by the model size.

Model size	2	3	5	7	9	11
Accuracy	.76	.75	.76	.73	.71	.70

Experiment 1: Diagram size

Prior work showed that simple diagrams led to better choices than highly complex ones (Kleinberg & Marsh, 2021). However, this work did not explore what exactly makes a diagram complex. In this experiment, we build on this by systematically varying causal model size to determine how this factor influences decision accuracy.

Method

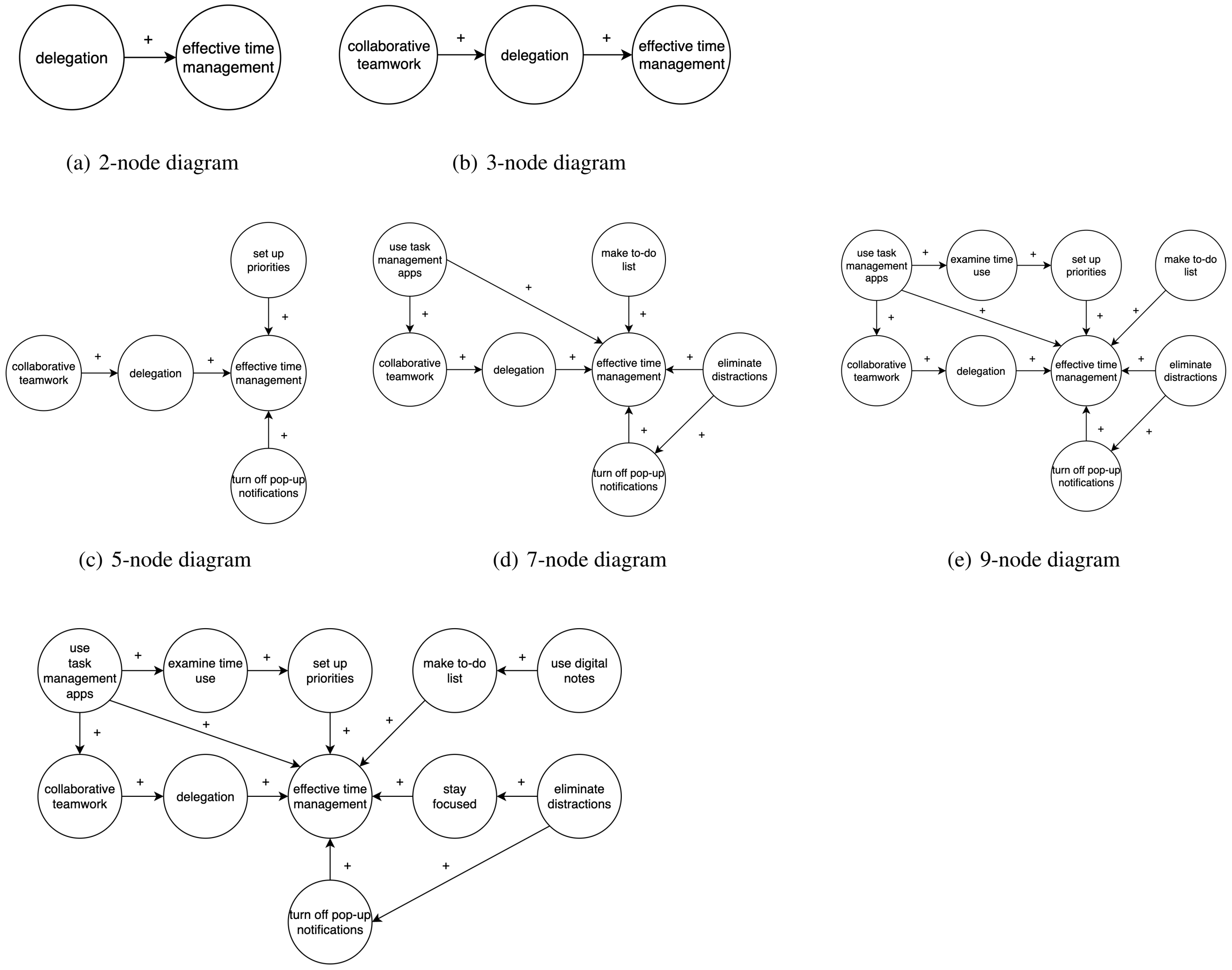
Participants We recruited a total of 600 participants due to the number of diagram variations tested. Two participants were excluded. Of the 598 participants in analysis, 299 were female, 292 were male, and 7 were non-binary.

