Predicting Insight during Physical Reasoning

Solim LeGris[1]**, Brenden M. Lake**[1,2]**, & Todd M Gureckis**[1]
{solim.legris, brenden, todd.gureckis}@nyu.edu
Department of Psychology[1] and Center for Data Science[2]
New York University

Abstract

When people solve problems, they may try multiple invalid solutions before finally having an insight about the correct solution. Insight problem-solving is an example of the flexibility of the human mind which remains unmatched by machines. In this paper, we present a novel experimental paradigm for studying insight problem-solving behavior in a physical reasoning domain. Using this paradigm and several data-driven analyses, we seek to quantify what it means to have an insight during physical problem-solving and identify behavioral traces that predict subjective insight ratings collected from human participants. This project aims to provide the first steps towards a computationally informed theory of insight problems solving.

Keywords: insight; problem-solving; physical reasoning

Introduction

Imagine you are watching TV and suddenly water starts dripping from the ceiling directly above the TV. Unfortunately, you can't move the TV quickly because it is fixed to the wall so you must find another solution to avoid damage. You have a bucket handy in the basement but you just can't get the bucket to stand on its own above or around the TV. While you're holding the bucket to prevent the water from damaging the TV, you're faced with an impasse. What could you do to avoid having to hold the bucket collecting the water? You're starting to get frustrated when suddenly you think of using your small surfboard. You quickly go grab it and put it flat on top of the TV such that the water drips into the bucket.

People often experience situations where a problem arises and the solution to their problem is not immediately obvious. Whether it is due to a failure to retrieve the right prior knowledge or not thinking about the problem in the right way, people may not initially know how to solve a problem, but at an unpredictable moment, they will suddenly see the solution.

Researchers have spent considerable effort in understanding and characterizing problem-solving behavior (Newell & Simon, 1972). Problem-solving behavior has been operationally divided into two distinct classes: analytical problem-solving and insight problem-solving (Gilhooly & Murphy, 2005). Typically, analytical problems have been problems for which a logical sequence of steps can be applied to reach a solution. Importantly, people can report on the steps they took to generate their solution. Conversely, when people solve insight problems, they tend to find the solution in a sudden and unpredictable moment of clarity and have trouble reporting on how they got the solution (Kounios & Beeman, 2014; Gilhooly & Murphy, 2005; Chronicle, MacGregor, & Ormerod, 2004). Insight problem-solving has been studied extensively in the cognitive science literature, with problems such as the nine-dot problem (Maier, 1930), the candle/box problem (Duncker & Lees, 1945), the radiation problem (Duncker & Lees, 1945), and others. *How* people solve insight problems and *why* some problems require insight is still not well characterized computationally and remains a standing issue.

We propose that a major hurdle in advancing our understanding of this cognitive phenomenon has been the use of low-resolution measurement devices: experiments that don't generate rich real-time behavioral data that could be used to help pinpoint the origin of insights. Similar to previous approaches (Jones, 2003; Stephen & Dixon, 2008; Berckley & Hattie, 2023) that have utilized high sampling rate of actions, eye movement and even facial expressions to study insight, we do so in a 2D game environment specifically designed for studying insight and focus directly on sampling behavior that supports problem solving. We designed a physical reasoning game that attempts to expose the underlying cognitive mechanisms supporting insight problem-solving by externalizing some aspects of cognition. Our game allows us to record at a high sampling rate the actions people take as well as the information people sample while solving problems. The task is designed to require constant interaction from the participants by introducing a "pressure to act", thereby limiting offline processing. We take inspiration from recent work investigating tool use in a virtual physical environment (Allen, Smith, & Tenenbaum, 2020). Expanding upon this work, we 1) introduce a dynamic and interactive component to the task and 2) focus on insight problem-solving behavior.

