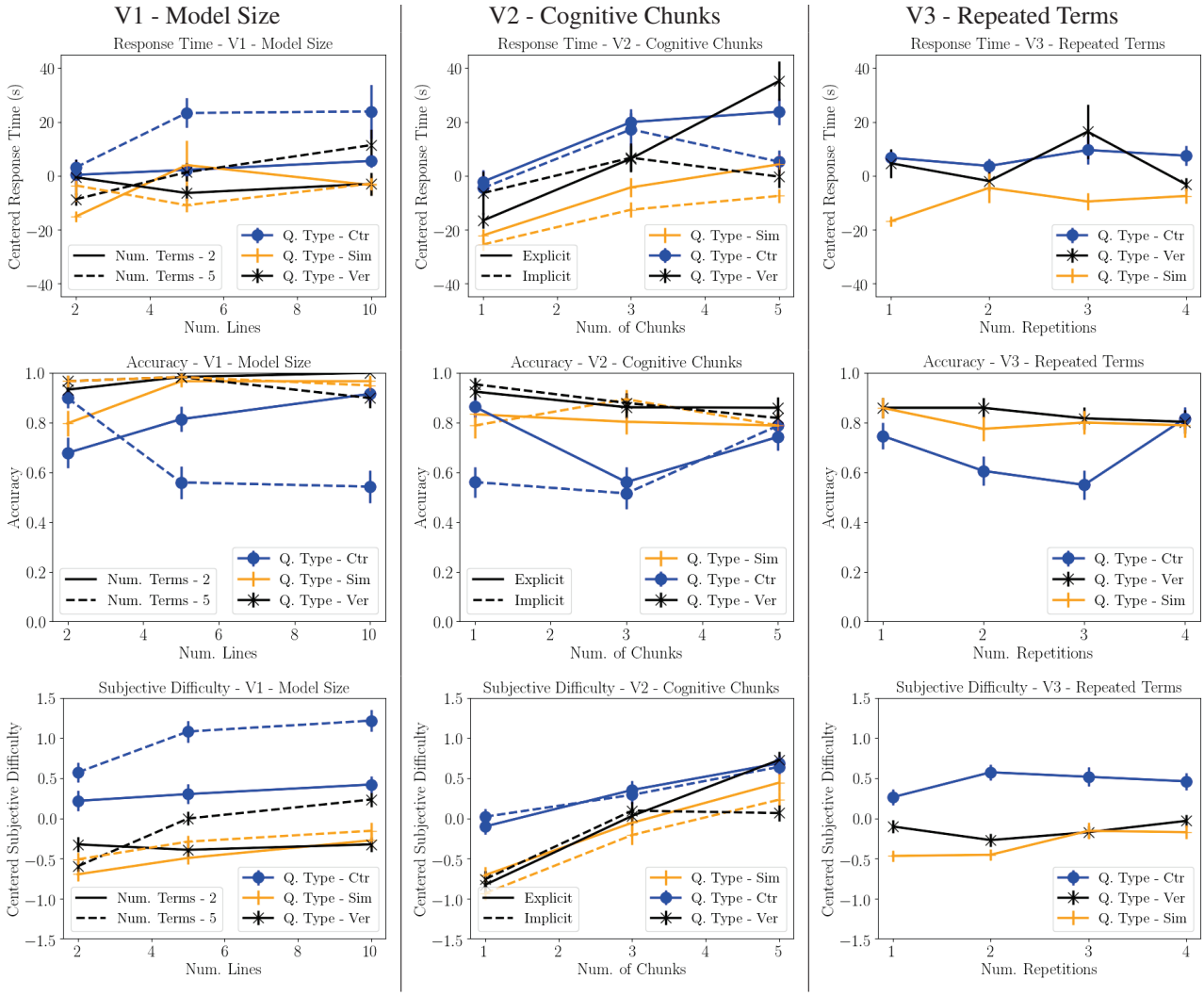

Figure 3: Accuracy, response time and subjective evaluation for all experiments in the recipe domain. Vertical lines signify standard errors. See Lage et al. (2019) for the corresponding set of figures in the clinical domain.

time on the order of 20 seconds, whereas increases in length have effects on the order of 10 seconds. Increases in variable repetition has an effect of at most a few second.

Finally, we found that participants had significantly longer response times when new cognitive chunks were made explicit rather than implicitly embedded in a line. This ran counter to our expectations since users had to process fewer but longer lines with implicit cognitive chunks compared to the same chunks defined explicitly.