Materials We developed diagrams of six sizes for each topic, simultaneously increasing the number of nodes and edges across the range of sizes. We chose the number of edges to maintain consistency in the number of root nodes (1 root for 2 and 3-node diagrams and 3 roots for 5, 7, 9, and 11-node diagrams). As shown in Figure 1(a), the 2-node diagram contains just the causal pathway corresponding to the target answer (e.g., delegation to achieve effective time management). This core remained consistent across all diagrams, and larger diagrams were created by adding additional detail or expanding on causal chains. Figure 1 shows all diagram variations for one topic.

Procedure Participants completed one question for each topic, with the order of questions being randomized by participant. Since we had 6 diagram levels, 4 diagram sizes were completed once and 2 were repeated for each participant. To assign the diagram sizes across questions, we counterbalanced the assignment of repeated items. This allowed us to present the diagrams in a balanced manner, ensuring that each participant receives a combination of small, medium, and large diagrams in different combinations.

Results

As we did not vary the number of edges separately from the number of nodes, we used the number of nodes as a model input to capture diagram size for the GEE analysis. The results showed a significant effect of diagram size on decision-making accuracy $\chi^2(5) = 13.780, p = .017$. We explored the main effect of number of nodes through a follow-up polynomial contrast comparison. We found a significant linear effect across the number of nodes $\chi^2(1) = 12.050, \beta = -.05, SE = .015, p = .003$. Overall, smaller diagrams resulted in higher average accuracy across topics (smallest diagram: M = .76, 95% *CI* [.73, .79]) compared to larger ones (largest diagram: M = .70, 95% *CI* [.66, .73]).



(a) 2-node diagram

(b) 3-node diagram

(c) 5-node diagram

(d) 7-node diagram

(e) 9-node diagram

(f) 11-node diagram

Figure 1: Diagrams of three sizes for Experiment 1 time management question.

Discussion

While prior work has shown a stark contrast in utility between the simplest and most complex diagrams, it has been unknown exactly how accuracy varies with model size, and whether there is a threshold over which models are too complex to be useful or whether accuracy continues to drop with increases in size. We now shed light on this by showing that the relationship is linear, with accuracy being reduced as a model becomes more complex. This suggests that there is a constant tradeoff between the amount of detail and information presented, and the use of models for decision-making purposes.

Importantly, this also means that *decision accuracy* is inversely related to computational measures of *model accuracy*. For human choices, increases in model size lead to a reduction in decision accuracy, while for computational methods these same increases would mean the model is considered more complete and has better recall of ground truth. Another potential reason for the drop in accuracy with larger diagrams is that as the number of nodes increases, it becomes harder to identify the correct match when the correct answer option is worded differently in the question. Thus there is a significant need for future work examining the interaction between causal models and decision-making. When adding nodes and edges comes at a cost, it is important to determine how to prioritize the information presented. Including nodes that are modifiable (e.g., in contrast to genetic factors, which may put one at risk but cannot be altered), or that are more easily or cheaply altered, or which can bring about an effect sooner may lead to different choices even at the same level of complexity. Similarly, different parts of a model could be emphasized, by collapsing causal chains or highlighting variables that are more intensely connected.

Experiment 2: Feedback loops

In Experiment 1 we examined the effect of a causal model's size on decision-making. However, models of the same size (number of nodes and/or edges) can be arranged in different ways that may contribute to accuracy. One feature that may influence people's perceptions of complexity is the inclusion of feedback, where nodes form a cycle such as in predator-prey dynamics. Such cycles are a core part of causal loop diagrams, which have been widely used to model system dynamics (Haraldsson, 2004; Binder, Vox, Belyazid, Haraldsson, & Svensson, 2004). Work in cognition has proposed that thinking about feedback loops could be indicative of individuals engaging in systems thinking (Hamilton, Salerno, & Fischer, 2022), meaning comprehensively considering how all parts of a system work together rather than focusing on individual components. However individuals do not often come up with such loops on their own (Levy, Lubell, & McRoberts, 2018; White, 2008), and education research suggests they are difficult for learners (Kastens et al., 2009). Nevertheless, loops may be an important intervention target as effects may be magnified. Thus we now examine whether the presence of feedback loops has an effect on decision accuracy.

