An Evaluation of the Human-Interpretability of Explanation

Isaac Lage* **Emily Chen*** **Jeffrey He*** **Menaka Narayanan*** **Been Kim†**

Samuel J. Gershman* **Finale Doshi-Velez***

Abstract

The evaluation of interpretable machine learning systems is challenging, as explanation is almost always a means toward some downstream task. In this work, we carefully control a number of properties of logic-based explanations (overall length, number of repeated terms, etc.) to determine their effect on human ability to perform three basic tasks: simulating the system's response, verification of a suggested response, and counterfactual reasoning. Our findings about how each of these properties affect the ability of humans to perform each task provide insights on how we might construct regularizers to optimize for task performance.

1 Introduction

The relatively recent widespread adoption of machine learning systems in real, complex environments has lead to an increased attention to interpretable machine learning systems. Many forms of explanation have been proposed, ranging from classical approaches such as decision trees [Breiman *et al.*, 1984] to input gradients or other forms of (possibly smoothed) sensitivity analysis [Selvaraju *et al.*, 2016; Ribeiro *et al.*, 2016; Lei *et al.*, 2016], generalized additive models [Caruana *et al.*, 2015], procedures [Singh *et al.*, 2016], falling rule lists [Wang and Rudin, 2015], exemplars [Kim *et al.*, 2014; Frey and Dueck, 2007] and decision sets [Lakkaraju *et al.*, 2016]—to name a few.

However, there has been relatively little attention given to what kind of explanation is best in what situation, and relatedly, what are the properties of a good explanation. In this work, we make modest but concrete strides toward the larger goal of quantifying what makes explanations human-interpretable via carefully controlled experiments in which we systematically vary properties of an explanation to measure their effect on the performance of several basic tasks. Our results help inform what kinds of regularizers might help optimize the utility of the explanation in a variety of situations.

2 Related Work

The field of interpretable machine learning is large. Within interpretable machine learning, most user-study based evaluations fall into an A/B testing framework, in which the user-study is used to argue that a propsed form of explanation is better than some alternative (e.g. [Kim *et al.*, 2014; Lakkaraju *et al.*, 2016; Subramanian *et al.*, 1992; Huysmans *et al.*, 2011; Hayete and Bienkowska, 2004; Freitas, 2014]). These are valuable for demonstrating the value of the proposed form of explanation, but generalizing from these studies is often non-obvious.

In contrast, our work falls into a smaller category of literature that carefully varies explanation properties to understand how those properties affect performance. In this vein, Kulesza *et al.* [2013]