Although multiple cognitive processes underlie insight problem-solving, we focus on understanding and quantifying behavioral signatures of insight by adopting a data-driven approach. More specifically, we are interested in dissociating between three potential sources of insight: **individual differences in problem-solving behavior**, **general properties of problems** that might make it more likely that people will initially reason about them in the wrong way, and **aspects of behavior that support having an insight**. Similarly to previous research, we think of insight as the ability to find a solution after having initially thought about the problem in

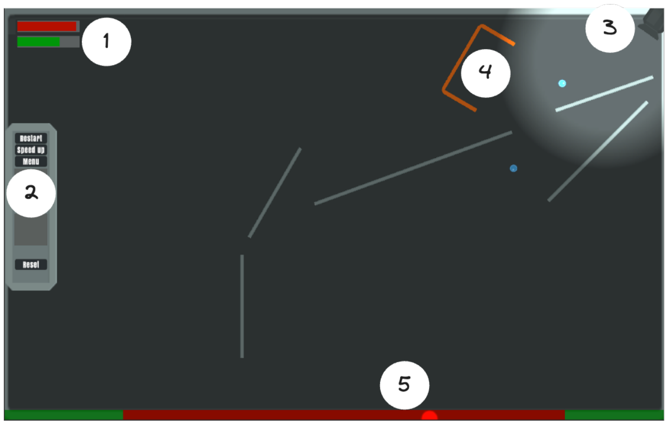

Figure 1: Participant replay screenshot. For illustrative purposes, the full problem is slightly illuminated. At the top (1), red and green bars show accumulated penalties and rewards. Participants must use tools (4) dragged from the inventory (2) to redirect balls shot from a cannon (3) into green zones. The inventory and the red and green bars at the top are constantly visible throughout gameplay. Visual feedback (i.e., ball glow) (5) is given when a ball falls into a red or green zone.

the wrong way, suggesting inadequate or sub-optimal solutions (Öllinger, Jones, & Knoblich, 2014; Knoblich, Ohlsson, & Raney, 2001). As a first step towards understanding the computational basis of insight problem-solving, we examine real-time behavioral signatures and fit statistical models predicting participant self-reports of insight.

Experiment

A physical reasoning game

Games are becoming increasingly popular experimental tools for understanding the mind (van Opheusden et al., 2023; Brändle, Binz, & Schulz, 2021; Allen et al., 2020). A major advantage of using games to study the mind is that they afford tasks that are closer to real-world complexity than traditional psychological experiments yet they are controllable and engineered environments (Allen et al., 2023). With this in mind, we developed a physical reasoning game to study insight problem-solving behavior.

Our game is inspired by previous games used to study physical reasoning (Allen et al., 2020) as well as mobile phone games such as Enigmo. In our 2D game (see Figure 1)[1], a ball cannon shoots balls at a constant rate of 1Hz. The participant must use available tools to redirect them into one or multiple green zones at the bottom while avoiding the red zones. The goal is the same for each of the twenty problems we created: maximize rewards (balls in green zone) and minimize penalties (balls in red zone or that time out). Penalties and the constant flow of balls impose a pressure to act, encouraging participants to act while they think. The trial ends when either the rewards or penalties bar reaches full capacity, resulting in a win or a loss, respectively. The maximum capacity of each bar is set to 95 balls.

Each problem/level varies in the following dimensions:

- Position, angle, and length of walls
- Length, count, and position of green and red zones
- Position and angle of ball cannon
- Number of tools $k \in \{0,1,2\}$ available for each type
- Initial velocities of balls

The game provides two types of tools which we name containers and platforms. Some problems only allow the use of one type while others allow the use of both types. We designed the problems to vary in difficulty and obviousness of solution. While certain problems simply require the participant to reason about the physics of the scene and implement an easy solution, such as blocking a moving object, other problems require participants to come up with creative and less obvious solutions. The problems can be solved using intuitive actions like **blocking**, **containing** and **redirecting**. These concepts are useful to solve the problems but many of them require going beyond these obvious uses and actions. For example, participants might need to notice an inconspicuous gap and use it to redirect the balls appropriately towards a smaller and less obvious green zone. Yet, another problem might require using a container for purposes other than containing (i.e., as a platform). We aimed for varied solutions which differ from each other to avoid transfer between problems as much as possible.