Method

Participants We recruited a total of 800 participants due to the number of diagram variations tested. Three participants were excluded. Of the 797 participants in analysis, 399 were female, 394 were male, and 4 were non-binary.

Materials To investigate whether the presence of a feedback loop in diagrams affects decision-making accuracy, we developed 3, 5, 7, and 9-node diagrams with and without a feedback loop, each with 3, 5, 7, and 9 edges, respectively. We had a total of 8 unique diagram structures, four of which had a feedback loop and four of which did not. Figure 2 shows an example of diagrams at the same level of complexity (number of nodes) where one contains a feedback loop and the other does not. All feedback loops represented positive relationships. To develop these diagrams, we began by creating the smallest (3-node) diagrams with and without feedback loops. These diagrams included the causal pathway corresponding to the target response, and added one additional node. For diagrams with feedback, the loop was always connected to the target outcome (e.g., "stay focused" loop in the diagram about effective time management, shown in Figure 2(b)), so that the loop itself could remain the same for all diagram sizes. In the example shown staying focused helps to increase the chances of effective time management, while at the same time, effective time management positively contributes to staying focused.

Procedure The procedure was the same as in Experiment 1, with the difference being the diagrams shown with each question. Participants were randomly assigned to receive each question with a varying size and feedback loop (present/absent) condition in a counterbalanced manner.

Table 3: Decision-making accuracy by model size for both conditions with and without feedback loop.

Model size	3	5	7	9
Accuracy with loop	.79	.74	.73	.72
Accuracy without loop	.71	.66	.70	.68

Results

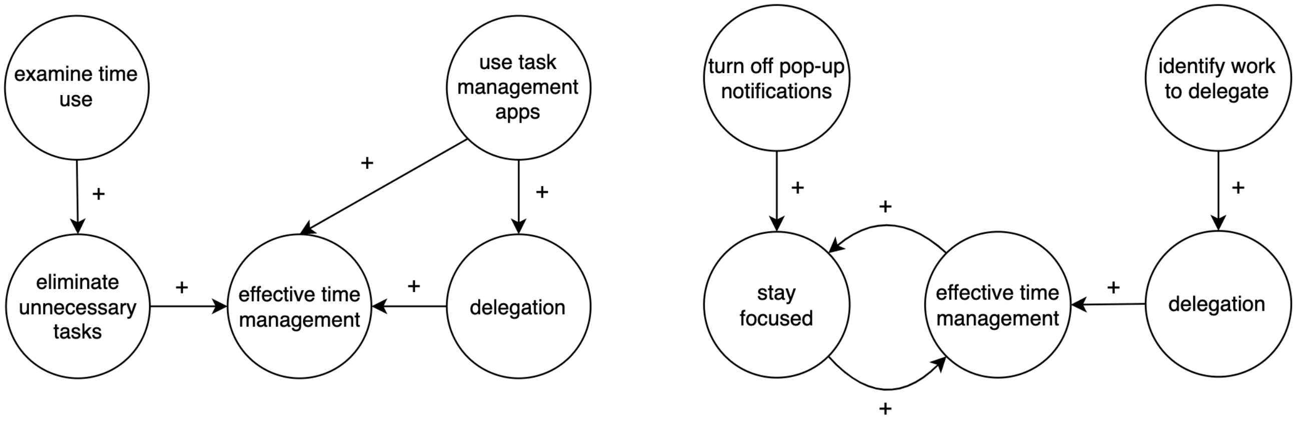
We found a significant main effect of the presence of a feedback loop, $\chi^2(1) = 30.595, p < .001$, as well as a main effect of number of nodes, $\chi^2(3) = 14.287, p = .003$. Decision-making accuracy was significantly higher in participants who received diagrams with feedback loops (*M* = .74, 95% *CI* [.73, .76]) compared to those who received diagrams without feedback loops (*M* = .69, 95% *CI* [.67, .70]). As in Experiment 1, smaller diagrams resulted in higher average accuracy across topics (smallest diagram: *M* = .75, 95% *CI* [.72, .78]) compared to larger ones (largest diagram: *M* = .70, 95% *CI* [.67, .74]). There was not a statistically significant interaction between the number of nodes and the presence of feedback loops $\chi^2(3) = 4.746, p = .191$.