*Harvard

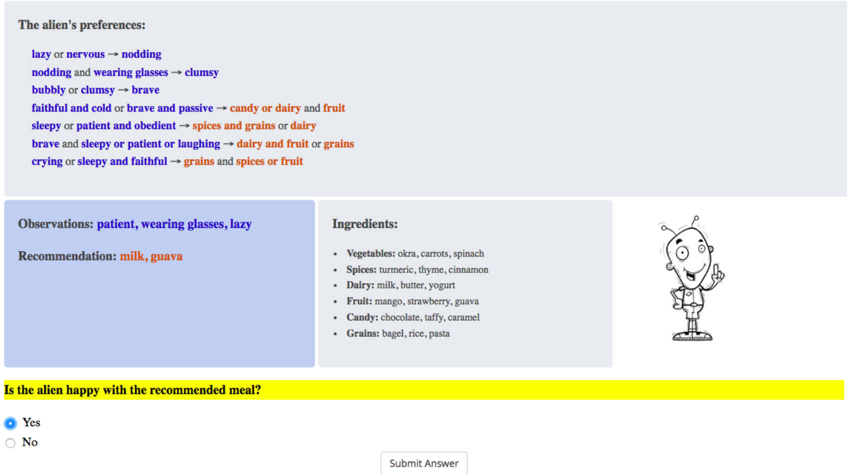
†Google Brain

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

performed a qualitative study in which they varied the soundness (nothing but the truth) and the completeness (the whole truth) of an explanation in a recommendation system setting. They found completeness was important for participants to build accurate mental models of the system. Allahyari and Lavesson [2011]; Elomaa [2017] also find that larger models can sometimes be more interpretable. Schmid *et al.* [2016] find that human-recognizable intermediate predicates in inductive knowledge programs can sometimes improve simulation time. Poursabzi-Sangdeh *et al.* [2017] manipulate the size and transparency of an explanation and find that longer explanations and black-box models are harder to simulate accurately on a real-world application predicting housing prices.

3 Methods

We conduct three experiments to test how increasing different types of complexity affects the usability of explanations for three tasks. The explanations we use are *decision sets*, as they are easy for humans to scan and parse [Lakkaraju *et al.*, 2016], but also have a large number of potentially-relevant properties that might make them easier or harder to use. The tasks we consider are: simulating the system's response, verification of a suggested response, and counterfactual reasoning because they are general enough to be relevant across a variety of domains [Doshi-Velez and Kim, 2017]. Through this work, we hope to gain an understanding of which types of complexity significantly increase the cognitive burden of using explanations and in which contexts. In this section we describe the specific explanations, types of complexity and tasks we study, as well as the metrics we record and the study design. See Narayanan *et al.* [2018] for more details.



The alien's preferences:

lazy or nervous → nodding
nodding and wearing glasses → clumsy
bubbly or clumsy → brave
faithful and cold or brave and passive → candy or dairy and fruit
sleepy or patient and obedient → spices and grains or dairy
brave and sleepy or patient or laughing → dairy and fruit or grains
crying or sleepy and faithful → grains and spices or fruit

Observations: patient, wearing glasses, lazy

Recommendation: milk, guava

Ingredients:

- **Vegetables:** okra, carrots, spinach
- **Spices:** turmeric, thyme, cinnamon
- **Dairy:** milk, butter, yogurt
- **Fruit:** mango, strawberry, guava
- **Candy:** chocolate, taffy, caramel
- **Grains:** bagel, rice, pasta

Is the alien happy with the recommended meal?

● Yes
○ No

Figure 1: Screenshot of or interface for the verification task. The bottom left box shows the observations we give participants about the alien, and a meal recommendation. They must then say whether the machine learning system agrees with the recommendation based on the explanation. Each task is coded in a different color (e.g. yellow) to visually distinguish them.

Domain In our study, participants were told they were assisting with a meal recommendation system. The system made recommendations for an alien to avoid any biasing effects of prior knowledge. All participants were given a set of *observations* about the alien–the input to the machine learning system, and the alien's *preferences*–the explanation of the system. Some tasks also required a *recommendation*–the output of the machine learning algorithm. Figure 1 shows our experimental interface. The form of this design was chosen based on feedback from pilot studies.

Tasks We consider three canonical tasks suggested in Doshi-Velez and Kim [2017] that are likely to be relevant for a variety of specific situations:

- **Simulation** Forward simulating the system's recommendation given an explanation and a set of input observations. Participants were given observations about the alien and the alien's preferences and were asked select ingredients that would satisfy the alien. See Figure 3.

- **Verification** Verifying whether the system's recommendation is consistent given an explanation and a set of input observations. Participants were given a recommendation as well as the observations and preferences and asked whether it would satisfy the alien. See Figure 1.
- **Counterfactual** Determining whether the system's recommendation changes given an explanation, a set of input observations, and a perturbation that changes one dimension of the input observations. Participants were given a change to one of the observations in addition to the observations, preferences and recommendation and asked whether the alien's satisfaction with the recommendation would change. See Figure 4.

