Using Puzzle Video Games to Study Cognitive Processes in Human Insight and Creative Problem-Solving

Vasanth Sarathy **Nicholas Rabb** **Daniel M Kasenberg** **Matthias Scheutz**
Department of Computer Science
Tufts University
177 College Ave, Medford, MA 02155 USA
{vasanth.sarathy, nicholas.rabb, daniel.kasenberg, matthias.scheutz}@tufts.edu

Abstract

Classical approaches to studying insight problem-solving typically use specialized problems (e.g., nine-dot problem, compound-remote associates task) as stimuli together with verbal reports from subjects during problem-solving to reveal their thought processes, possibly adding other task-related metrics such as completion rate and physiological measures like eye fixation and neural activity. This approach has led to the claims that insight and creative thought require impasse and mental restructuring. What is missing from this literature is a cognitive *process* model of insight, and one reason for the lack of such a model is the lack of a unified, scalable, and tunable experimental framework with which to study human creative problem-solving with higher fidelity. In this paper, we introduce ESCAPE, an experimental paradigm using puzzle video games as stimuli which allow for the collection of process data that can serve as a basis for computational models. We have specifically developed a set of puzzle games based on this paradigm and conducted experiments that demonstrate the utility of the approach by revealing a set of computational principles that need to be accounted for by a theory of creative problems and the computational models based on it.

Keywords: insight problem solving; discovery

Introduction and Motivation

An open problem in creative problem-solving research is the understanding of the underlying cognitive processes and building "process models" of various intelligence-related capabilities – process in the sense of Marr's algorithmic level (Marr, 1982). For instance, in insight problem solving[1], we know that humans can and do restructure their problem representations to overcome constraints that may have been self-imposed (Kounios & Beeman, 2014). This led to useful conceptualizations of the process of insight from initial problem representations to reaching an impasse and then to restructuring the problem representations (MacGregor, Ormerod, & Chronicle, 2001; Oellinger, Jones, & Knoblich, 2014; Langley & Jones, 1988; Hélie & Sun, 2010). However, despite nearly a century of insight research (Maier, 1931), there has been little to no illumination of the underlying processes and computational aspects of how we restructure a problem representation, or for that matter what cognitive processes are involved in managing and overcoming an impasse. In other words, we do not know *how*, algorithmically, we humans solve insight problems, and when we do not solve them, what our strategies and approaches entail – i.e., how we conceptualize and approach a problem.

Figure 1: Screenshot from the puzzle platformer "Braid"[3]. We propose leveraging puzzle games like this as a basis for understanding the cognitive *processes* underlying creative problem-solving. Here, the user must restructure the notion of time to discover and use game mechanics. To avoid prior knowledge biases associated with visual cues in Braid, we have developed our own abstract puzzle game.

A process model is critical because it helps us develop a deeper understanding of the computational properties of problem-solving, which in turn will allow us to stimulate this behavior in humans and also build smarter AI. Current approaches to AI – particularly those involving large language models – have shown tremendous promise in a number of applications and even displayed emergent reasoning and problem-solving capabilities (Yao et al., 2023; Xie et al., 2023; Besta et al., 2023; Tian et al., 2023; Naeini, Saqur, Saeidi, Giorgi, & Taati, 2023). They are however incapable of lifting inductive biases present in their training data, something which humans as well as non-human animals are readily able to do. Changing our perspective, restructuring a representation, lifting constraints that we thought were relevant, etc., are all different capabilities required for creative problem-solving that modern-day AI cannot do because of the significant weight of pretraining (Steed, Panda, Kobren, & Wick, 2022). While training with large amounts of data

[1]Insight Problem Solving is a distinct family of problem-solving that overlaps with creativity. Here, we will use insight and creative problem-solving somewhat interchangeably, recognizing and respecting that they are indeed different. We believe the contributions in this work will benefit both insight and creative problem-solving researchers more generally.

[3]Courtesy: https://store.steampowered.com/app/26800/Braid/

endows LLMs with certain emergent capabilities, it also fortifies underlying biases. To get past these challenges, we will need new computational models (potentially inspired by human problem-solving) to unlock these additional capabilities. Current research in insight and creative problem solving is limited in many different respects: specific, bespoke tasks (Tulver, Kaup, Laukkonen, & Aru, 2023), limited scalability, they track performance and not actions, to name a few. Thus, understanding human problem solving requires a *restructuring* of how we conduct creative/insight problem research itself. In this paper, we propose a new paradigm for conducting human-subject research in creative problem-solving using an Experimental Setup for Capturing Problem-solving Experience (ESCAPE).