The different dimensions along which our problems vary allowed us to restrict the solution space. We designed several of our problems to have locally sub-optimal solutions. We define a sub-optimal solution as a solution which, even if it were implemented at the first time step of the game, would win but not achieve the maximum score for that problem. Moreover, we aimed to structure the problems such that the sub-optimal solutions appeared more intuitive than the harder, but better optimal ones.

At any time point during gameplay, only a small circular window around the participant's cursor is illuminated (see Figure 1). Given that the scene is always dark other than around the cursor, the participant must move their cursor to gather information, shift their attention to the different elements of the scene, and search for solutions. This feature allows us to externalize cognitive processes supporting problem-solving. For example, by tracing the focus of this foveated window, we can identify elements within the scene that were considered to generate solutions. Participants simultaneously engage in information search and problem-solving by progressively constructing a mental representation of the environment and using it to inform their search for a solution.

[1]See the project website for a demo of our task and additional visualizations: `https://gureckislab.org/papers/#/ref/legris2024physicalsolving`

Methods

Participants. We recruited 140 participants (62.9 % male, 35 % female, 2.1 % other) on Prolific. Participants varied in age from 19 to 70 (M= 39.6, SD=11.9) and in weekly gaming time (M=7 hours, SD=4.49 hours). They were compensated $7.50 for an expected 30 minutes of their time plus a potential bonus of up to $1.50 calculated using performance on a randomly selected problem.

Design. Each participant completed ten randomly sampled problems out of the twenty total tasks in a random order and was given a single attempt at each problem.

Procedure. In the first part of the experiment, participants were given general instructions about the experiment. They were informed that:

- After every problem, they would be asked about whether they experienced an Aha! moment (defined below).
- Following the tutorial, they would be tested on their understanding of the instructions before proceeding to the second part of the experiment.
- A perfect solution exists for every level and it does not require any further movement of the tools once implemented.

The participants then completed a tutorial with step-by-step instructions explaining every element of the game, controls, and scoring. Finally, the participants were asked to answer seven comprehension questions to advance to the experiment. If they failed the quiz three times, the participants were returned to the general instructions and tutorial. This was done to ensure that participants were comfortable using the game interface and would follow instructions carefully.

Gameplay. At the beginning of each problem, the game environment is completely blacked out. Participants only see the inventory and score bars. After a three-second countdown, balls are shot at a constant rate into the environment and a light appears around the cursor allowing the participants to discover the structure and dynamics of the problem. During each trial, the red and green zones at the bottom are hidden but visual feedback (temporary green or red glowing light around the ball) is shown whenever a ball falls into one of the zones. This provides clues to the participant about the position of the zones and signals whether points were gained or lost. Additionally, at the top of the screen, green and red bars show accumulated rewards and penalties throughout the trial. Each level ended when the participant either reached the maximum number of green or red points. The participant's final score was shown at the end of each problem.

Post-problem insight rating. After each problem, the participants were asked whether they experienced an Aha! moment while solving the problem. Aha! moments were described as *initially not knowing the solution and subsequently having a sudden realization while solving the problem*. Conversely, not having an Aha! moment was described as *finding the solution incrementally without any sudden realization or not solving the problem*. If the participant answered that they had an Aha! moment, they were then asked to rate the strength of their experienced Aha! moment on a scale of 1 to 5. Participants who did not report an insight were simply assigned an insight rating of 0.