Types of Complexity To carefully control various properties of the explanation and the context, we generated decision sets by hand that mimic those learned by machine learning systems. We conducted 3 experiments that varied the following sets of properties (all other factors were kept constant):

- **V1: Explanation Size.** We varied the size of the explanation across two dimensions: the *total number of lines* in the decision set (2, 5, or 10), and the *number of terms within the output clause* (2 or 5). For example, Figure 1 has 4 lines (in addition to the 3 lines defining explicit cognitive chunks), and 3 terms in each output clause.
- **V2: Creating New Types of Cognitive Chunks.** We varied the total number of cognitive chunks (1, 3, or 5), and whether they were implicitly or explicitly defined. For example, Figure 1 has 3 explicit cognitive chunks, and Figure 4 has 3 implicit cognitive chunks.
- **V3: Repeated Terms in an Explanation.** We varied the number of times that input terms were repeated in the decision set (2, 3, 4 or 5). For example, each of the observations in Figure 1 appears twice in the explanation (the observations used in the explicit cognitive chunks appear only once, but the final chunk appears twice).

Expt	Condition	Time Param	P Val	Subjective Param	P Val	Accuracy Param	P Val
V1	Num. Lines	1.01	0.00317	0.0491	**5.57E-11**	0.00598	0.842
V1	Num Out. Terms	1.57	0.0378	0.116	**2.54E-12**	-0.117	0.0771
V1	Counterfactual	13.7	**1.79E-06**	1.04	**1.5E-59**	-1.7	**1.24E-11**
V1	Verification	4.11	0.121	0.169	0.00475	0.476	0.174
V2	Implicit	-7.93	**0.000489**	-0.121	0.0171	-0.179	0.222
V2	Num. Chunks	5.88	**4.45E-17**	0.254	**3.76E-54**	-0.0364	0.416
V2	Counterfactual	19.9	**6.65E-12**	0.52	**1.61E-16**	-0.773	**4.36E-06**
V2	Verification	15.4	**1.27E-08**	0.092	0.14	0.532	0.00904
V3	Num. Reps	0.884	0.463	0.0676	0.00411	-0.0473	0.524
V3	Counterfactual	16.6	**1.91E-06**	0.767	**2.41E-30**	-0.67	**0.00066**
V3	Verification	13.7	**3.03E-05**	0.169	0.00899	0.196	0.371

Table 1: Significance tests for each factor. Linear regression weights were computed for response time and subjective evaluations (treated as continuous). A regression was computed for each dependent variable for each experiment (9 total). Highlighted p-values are significant at $\alpha = 0.05$ with Bonferroni correction across all tests of all experiments. The 'param' (or coefficient) indicates the magnitude and direction of the effect. The null hypothesis is that the coefficient is zero. Counterfactual and verification coefficients describe changes with respect to simulation.

Metrics To quantify the interpretability of the explanations, we recorded three metrics: response time, accuracy, and subjective satisfaction. Response time was measured as the number of seconds to respond; accuracy was the correctness of the response; and subjective satisfaction was measured on a 1 to 5 Likert scale from very easy to use to very hard to use. Subjective ratings were collected after the response so as not to add to the response time.