Consistency across domains: Magnitudes of effects change, but trends stay the same. In all experiments, the general trends are consistent across both the recipe and clinical domains. Sometimes an effect is weaker or unclear, but never is an effect clearly reversed. There were 21 cases of factors that had a statistically significant effect on a dependent variable in at least 1 of the 2 domains. For 19 of those, the 95% confidence interval of both domains had the same sign (i.e., the entire 95% confidence interval was positive for both domains or negative for both domains). For the other 2 (the effect of verification questions on accuracy and response time for experiment V1), one of the domains (clinical) was inconclusive (interval overlaps zero).

Consistency across tasks: Relative trends stay the same, different absolute values. The effects of different types of complexity on response time were also consistent across tasks. That said, actual response times varied significantly between tasks. In Figure 3, we see that the response times for simulation questions are consistently low, and the response times for counterfactual questions are consistently high (statistically significant across all experiments except V2 in the Recipe domain). Response times for verification questions are generally in between, and often statistically significantly higher than the easiest setting of simulation.

Consistency across metrics: subjective difficulty of use follows response time, less clear trends in accuracy. So far, we have focused our discussion on response time. In Ta-

ble 3, we see that subjective difficulty of use largely replicates the findings of response time. We see that simulation questions are significantly easier for participants to answer than counterfactuals. We also see a statistically significant effect of the number of cognitive chunks, model length, and number of output terms. The finding that implicit cognitive chunks are less difficult to use than explicit cognitive chunks appeared only in the recipe domain.

Unlike response time and subjective difficulty of use, where the trends were significant and consistent, the effect of different types of complexity on accuracy was less clear. None of the effects were statistically significant, and there are no clear trends. We believe that this was because we asked participants to be accurate primarily and fast secondarily, effectively pushing any effects into response time. But even when participants were coached to be accurate, some tasks proved harder than others: counterfactual tasks had significantly lower accuracies than simulation tasks.

Discussion

Observation: Consistent patterns provide guidance for the design of interpretable machine learning systems. In this work, we found consistent patterns across tasks, and domains for the effect of different types of decision set complexity on human-simulatability. These patterns suggest that, for decision sets, the introduction of new cognitive chunks or abstractions had the greatest effect on response time, then model size (overall length or length of line), and finally there was relatively little effect due to variable repetition. These patterns are interesting because machine learning researchers have focused on making decision set lines orthogonal (e.g. Lakkaraju, Bach, and Leskovec, 2016), which is equivalent to regularizing the number of variable repetitions, but perhaps, based on these results, efforts should be more focused on length and if and how new concepts are introduced. This knowledge can help us expand the faithfulness of the model to what it is describing with minimal sacrifices in human ability to process it.

Observation: Consistency of the results across tasks and metrics suggests studies of interpretability can be conducted with simplified tasks at a lower cost. While the relative ordering of the types of complexity was the same for both the simulation and counterfactual tasks, the counterfactual tasks were more difficult for people to answer correctly. This suggests that we may be able to simplify human-subject studies by using simpler tasks that capture the relevant notions of interpretability. To measure human-simulatability, for example, it seems that the simulation task can be used in future experiments to obtain the same results at a lower cost. A second possibility for simplifying tasks is to rely on people's subjective evaluations of interpretability. In our results, the correlation between peoples' subjective evaluation of difficulty and the more objective measure of response time suggests that this can be done. However this should be followed up since it may be an artifact of running this study on Amazon Mechanical Turk where faster response times correspond to higher pay rates.

Future Work: Using Amazon Mechanical Turk to evaluate interpretability requires additional study. While simplifying tasks is one possible way to make interpretability studies on Amazon Mechanical Turk more effective, finding other ways to do so remains an open question. In our experiments, the criteria we used to exclude participants who were not able to complete the tasks effectively at the beginning of the experiment excluded over half of the participants. This is likely because the models and tasks were challenging for laypeople since they were designed with expert users in mind. Whether and how Amazon Mechanical Turk studies should be used to evaluate notions of interpretability associated with domain experts warrants future study.

Future Work: The unexpected preference for implicit cognitive chunks should be unpacked in future experiments. An unexpected finding of our study that warrants further investigation is that implicit cognitive chunks were easier for people to process than explicit ones. This could be because explicitly instantiating new concepts made the decision set harder to scan, or because people prefer to resolve the answer in one long line, rather than two short ones. Follow-up studies should investigate whether this finding persists when the familiarity of the concept or the number of times it is used within the decision set increases. These insights could guide representation learning efforts.

Limitations. There are several areas of our experiment design that future studies could explore further. Whether interfaces for interpretable machine learning models can be optimized to present the same information more effectively is an open question. In this project, we fixed ours to something reasonable based on pilot studies, but this warrants further study. Future studies measuring interpretability with humans should also explore whether the results of this study generalize to other classes of models–logic-based or otherwise, as well as whether certain model classes are inherently more interpretable or are preferred in certain cases. Finally, while this work relied on a set of synthetic tasks building on the notion of human-simulatability, future work will need to connect performance on these basic tasks to real-world tasks, like finding errors or deciding whether to trust a model, that are more difficult to run in controlled settings.