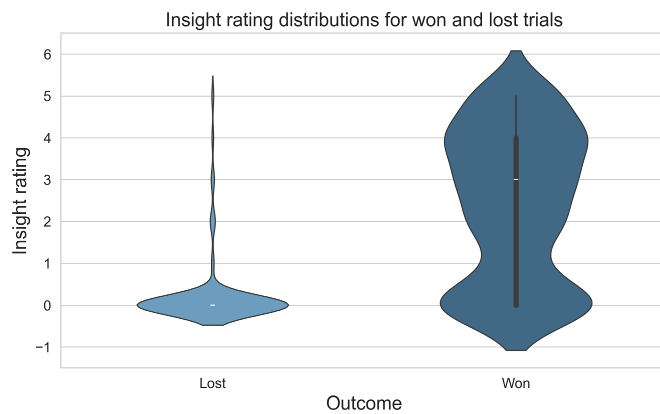

Figure 2: Violin plots for reported insight ratings for both won and lost trials.

Data collection

In total, we obtained 1329 complete trials from our 140 participants. We first collected data from 40 participants on 20 problems and computed the average insight score for each problem. We then split the ordered problems into a high insight group and a low insight group, randomly sampling 2 problems from each category to be used as held-out problems. Additionally, we replaced one problem with a variation for which we had not previously collected any data and included it in the held-out problems for a total of five held-out problems. Finally, we also randomly held out 10% of participants entirely for testing. Our final dataset has a split of 71% in the training set and 29% in the held-out dataset which we omit from this study and keep for future work.

Results

General behavioral results

We found strong variation in task difficulty, covering a wide range of problem solution rates and scores. Participants had an average solution rate (frequency of wins) across problems of 48% (SD = 20%) and solution rates ranged from the hardest problem at 10% to the easiest problem at 88%. Although solution rates are indicative of problem difficulty, mean problem score is also informative since a problem could, for example, have a high solution rate but low mean score. This would indicate that people tend to find a solution but in doing so accumulate a lot of penalties. Problems that have more obvious solutions but require precise fine-tuning are an example of such instances. A participant won if they obtained a score of 1 or more up to a maximum of 95. A score of 0 corresponds to an unsuccessful attempt at a problem. For example, if a participant obtained a score of 65, then they accumulated

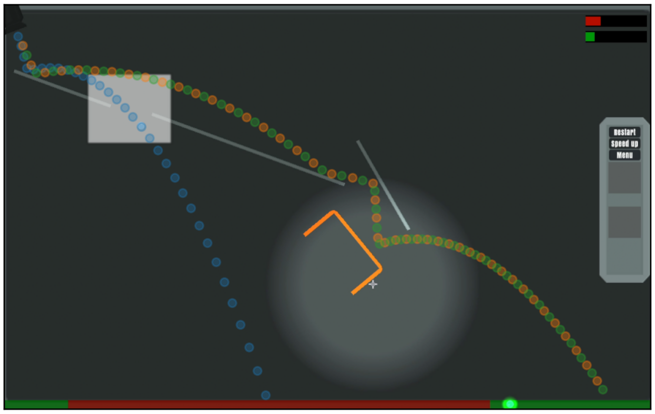

Figure 3: Participant replay screenshot with ball traces. For illustrative purposes, the full problem is slightly illuminated. The white square in the top left corner represents a region of interest (ROI), a part of the game scene relevant to finding the optimal solution. As illustrated by the ball traces, 1/3 of the balls (blue here) fall through the gap where the ROI is situated while 2/3 of the balls (orange and green here) are shot further. The solution shown is suboptimal since not all balls fall into a green zone. The optimal solution requires placing the tool in the ROI such that it redirects both streams of balls into the left green zone.

95 reward points and 30 penalty points. The mean problem score was 20.81 (SD = 27.32), with a minimum mean problem score of 4.11 and a maximum of 58.25.

We found that participants report an insight in 37% of trials and that 44% of trials were wins. Participants that won report an insight rating of 2.40 (SD = 1.81) on average while participants that lost report an insight rating of 0.23 (SD = 0.84) on average. We also observed substantial variation in reports of insight per problem (see Figure 4). The average insight rating across tasks was 1.24 (SD = 1.75) with a minimum of 0.21 and a maximum of 2.42.

