What common error patterns can tell us about human problem solving

Caroline Ahn (ahncj@bu.edu)
Graduate Program for Neuroscience, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Quan Do (qdo@bu.edu)
Graduate Program for Neuroscience and Center for Systems Neuroscience
Boston University, Boston, MA

Leah Bakst (lbakst@bu.edu)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Michael Pascale (mpascale@bu.edu)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Jingxuan Guo (jxguo21@bu.edu)
Graduate Program for Neuroscience, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Joseph T. McGuire (jtmcg@bu.edu)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Michael E. Hasselmo (hasselmo@gmail.com)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Chantal E. Stern (chantal@bu.edu)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Abstract:

This study examines human abstract reasoning using the Cognitive Abstraction and Reasoning Corpus (CogARC), a visuospatial task inspired by an AI competition and adapted here to assess human problem-solving strategies. We analyzed online behavioral data from 233 participants who engaged in few-shot learning to learn input-output transformation rules from limited examples and apply these to novel problems. Our human subjects (M = 78.9% accuracy) significantly outperformed competing AI programs in the task. While the performance data shows considerable subject- and task-level variability, DBSCAN clustering of first attempt solutions also reveals that on certain tasks, a substantial proportion of participants made similar errors. The findings suggest shared cognitive biases in human abstract reasoning and suggest directions for future research to explore the representational space of problem-solving.

Introduction

One of the hallmarks of human intelligence is our ability to extract generalizable rules from limited information and apply these rules across contexts in a speedy and flexible manner. This allows us to predict outcomes in novel situations on the fly, make analogies to understand and communicate unfamiliar concepts, and come up with creative solutions for complex problems, which artificial intelligence is, to date, less able to do[1].

However, human abstract reasoning is not infallible. We believe the same mechanisms that enable speed and flexibility in reasoning can also lead to systematic error patterns. In the current study, we set out to characterize shared error patterns in human abstract problem solving to better understand the cognitive processes underlying abstraction and generalization, asking how humans determine what is the correct level of abstraction to efficiently learn and transfer rules for a task. Exploring the differences between successful and erroneous rule learning may shed light on the shared biases or assumptions that lead to different outcomes, paving the way for future theoretical investigations.

To study this aspect of human problem solving, we tested human participants with the Cognitive Abstraction and Reasoning Corpus (CogARC), a visuospatial task of abstract reasoning. The original ARC task dataset was introduced on the website Kaggle as a benchmark for AI abstract reasoning and generalization but has also been used for studying strategies in problem solving in humans[2,3,4]. In our CogARC variant of the task, subjects generate solutions dynamically on an interactive browser-based interface (**Fig. 1**). Our work is distinguished from previous studies in humans also inspired by the AI benchmark by our focus on how task rules and human biases can drive different distributions of shared errors. We are also interested in understanding error correction after feedback.

Methods

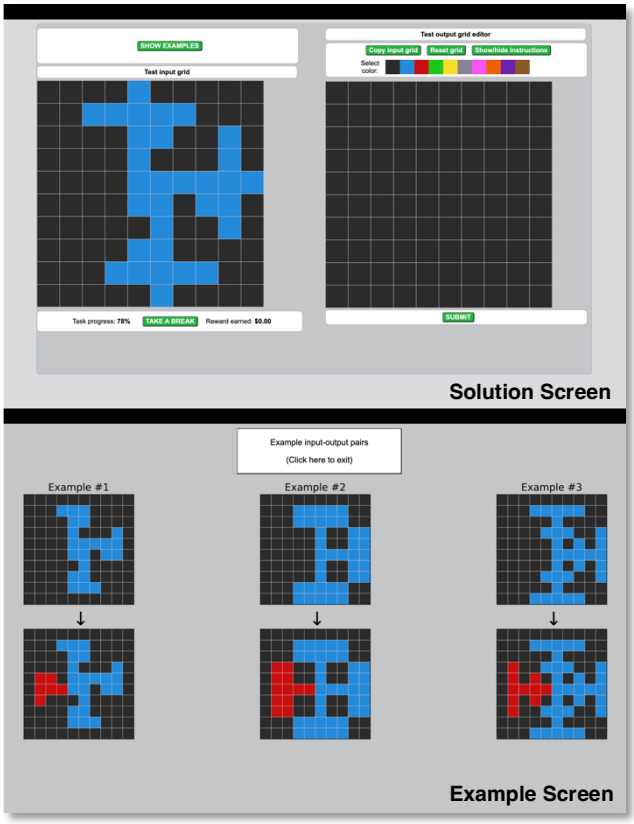

Figure 1: An example of the browser-based task interface. Participants could toggle between the 'Example' screen and 'Solution' screen.