Conclusion

In this work, we investigated the effect of different types of complexity for decisions sets on the interpretability of the models, measured through three different tasks based on human-simulatability. Our results suggest that the number of cognitive chunks, then the model size are the most important types of complexity to regularize in order to learn interpretable models, while the number of variable repetitions has relatively little effect. These results are consistent across tasks and domains, which suggests that there are general design principles that can be used to guide the development of interpretable machine learning systems.

Acknowledgments

The authors acknowledge PAIR at Google, a Google Faculty Research Award, and the Harvard Berkman Klein Center for their support. IL is supported by NIH 5T32LM012411-02.

References

Allahyari, H., and Lavesson, N. 2011. User-oriented assessment of classification model understandability. In *11th scandinavian conference on Artificial intelligence*. IOS Press.

Bussone, A.; Stumpf, S.; and O'Sullivan, D. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on*, 160–169. IEEE.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. ACM.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc. 2172–2180.

Clark, P., and Boswell, R. 1991. Rule induction with cn2: Some recent improvements. In *European Working Session on Learning*, 151–163. Springer.

Cohen, W. W. 1995. Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning*, 115–123.

Elomaa, T. 2017. In defense of c4. 5: Notes on learning one-level decision trees. *ML-94* 254:62.

Feldman, J. 2000. Minimization of boolean complexity in human concept learning. *Nature* 407(6804):630–633.

Frank, E., and Witten, I. H. 1998. Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, 144–151. Morgan Kaufmann Publishers Inc.

Freitas, A. A. 2014. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter* 15(1):1–10.

Gottwald, R. L., and Garner, W. R. 1972. Effects of focusing strategy on speeded classification with grouping, filtering, and condensation tasks. *Perception and Psychophysics* 179–182.

Horsky, J.; Schiff, G. D.; Johnston, D.; Mercincavage, L.; Bell, D.; and Middleton, B. 2012. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. *Journal of biomedical informatics* 45(6):1202–1216.

Hutton, A.; Liu, A.; and Martin, C. 2012. Crowdsourcing evaluations of classifier interpretability. In *AAAI Spring Symposium Series*. AAAI Press.

Huysmans, J.; Dejaeger, K.; Mues, C.; Vanthienen, J.; and Baesens, B. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*.

Kulesza, T.; Stumpf, S.; Burnett, M.; Yang, S.; Kwan, I.; and Wong, W.-K. 2013. Too much, too little, or just right? ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, 3–10. IEEE.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S.; and Doshi-Velez, F. 2019. An evaluation of the human-interpretability of explanation. *arxiv* abs/1902.00006.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. ACM.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Leskovec, J. 2017. Interpretable & explorable approximations of black box models. *arxiv* abs/1707.01154.

Lipton, Z. C. 2016. The mythos of model interpretability. In *ICML 2016 Workshop on Human Interpretability in Machine Learning*.

Miller, G. A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review* 63(2):81.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1 – 38.

Plumb, G.; Al-Shedivat, M.; Xing, E.; and Talwalkar, A. 2019. Regularizing black-box models for improved interpretability. *arxiv* abs/1902.06787.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2017. Manipulating and measuring model interpretability. In *NIPS Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Ross, A., and Doshi-Velez, F. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI Conference on Artificial Intelligence*. AAAI Press.

Schmid, U.; Zeller, C.; Besold, T.; Tamaddoni-Nezhad, A.; and Muggleton, S. 2016. How does predicate invention affect human comprehensibility? In *International Conference on Inductive Logic Programming*, 52–67. Springer.

Seabold, S., and Perktold, J. 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, 61.

Singh, S.; Ribeiro, M. T.; and Guestrin, C. 2016. Programs as black-box explanations. *arXiv preprint arXiv:1611.07579*.

Subramanian, G. H.; Nosek, J.; Raghunathan, S. P.; and Kanitkar, S. S. 1992. A comparison of the decision table and tree. *Communications of the ACM* 35(1):89–94.

Tintarev, N., and Masthoff, J. 2015. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*. Springer. 353–382.

Treisman, A. M., and Gelade, G. 1980. A feature-integration theory of attention. *Cognitive Psychology* 97–136.

Ustun, B., and Rudin, C. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102(3):349–391.

Wang, F., and Rudin, C. 2015. Falling rule lists. In *Artificial Intelligence and Statistics*, 1013–1022.

Wu, M.; Hughes, M.; Parbhoo, S.; Zazzi, M.; Roth, V.; and Doshi-Velez, F. 2018. Beyond sparsity: Tree regularization of deep models for interpretability. In *AAAI Conference on Artificial Intelligence*. AAAI Press.