Discussion

In this experiment we replicate the effect of diagram size from Experiment 1, and further find that feedback loops improved use of causal models. In one sense loops could be considered more complex than a chain with the same number of nodes and edges. Understanding loops requires thinking about multiple nodes simultaneously and across time, while chains could be considered in a sequence. There is a potential mechanisms for why instead feedback loops improved decisions in our experiment. As proposed by Hamilton et al. (2022), feedback loops may be prompting individuals to consider the system as a whole, which could lead to better understanding. Thus even if individuals rarely think of such loops on their own, diagrams that include them may be beneficial by making them salient. Considering the system as a whole may lead to better understanding the effects of interactions and how choices interact. Nevertheless, as we integrated only a single feedback loop involving only two nodes, further research is needed to explore the impact of more complex and multiple feedback loops on decision-making.

General Discussion

Causal models could form the basis for better choices by individuals and policy-makers by providing information on what interventions may be most effective and enabling them to predict the results of their actions. Yet despite the work in computer science (on methods to find causal models) and cognition (on how people reason with causal models) no work


examine time use
+
eliminate unnecessary tasks
+
effective time management
use task management apps
+
+
+
delegation
turn off pop-up notifications
+
+
stay focused
effective time management
+
+
identify work to delegate
+
delegation

(a) 5-node diagram without a feedback loop

(b) 5-node diagram with a feedback loop

Figure 2: Diagrams with 5 nodes with and without feedback loop for Experiment 2 time management question.

has aimed to understand how the models these methods infer match up to the models people can use.

Our work suggests a critical need for new metrics for evaluating causal inference methods. As we found when reviewing the literature, metrics corresponding to completeness of a structure are by far the dominant method for evaluation, and no existing methods have been evaluated in terms of whether their output can successfully guide human decisions. Future work is needed to develop the specific metrics, however our work provides a few clear avenues for follow-up. While we find that complexity hampers decision-making accuracy, it still remains to determine whether nodes and edges may be perceived as contributing differently to complexity and further whether altering the types of nodes within a complexity level can improve decisions. That is, even for the same diagram size and structure, nodes could be easier or harder to intervene on, could work at different timescales, or could have different causal strength (probability of bringing about the effect). Determining how these factors contribute to choices would enable methods that can calculate the utility of a given diagram and ultimately use this scoring method to find the optimal simplification of a causal model for a given user.

Our findings also provide insight into the process of decision-making with information. Prior work has examined how information presentation format can impact the quality of decision-making, with graphical and spatial formats yielding the best results (So & Smith, 2003; Speier, 2006). Research has also examined how much information people prefer to receive when making a choice (Fernbach, Sloman, Louis, & Shube, 2013) and how the complexity of a task influences how much information they seek out (Byström & Järvelin, 1995). We now show that giving people the more complex information they believe they need does not lead to better choices and to the contrary, the most parsimonious model should be provided if the goal is to influence a specific choice.

Our study has limitations. Firstly, it only included US participants, limiting the generalizability of the findings to other populations or cultures. Another limitation is related to the use of Prolific as the primary recruitment platform. Although Prolific is known for its high-quality research participants, it is possible that the use of this platform may introduce some biases into the sample. For example, Prolific users may be more tech-savvy or more likely to be interested in research studies, which could impact the generalizability of our results. Future studies should address these limitations by using more diverse samples and employing alternative recruiting strategies to ensure the generalizability of the findings. Lastly, we did not gather information about participants' prior knowledge of the decision-making topics. It is possible that people who knew more about these decision-making topics could handle additional complexity better than people who know less. Alternatively, people with more a priori knowledge may handle additional complexity more poorly because it conflicts with their existing knowledge. This is an important question for future research.