We collected online behavioral data on the CogARC task from 233 participants (52.27% male, 47.27% female, 0.45% other) on the crowdsourcing website Amazon Mechanical Turk (MTurk). Participants ranged in age from 20 to 35 years (M = 29.6, SD = 4.1). 75 problems were selected from the original ARC task database to represent rules of a variety of types and difficulty levels. In each trial, participants learned input-output transformation rules by studying two to six example pairs, then applied the learned rules to a test input by editing tiles in the test output grid. Participants were allowed up to three attempts per problem to reach the correct solution. To study the strategies underlying problem solving in the task, and to identify common

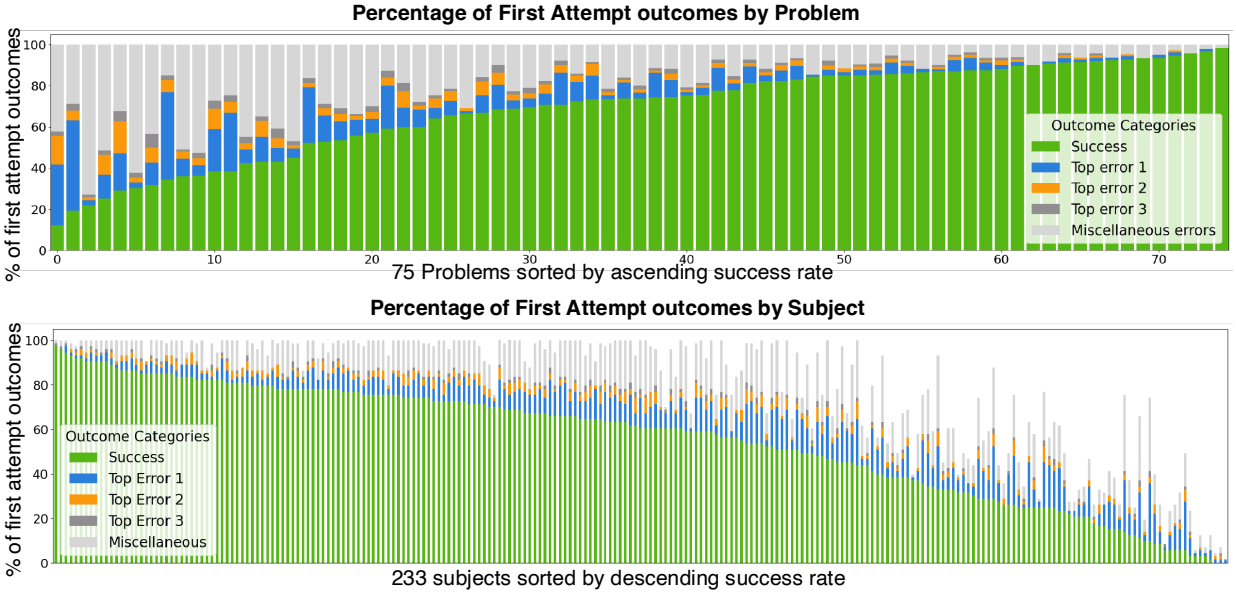

Figure 2: a) The percentage of first attempt outcomes per problem. Problems are ordered in ascending order by success rate. Problems each had between 155 and 200 participants. **b)** The percentage of first attempt outcomes per subject, ordered by success rate in descending order. Percentage expressed out of the full 75 problems; participants completed between 1 and 75 problems.

strategies or errors shared among participants, we used DBSCAN clustering to identify the top three common errors for the first attempt[6]. We also looked at the final success rates for each error category.