Qualitative inspection of participant replay data revealed a wide range of creative solutions for each problem but also suggested that people often get stuck. Similar to results commonly described in the literature, we observed instances of mental set. Mental set occurs when the problem-solver persists at trying solutions within a restricted part of the solution space, not exploring other potential options (Ollinger, Jones, & Knoblich, 2008). As shown and described in Figure 3, participants sometimes find suboptimal solutions and subsequently fail to find other potential and better solutions. Additionally, we also observed instances of functional fixedness: the failure to use objects in ways other than their intended use. This is particularly exemplified in problems where the container tool must be used as a platform to redirect balls. Finally, we also observed general failures of noticing how certain elements of the problems can be used to solve it, even when the participant explored the full scene with their cursor.

Moreover, we found that participants implement two qualitatively different kinds of solutions across levels: *static* and *dynamic* solutions. Static solutions correspond to solutions where the participant, once satisfied with their solution, does not move or rotate the tools until the end of the trial. Conversely, in dynamic solutions, participants move and rotate the tools to redirect the balls. For example, a participant might wait for the container to fill up then move it to a green zone to drop the balls in it, repeating this sequence of actions until they win or lose the game. We algorithmically determined which participants had implemented a static solution by determining whether there existed a sufficiently long window (set to approximately 10 seconds) in their replay data within which the cumulative displacement of tools was smaller than 1 unit and rotation was smaller than 2 degrees until the end of the trial. These values were selected to ensure that only solutions with minor or inconsequential adjustments maintained for at least 10 seconds from the end of the trial were counted as static. We determined that 76% of trials from our training set correspond to static solutions and that 99% of successful trials (i.e., wins) were trials where the participant found a static solution. Given that we were only interested in information preceding or surrounding the moment of insight, we truncated all trials where a static solution was found to the time point when the solution was implemented. For trials where no static solution was found, we use the full behavioral trace. All subsequent analyses were made using data cleaned up in this way.

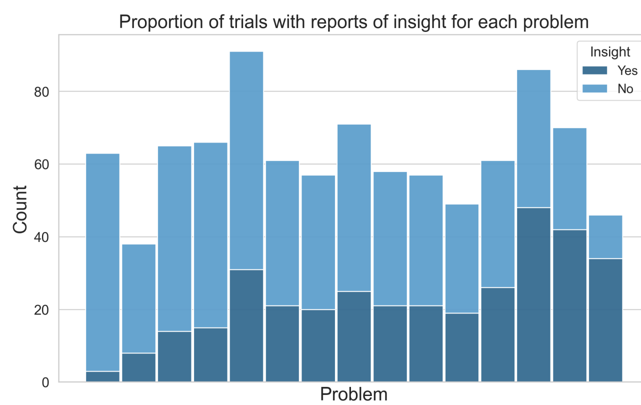

Figure 4: All problems attempted by participants, ordered by proportion of insight reports. We observe a wide spread of insight reports between problems.

Predicting insight

To determine aspects of behavior and the different problems that are predictive of reporting an insight, we fit mixed effects logistic regression models using features computed from participant replay data. We partitioned the data into insight and no-insight based on whether participants reported an Aha! experience or not. From qualitative inspection of participant replays, we identified three sources of information from which we computed various features hypothesized to be predictive of insight. In total, we generated ten features for all trial data.

Each feature's associated fit parameter name is indicated in parentheses.

General information gathering. Given that our experimental paradigm requires participants to actively seek information that will enable them to solve each problem, we propose two features relating to this behavior.

Accumulated information (β_1): This feature corresponds to how much of the problem has been visually explored. How much one explores is bound to influence what elements of the scene are part of one's mental representation and will thus inform search for a solution. We thus computed the percentage of the scene illuminated using cursor movement from participant replays. This measure excludes green zones and regions of interest (defined below).

Initial search (β_2): This feature corresponds to how much time the problem-solver spent exploring before dragging a tool from the inventory. It is meant to capture how much time was spent getting a general sense of the structure of the problem before acting. We thus computed the number of time steps spent moving the mouse before a tool dragged from the toolbox and normalize by total time spent solving the problem.