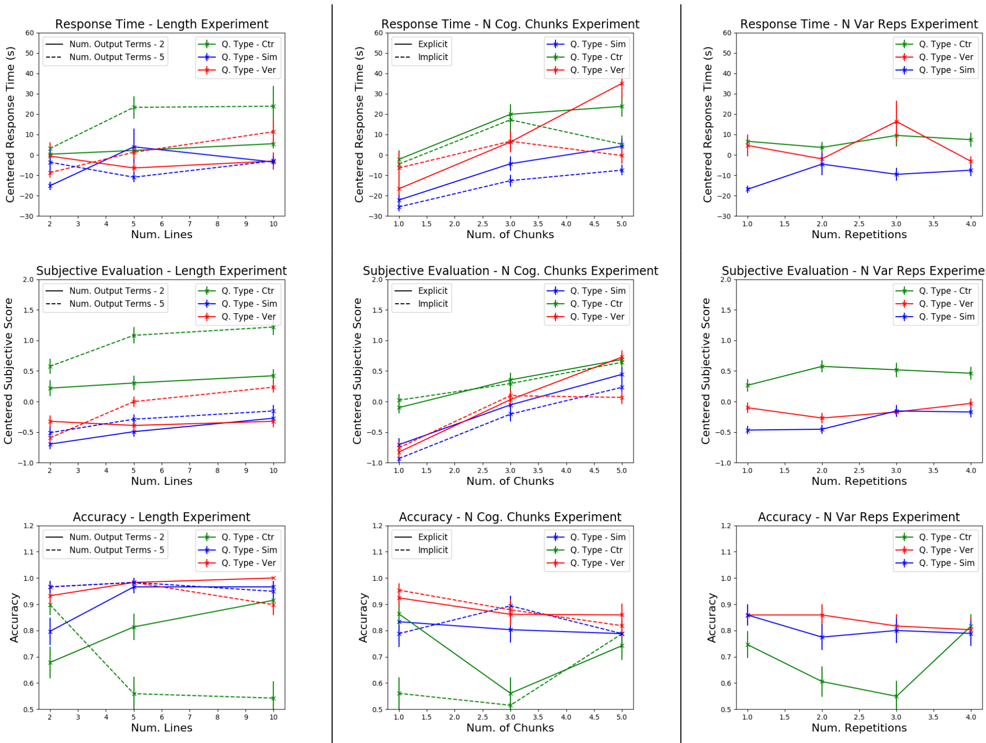
Experimental Design Each experiment consisted of of a block of three questions—one for each task—for each combination of complexity settings. For example, V1 had 6 blocks–1 for each combination of number of lines and terms in the ouput clause. Within a block, the logical structure of the questions was identical, but the questions used different literals (i.e. we changed the words used in the observations) so that they would appear distinct to participants. (In pilot studies, we found this level of obsfuscation was sufficient.) We randomized question order to avoid learning effects, but they were always presented in this order: verification, simulation, counterfactual.

Participants Participants were recruited via Amazon Mechanical Turk. Before starting the experiment, they were given a tutorial on each task and practice questions. They were told that their primary goal was accuracy, and their secondary goal was speed. Participants could only participate in one experiment (V1, V2, or V3). The IRB of Harvard University approved this study.

4 Results and Discussion

We recruited 150 subjects for each of our three experiments for a total of 450 subjects. Participants who did not answer sufficient practice questions correctly were not included in the analysis. We also excluded 5 questions that took longer than 10 minutes to answer from the analysis. US and Canadian participants with moderate to high education dominate this survey, so results may differ for different populations. Our exclusion criteria may artificially increase accuracy, but this criterion helped filter the participants breezing through the experiment for their payment. See appendix for table 2 that summarizes the demographics of our subjects across the experiments, and table 3 that describes the total number of participants that met our inclusion criteria.

Figure 2 presents the response time, subjective score and accuracy across all three experiments. Response time is shown for subjects who correctly answered the questions. Response time and subjective evaluation were normalized across participants by subtracting the participant-specific mean. The statistical significance of each type of complexity's affect on performance was assessed via linear regression for response time and subjective score, and logistic regression for accuracy carried out with the statsmodels library [Seabold and Perktold, 2010]. Table 1 summarizes these results. Bold indicates significant at $\alpha = 0.05$ after Bonferroni correction across all tests of all experiments, where the null hypothesis is that the coefficient is zero.



(a) V1 (b) V2 (c) V3

Figure 2: Accuracy, response time and subjective evaluation. Vertical lines signify standard errors.

The magnitude of performance effect varies by the type of complexity. Across all tasks, across all properties, we see a general trend that increasing complexity increases response time and subjective scores (higher means greater dissatisfaction). However, changes in response time due to the number of variable repetitions are much smaller than changes due to number of cognitive chunks. Subjective response replicates most trends in response time with statistical significance, but fails to replicate the counterintuitive finding described below. Finally, there exist fewer trends with respect to accuracy. We hypothesize this may be because participants were tasked to be as fast as possible while being accurate. That said, counterfactual questions have significantly lower accuracies across all experiments.