Results

Human performance on the CogARC task (M = 78.9%, SD = 19.4% accuracy) greatly outperformed AI programs that were submitted as part of the Kaggle challenge (29.33% accuracy, 22/75 tasks solved from top performer[5]). Clustering and analysis of first attempts show both problem-level and subject-level variability in patterns of common errors. The 75 problems presented in CogARC show a broad distribution of difficulty according to success rate. Interestingly, some tasks have higher ratios of shared errors compared to others (**Fig. 2**). The problem with the highest percentage of shared common errors had 44.07% of subjects submit similar wrong answers during their first attempt. All participants made at least one common error during their first attempts.

An in-depth look into the top errors for one of the problems, selected among the top five most difficult to solve, offers some insight into error generation and post-error processing (**Fig. 3**). In this problem, the rule is to complete a 'pinwheel' type pattern in red. The three most common errors for this problem all demonstrate an attempt at the "pattern completion" necessitated by this problem, and are similar in shape to the correct solution, yet fall short for different reasons – Top Error 1 the red portion is attached to the wrong central location. In Top Error 2, the red portion is mirrored instead of rotated. Top Error 3 is correct but employs the wrong color and is associated with a higher final success rate compared to the first two errors. This suggests that color might be a bias that is easier to correct or overcome than location or rotation. This result suggests that in this specific problem, there exists a variety of biases among subjects that lead to shared erroneous rule representations. Future work will investigate if this pattern holds true for all problems. In summary, the analyses of patterns of error generated by human subjects across CogARC problems reveal insights into the common biases used to rapidly and flexibly solve abstract reasoning problems. Further analysis into these error patterns can help us understand human cognitive strategies, as well as inform frameworks of cognitive biases in computational models of reasoning.

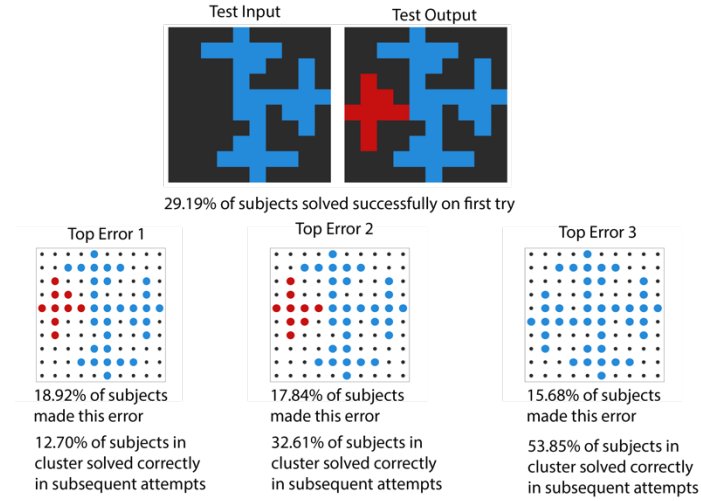

Figure 3: Breakdown of first attempt results for the 185 subjects that submitted solutions for the 'pinwheel' problem.

Acknowledgements

This work is supported by the Office of Naval Research ONR MURI N00014-19-1-2571, ONR MURI N00014-16-1-2832, and a Rajen Kilachand Fund for Integrative Life Science and Engineering Award.

References

1. Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in artificial intelligence*, *4*, 622364.

2. Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

3. Johnson, A., Vong, W.K., Lake, B.M. and Gureckis, T.M. (2021). Fast and flexible: Human program induction in abstract reasoning tasks. *arXiv preprint arXiv:2103.05823*.

4. Acquaviva, S., Pu, Y., Kryven, M., Sechopoulos, T., Wong, C., Ecanow, G., Nye, M., Tessler, M. and Tenenbaum, J. (2022). Communicating natural programs to humans and machines. *Advances in Neural Information Processing Systems*, *35*, pp.3731-3743.

5. Odouard, V.V. and Mitchell, M. (2022). Evaluating understanding on conceptual abstraction benchmarks. *arXiv preprint arXiv:2206.14187*.

6. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.