Regions of interest. Each problem in the game has a different structure and certain elements are crucial to finding the best solution while other elements are irrelevant/distracting. Focusing on the distractors and failing to identify the relevant elements should lead to losing or implementing a suboptimal solution. For each problem, we manually identified areas using bounding boxes that we call regions of interest (ROIs, see Figure 2 for an example). We also include in this section features related to green zones.

ROI/green zone information (β_3, β_4): Similarly to accumulated information, this feature corresponds to how much of ROIs were illuminated by the cursor. We computed the same measure for green zones separately.

Time spent in ROIs/green zones (β_5, β_6): This feature serves as a proxy for time spent processing relevant elements of the scene. It is computed as the number of time steps spent looking at ROIs relative to the total time spent finding a solution. Again, we compute the same measure for green zones separately.

Number of switches between ROIs and green zones (β_7): This feature represents the number of times the participant switched between looking at the different regions, considering the combined set of ROIs and green zone regions. We count mouse movements between ROIs and green zones as switches if they occurred in 1s or less.

Outlier mouse movement. It is a characterizing aspect of insight moments that they are sudden and unpredictable (Danek, 2018; Ohlsson, 1992). Thus, we hypothesized that this suddenness could be captured using outlier mouse movements. Mouse movement was quantified as the distance the cursor moved between each frame. Analyses of mouse data suggested that mouse movement from frame to frame was exponentially distributed. Thus, for every participant's mouse data, we fit an exponential distribution and identified as outliers any value in the top 2.5%. We only consider the last 10 seconds of a trial.

Table 1: Fixed effects odds ratios and 95% confidence intervals for full mixed effects model
Signif. codes: '***': p < 0.001 and '**': p<0.01

	OR	2.5%	97.5%
Acc. information** (β_1)	0.73	0.58	0.92
Initial search*** (β_2)	1.56	1.29	1.89
Green zone information (β_3)	1.15	0.96	1.40
Time spent in ROIs*** (β_5)	1.68	1.33	2.11
Time spent in green zones (β_6)	0.89	0.70	1.10
Number of switches (β_7)	1.21	0.95	1.60
Percentile largest mouse movement (β_8)	1.17	1.00	1.89

Percentile of largest mouse movement in the last 10 seconds (β_8): This feature corresponds to how extreme the largest mouse movement was in the last 10 seconds.

Number of outlier mouse movements into and out of ROIs/green zones in the last 10 seconds of the trial (β_9, β_{10}): This feature quantifies how many outlier movements in the last 10 seconds of a trial were from or into ROIs or green zones, separately.

Regression models. All features used for statistical modeling were z-scored. We fit a mixed effects logistic regression model with random intercepts for problem and participant predicting insight. Log likelihood ratio tests found significant effects of accumulated information (β_1), $\chi^2(1) = 7.70$, $p < .01$, initial search (β_2), $\chi^2(1) = 20.71$, $p < .001$ and time spent in ROIs (β_5), $\chi^2(1) = 19.02$, $p < .001$. All other fixed effects were not found to be significant. We report odds ratios for all effects which were found to increase or decrease odds by more than 10% (see Table 1). Odds ratios are obtained by computing the inverse natural logarithm of the log odds for each predictor. Inspecting odds ratios of the significant fixed effects, we found that time spent in ROIs has the strongest positive effect, increasing odds of reporting an insight by 1.67 for each unit (1 SD) of increase of the predictor. Similarly, the odds of reporting an insight increase by 1.56 for every unit of increase of initial search. Conversely, accumulated information was found to decrease the odds of reporting an insight by 0.73 for every unit of increase.

Inspecting random effects, we find that baseline (i.e., all features held equal at mean value) probability of insight for 95% of participants (SD=0.89) lies in the interval (0.08, 0.74) suggesting that some participants almost never report an insight while others do so very often. Additionally, we find that problem random intercepts explain some of the variance in odd of reporting an insight (SD=0.53).