Additionally, more research is needed to investigate the impact of other causal diagram features on decision-making accuracy. For instance, a greater number of edges can increase the complexity of the diagram by creating a larger number of connections and interconnections between the nodes, making it more difficult to trace paths through the diagram. Positive and negative edge connections can also contribute to model complexity and affect decision-making. Furthermore, root nodes could potentially impact the interpretation of the diagram by emphasizing the importance of just a few nodes, or suggesting that the intended goal is difficult to obtain because of the long sequence of events that has to occur before it happens. In future work, we aim to explore these and other potential structural features.

Our findings can guide future evaluations of causal discovery methods and development of new algorithms by providing new features to optimize. To guide practical decision-making, we need methods that can identify the smallest possible models with the information needed for a choice. Such research can lead to better use of machine learning by explicitly designing for the human in the loop.

Acknowledgments

This work was supported in part by the NSF under award number 1907951.

References

Assaad, C. K., Devijver, E., & Gaussier, E. (2022). Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, *73*, 767–819.

Binder, T., Vox, A., Belyazid, S., Haraldsson, H., & Svensson, M. (2004). Developing system dynamics models from causal loop diagrams. In *Proceedings of the 22nd international conference of the system dynamic society* (pp. 1–21).

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information processing & management*, *31*(2), 191–213.

Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K. S., & Liu, H. (2022). Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, *3*(6), 924–943.

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research*, *39*(5), 1115–1131.

Hamilton, M., Salerno, J., & Fischer, A. P. (2022). Cognition of feedback loops in a fire-prone social-ecological system. *Global Environmental Change*, *74*, 102519.

Haraldsson, H. V. (2004). *Introduction to system thinking and causal loop diagrams*. Department of chemical engineering, Lund University Lund, Sweden.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *NeurIPS*.

Kastens, K. A., Manduca, C. A., Cervato, C., Frodeman, R., Goodwin, C., Liben, L. S., … Titus, S. (2009). How geoscientists think and learn. *Eos, Transactions American Geophysical Union*, *90*(31), 265–266.

Khan, M. M., & Vice, J. (2022). Toward accountable and explainable artificial intelligence part one: Theory and examples. *IEEE Access*, *10*, 99686-99701. doi: 10.1109/ACCESS.2022.3207812

Kleinberg, S., & Marsh, J. (2021). It's complicated: Improving decisions on causally complex topics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Korman, J., & Khemlani, S. (2020). Explanatory completeness. *Acta Psychologica*, *209*, 103139.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Levy, M. A., Lubell, M. N., & McRoberts, N. (2018). The structure of mental models of sustainable agriculture. *Nature Sustainability*, *1*(8), 413–420.

Parikh, H., Varjao, C., Xu, L., & Tchetgen, E. T. (2022). Validating causal inference methods. In *International conference on machine learning* (pp. 17346–17358).

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Peyrard, M., & West, R. (2020). A ladder of causal distances. *arXiv preprint arXiv:2005.02480*.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., … Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, *2*(7), 369–375.

Reisach, A., Seiler, C., & Weichwald, S. (2021). Beware of the simulated DAG! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, *34*, 27772–27784.

So, S., & Smith, M. (2003). The impact of presentation format and individual differences on the communication of information for management decision making. *Managerial Auditing Journal*.

Speier, C. (2006). The influence of information presentation formats on complex task decision-making performance. *International Journal of Human-Computer Studies*, *64*(11), 1115-1131. doi: 10.1016/j.ijhcs.2006.06.007

Voight, B. F., Peloso, G. M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M. K., … others (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mMndelian randomisation study. *The Lancet*, *380*(9841), 572–580.

White, P. A. (2008). Beliefs about interactions between factors in the natural environment: A causal network study. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(4), 559–572.

Zerilli, J. (2022). Explaining machine learning decisions. *Philosophy of Science*, *89*(1), 1–19. doi: 10.1017/psa.2021.13