Surprisingly, implicit cognitive chunks were faster for people to process than explicit cognitive chunks. Participants took significantly longer to answer when new cognitive chunks were made explicit rather than implicitly embedded in a line. This is surprising because we might have expected that even if the explanation took longer to process, it would have been subjectively easier to follow. It would be interesting to unpack this effect in future experiments.

The type of question significantly impacted response time. Simulation questions were consistently the fastest to answer, and the response times for counterfactual questions were consistently highest. Verification and counterfactual reasoning both involve an extra checking step that seems to introduce a higher cognitive burden. These results suggest that it is important to understand the downstream application when designing interpretable machine learning systems, because the questions people ask of explanations affect how easy they are to use.

Acknowledgements 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

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

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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*, pages 1721–1730. ACM, 2015.

Finale Doshi-Velez and Been Kim. A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608*, 2017.

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

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

Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

Boris Hayete and Jadwiga R Bienkowska. Gotrees: Predicting go associations from proteins. *Biocomputing 2005*, page 127, 2004.

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

B. Kim, C. Rudin, and J.A. Shah. The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In *NIPS*, 2014.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 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*, pages 3–10. IEEE, 2013.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 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*, pages 1675–1684. ACM, 2016.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *ArXiv e-prints*, February 2018.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *NIPS Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments*, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.

Ute Schmid, Christina Zeller, Tarek Besold, Alireza Tamaddoni-Nezhad, and Stephen Muggleton. How does predicate invention affect human comprehensibility? In *International Conference on Inductive Logic Programming*, pages 52–67. Springer, 2016.

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

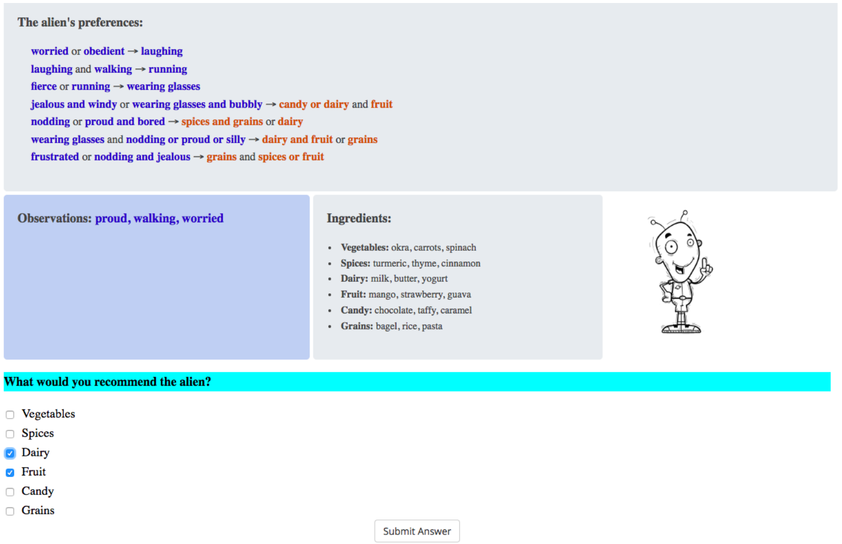
Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391*, 2016.

Sameer Singh, Marco Tulio Ribeiro, and Carlos Guestrin. Programs as black-box explanations. *arXiv preprint arXiv:1611.07579*, 2016.

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

Fulton Wang and Cynthia Rudin. Falling rule lists. In *Artificial Intelligence and Statistics*, pages 1013–1022, 2015.