In Table 2, we report within-sample model accuracy and F1 scores. F1 scores measure predictive accuracy by accounting for both precision and recall. A log likelihood ratio test con-

Table 2: Model accuracy comparisons

	Accuracy	F1
Fixed effects only	69.4%	0.50
Participant random intercepts	76.5%	0.61
Mixed effects model	77.6%	0.66

firms a statistically significant fit of the mixed effects logistics regression model compared to a fixed effects only model, $\chi^2(2) = 43.47$, $p < .001$. We observe strong random effects of participant resulting in an increase of 7.1% in predictive accuracy compared to the fixed effects only model. Finally, we find a marginal increase in predictive accuracy when we include our full set of features compared to the random intercepts only model.

Since insight and winning are tightly related (that is, in our task many insights result in winning), we wondered whether any features could predict insight even amongst those who won. Thus, we ran regressions that fit to the successful trials only. This subset of data contains 44% of the training set (417 trials). The insight report rate for this subset was found to be 74%, which is much higher than that of the full data set. Since a static solution was implemented in nearly all successful trials, we truncated the data at the time point when the static solution was implemented. We fit a mixed effects logistic regression on this subset with an additional predictor: whether the solution was optimal or not. To determine optimality of implemented solutions, we algorithmically verified for each trial that, within some margin of error, all balls were redirected to green zones for at least 5 seconds from the end of truncated trials. Likelihood ratio tests found significant effects of initial search (β_2), $\chi^2(1) = 4.44$, $p < .05$, percentile of the largest mouse movement in the last 10s of a (truncated) trial (β_8), $\chi^2(1) = 10.67$, $p < .01$, and optimal solution, $\chi^2(1) = 9.22$, $p < .01$. The strongest predictor of insight was optimality, increasing the odds of reporting an insight by 5.79. The percentile of the largest mouse movement was also found to be a strong predictor of insight, increasing the odds by 2.17. Interestingly, this model suggests that for successful trials, there is no variability between problems in odds of insight.

Discussion

We find that our best statistical model predicts insight with an accuracy of 77.6% on our training set. As is typical in psychological experiments, individual variability is one of the strongest contributors to predictive accuracy for our tasks. Nonetheless, we identify several significant predictors of insight in our task. We find that time spent in ROIs is the strongest predictor of insight in our main mixed effects logistic regression. Given that ROIs capture the relevant elements of each problem, it is expected that participants that spent more of their time looking at ROIs are likely to have found a solution and this may have resulted in an insight. Our model also suggests that differences in strategy lead to different outcomes in insight reports. Initial search, which corresponds to time spent gathering information about the problem, is positively predictive of reporting an insight. A potential explanation is that participants that spend a lot of time searching before acting don't initially find the solution obvious and/or are searching for a solution in a directed manner. Given that this feature is also significant in the model fit on successful trials only, it is disambiguated from solely predicting success. This supports our hypothesis that it is predicting insight because success was preceded by not knowing immediately what the solution was. Counter to our initial hypothesis, we find that accumulated information is a negative predictor of insight. A large proportion of each problem is empty space that has no relevance to the solution. Although still unclear, participants that uncover the full scene may also be those that are searching for clues in irrelevant regions because they have not found a solution. Finally, in the model fit on successful trials only, we find that participants that found an optimal solution are substantially more likely to report an insight. This result suggests that optimal solutions in our problems appear less obvious to participants. Accordingly, when such solutions are found, they are likely to have been found as a result of an insight.

These results only begin to uncover the behavioral signatures of insight that will allow us to characterize insight problem-solving computationally. Although we achieve relatively good accuracy at predicting insight, more research is required to disambiguate the different sources of insight. A limitation of the current work is that our analyses relied on participant self-reports without reference to an independent measure of insight such as strategy change within a trial which is difficult to capture because of the complexity of the dynamics of our task.