ESCAPE uses a sequence of puzzle video games[4] as task stimuli of the sort shown in Figure 2. We propose that these computerized puzzles limit (and standardize) the space of human actions, thereby enabling scaling and comparison across a broad range of populations, allowing researchers to equalize across the subject pool. They enable precise temporal mapping of actions, their preconditions and effects, world states and goals – all aspects of the problem representation. They allow researchers to calibrate across different subject expertise by systematically escalating the difficulty of puzzles, thereby enabling finding the sweet spot of puzzle difficulty. This allows the researcher to move past the prior knowledge of individual subjects and ensure that the experiment is at the appropriate level of difficulty for the subject.

In the rest of the paper, we will provide an overview of the framework, with a specific set of puzzle video games we designed together with some preliminary results showing the promise offered by the ESCAPE paradigm.

Background and Related Work

The phenomena of insight is often associated with the "Aha!" experience and represents a clear and somewhat sudden understanding of how to solve the problem (Kounios & Beeman, 2014). The emergence of insight has been associated with not only creative processes (e.g., incubation) but also the process of restructuring or representational change (Ohlsson, 1984, 1992), where an initial problem representation will need to be revisited and transformed to arrive at a solution.

Insight problem solving has been traditionally studied with specific, bespoke tasks called insight problems which are designed to trigger insight. A few examples of these problems include the 9-dot problem, 8-coin problem, matchstick problems and more recently compound remote associates (CRA) task (Tulver et al., 2023; Danek, Wiley, & Öllinger, 2016; Oellinger, Fedor, Brodt, & Szathmary, 2017). Typically a combination of verbal protocol and post-problem questionnaires are used to assess whether or not the subject solved the problem with insight. Task completion rates are often used to calibrate and measure difficulty. Despite the success of this paradigm, it has several limitations: several tasks are more conceptual and do not allow for timed capture of behaviors leading to insight. Moreover, while all these tasks elicit insight, they are not functionally the same, and as such making it more difficult to glean *insights* about how we can capture the computational essence of the particular processes of restructuring and post-impasse behavior. There is also no clear way of scaling the generation of puzzles so that new variations can be systematically tested. There is no notion of complexity levels associated with these puzzles and so researchers are unable to calibrate puzzles for different individuals who come in with different expertise and prior knowledge and bias. Finally, besides the CRA task (limited to word problems), other classical insight tasks are not readily useable within experimental paradigms such as fMRI.

Video games have been used in cognitive science for several decades including crucial contributions in the 1990s with the use of Space Fortress Game (Donchin, 1995) to present-day use in studying the impact of prior knowledge biases in human problem solving (Dubey, Agrawal, Pathak, Griffiths, & Efros, 2018). Much of the cognitive science use of video games has focused on the transfer of training from video games to other aspects of cognition or perception, using video games as interventions to improve problem-solving, measuring the role of game expertise, using games to measure other aspects of intelligence, and of course studying the impact of video-game playing on cognitive function (Boot, 2015; Kachergis & Austerweil, 2023; Zhang, Shen, Luo, Su, & Wang, 2009). We have yet to find a substantial body of work on using video games as core stimulus in human subject experiments to understand aspects of problem-solving. We have yet to find any work on using puzzle video games, specifically, to elicit insight or study creative behavior, and comparing against machine models.

Historically, AI research has used video games as a standard by which to measure progress in the field. Video games serve as a useful testbed to develop AI algorithms because the domain is limited making it easy to focus on specific aspects of the intelligence capabilities needed in a game, and there are human experts against which these algorithms can be pitted to measure success. There has also been recent work in developing environments for evaluating AI (particularly RL) algorithms in puzzle games (Renz, 2015; Bamford, 2021). However, using these puzzle video game environments for studying *human* creative problem-solving behavior has not been studied.

ESCAPE Framework

To facilitate discussion around our proposed framework, we will ground our discussion around a specific game: a 2D version of an escape-room puzzle that we created.

Puzzle Types

Figure 2 shows a set of 12 puzzles we use in our experiments as an illustration of the ESCAPE paradigm. These puzzles

[4]Readers are encouraged to try solving the puzzles here: http://tinyurl.com/insight-puzzles. We do not capture any data.