A Interface



The alien's preferences:

worried or obedient → laughing
laughing and walking → running
fierce or running → wearing glasses
jealous and windy or wearing glasses and bubbly → candy or dairy and fruit
nodding or proud and bored → spices and grains or dairy
wearing glasses and nodding or proud or silly → dairy and fruit or grains
frustrated or nodding and jealous → grains and spices or fruit

Observations: proud, walking, worried

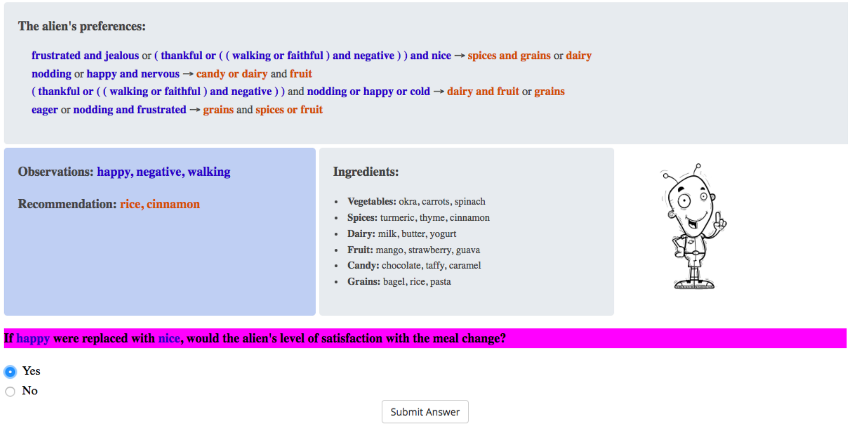
Ingredients:

- **Vegetables:** okra, carrots, spinach
- **Spices:** turmeric, thyme, cinnamon
- **Dairy:** milk, butter, yogurt
- **Fruit:** mango, strawberry, guava
- **Candy:** chocolate, taffy, caramel
- **Grains:** bagel, rice, pasta

What would you recommend the alien?

☐ Vegetables
☐ Spices
☑ Dairy
☑ Fruit
☐ Candy
☐ Grains

Figure 3: Screenshot of our interface for the simulation task. Participants must give a valid recommendation of ingredients that will satisfy the alien given the observations and the alien's preferences.


The alien's preferences:

frustrated and jealous or (thankful or ((walking or faithful) and negative)) and nice → spices and grains or dairy
nodding or happy and nervous → candy or dairy and fruit
(thankful or ((walking or faithful) and negative)) and nodding or happy or cold → dairy and fruit or grains
eager or nodding and frustrated → grains and spices or fruit

Observations: happy, negative, walking

Recommendation: rice, cinnamon

Ingredients:

- **Vegetables:** okra, carrots, spinach
- **Spices:** turmeric, thyme, cinnamon
- **Dairy:** milk, butter, yogurt
- **Fruit:** mango, strawberry, guava
- **Candy:** chocolate, taffy, caramel
- **Grains:** bagel, rice, pasta

If happy were replaced with nice, would the alien's level of satisfaction with the meal change?

● Yes
○ No

Figure 4: Screenshot of our interface for the counterfactual task. Participants must determine whether the alien's satisfaction with the meal changes under the change to the observations described in the magenta box given the observations and the alien's preferences.

B Participants

Feature	Category : Proportion		
Age	18-34 : 57.7%	35-50 : 36.7%	51-69 : 5.6%
Gender	Male : 63.8%	Female : 35.7%	
Education	High School : 34.7%	Bachelor's : 52.6%	Masters and Beyond : 9.2%
Region	US/Canada : 93.9%	Asia : 4.1%	

Table 2: Participant Demographics. There were no patients over 69 years old. 3.5% of participants reported "other" for their education level. The rates of participants from Australia, Africa, Europe, Latin America, and South America were all less than 1.0%. (All participants were included in the analyses, but we do not list specific proportions for them for brevity.)

Experiment	Number of Participants
Explanation Size (V1)	N=59
New Cognitive Chunks (V2)	N=62
Variable Repetition (V3)	N=70

Table 3: Number of participants who met our inclusion criteria for each experiment.