Although our ultimate goal is to define insight separately from the subjective experience that people report post-hoc, we believe that the current approach is a first step to understand the cognitive underpinnings of insight problem-solving. Despite the difficulty of observing and studying this covert cognitive phenomenon, our real-time problem-solving task seeks to expose aspects of this behavior. We use the high-resolution data generated by participants to develop a predictive model of insight and find behavioral traces that are significant predictors of insight. These features suggest ingredients necessary for a computational model of insight which will ultimately lead to an improved understanding of the flexibility of human problem-solving ability.

Acknowledgments

We thank Rheza Budiono, Guy Davidson, Pat Intara, Emily Liquin, Alexa Tartaglini and other lab members for helpful feedback on the manuscript and project. We also thank the anonymous reviewers for their valuable feedback. This work was generously supported by NSF BCS grant 2121102 to TMG.

References

Allen, K. R., Brändle, F., Botvinick, M., Fan, J., Gershman, S. J., Gopnik, A., … al., E. (2023, February). *Using games to understand the mind.* doi: 10.31234/osf.io/hbsvj

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020, November). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proc. Natl. Acad. Sci. U. S. A.*, *117*(47), 29302–29310. doi: 10.1073/pnas.1912341117

Berckley, J., & Hattie, J. (2023, September). Making learning visible: Observable correlates of the aha! moment when moving from surface to deep thinking. *J. Creat. Behav.*, *57*(3), 439–449. doi: 10.1002/jocb.589

Brändle, F., Binz, M., & Schulz, E. (2021, February). *Exploration beyond bandits.* doi: 10.31234/osf.io/9fnmj

Chronicle, E. P., MacGregor, J. N., & Ormerod, T. C. (2004, January). What makes an insight problem? the roles of heuristics, goal conception, and solution recoding in knowledge-lean problems. *J. Exp. Psychol. Learn. Mem. Cogn.*, *30*(1), 14–27. doi: 10.1037/0278-7393.30.1.14

Danek, A. H. (2018, March). Magic tricks, sudden restructuring, and the aha! experience. In *Insight* (pp. 51–78). Routledge. doi: 10.4324/9781315268118-4

Duncker, K., & Lees, L. S. (1945). On problem-solving. *Psychol. Monogr.*, *58*(5), i.

Gilhooly, K. J., & Murphy, P. (2005, August). Differentiating insight from non-insight problems. *Think. Reason.*, *11*(3), 279–302. doi: 10.1080/13546780442000187

Jones, G. (2003, September). Testing two cognitive theories of insight. *J. Exp. Psychol. Learn. Mem. Cogn.*, *29*(5), 1017–1027. doi: 10.1037/0278-7393.29.5.1017

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001, October). An eye movement study of insight problem solving. *Mem. Cognit.*, *29*(7), 1000–1009. doi: 10.3758/bf03195762

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annu. Rev. Psychol.*, *65*, 71–93. doi: 10.1146/annurev-psych-010213-115154

Maier, N. R. F. (1930, April). Reasoning in humans. i. on direction. *J. Comp. Psychol.*, *10*(2), 115–143. doi: 10.1037/h0073232

Newell, A., & Simon, H. A. (1972). *Human problem solving*.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. *Advances in the psychology of thinking*, *1*, 1–44.

Ollinger, M., Jones, G., & Knoblich, G. (2008). Investigating the effect of mental set on insight problem solving. *Exp. Psychol.*, *55*(4), 269–282. doi: 10.1027/1618-3169.55.4.269

Öllinger, M., Jones, G., & Knoblich, G. (2014, March). The dynamics of search, impasse, and representational change provide a coherent explanation of difficulty in the nine-dot problem. *Psychol. Res.*, *78*(2), 266–275. doi: 10.1007/s00426-013-0494-8

Stephen, D. G., & Dixon, J. A. (2008). The Self-Organization of insight: Entropy and power laws in problem solving. *The Journal of Problem Solving*, *2*(1), 6. doi: 10.7771/1932-6246.1043

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023, June). Expertise increases planning depth in human gameplay. *Nature*, *618*(7967), 1000–1005. doi: 10.1038/s41586-023-06124-2