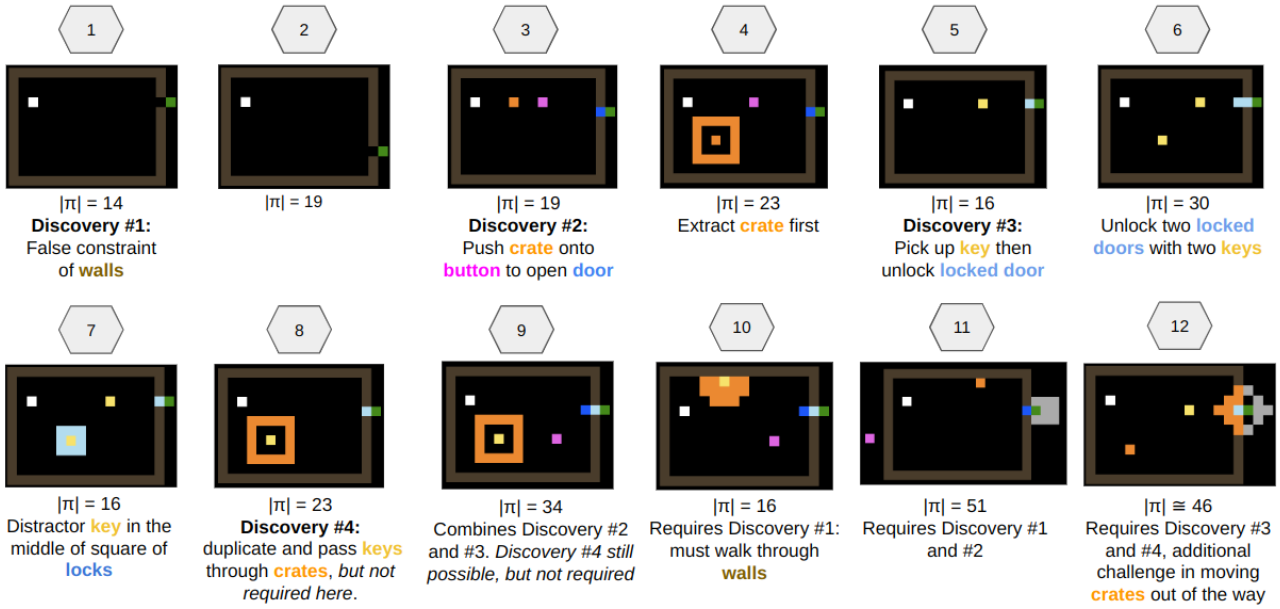

Figure 2: Twelve puzzle levels (left-to-right, top-to-bottom) of escalating difficulty together with an interpretation of the objects and their affordances that we, the developers, used when designing the levels. For each level we show the number of steps for the shortest plan ($|\pi|$). However, because the walls (brown squares) are a false constraint, until level 11, the solver can simply walk directly to the goal through the walls in about 14-16 steps. The numbers for $|\pi|$ shown in levels 1-9 are for solvers assuming the wall is real and therefore must make the required discoveries.

fall into six broad classes of puzzles, that capture the different aspects of the above-mentioned dimensions:

T1. **Simple:** The puzzles immediately induces the correct solution plan (E.g., Puzzles 1 and 2).

T2. **Tedious:** These puzzles typically do not readily evoke a solution plan. They consist of a large search space and require the solver to search the space of solutions, often backtracking (E.g., Puzzles 4, 8, 9, 13B). The classic push-box game, Sokoban, is another example of such puzzles.

T3. **Insight:** These puzzles induce a wrong initial problem representation (and undue self-imposed constraints), making the solution seem impossible at first. But once a correct representation is identified, the puzzle is immediately deemed solvable, at least at a high level, with a simple or tedious solution plan needing to be worked out. These puzzles require restructuring the mental representation of the problem (E.g., Puzzle 10 requires the wall to be (re)conceptualized as passable). Publicly available puzzle video games like Braid, Talos Principle, and Witness feature such insight puzzles.

T4. **Discovery:** These puzzles go beyond insight puzzles and require discovering something new about the world, which means the solver must explore their environment and find a hidden mechanic – a (novel) object, property, relation, affordance, or action effect (E.g., Puzzles 3, 5, 8 and 9). Video games like Braid, Infinifactory and Minecraft incorporate aspects of discovery.

T5. **Higher-Order Insights and Discoveries:** These puzzles combine elements of discovery, insight, and tedium, and often require the solver to compose together prior insights and discoveries (E.g., Puzzles 9, 11, 12).

T6. **Unsolvable:** These puzzles are designed so they cannot be solved. There are several reasons why we might need to study them: (1) it is not always practical (and typically undecidable) to know if a particular problem is unsolvable (Sarathy & Scheutz, 2018), (2) insight and discovery problems are often deemed solvable initially, followed by an impasse when they appear unsolvable, and when a solution is found, they appear trivial in hindsight (Sarathy, 2018). Without understanding this perception of (un)solvability, we cannot understand the cognitive processes underlying mechanisms like constraint formation, impasse detection, and restructuring (E.g., Puzzle 13B).

Dimensions of Problem-Solving Complexity

What makes a puzzle similar or different from another puzzle? What makes a puzzle difficult for a person? Here, we begin conceptualizing a set of dimensions or features that capture some crucial aspects of what makes certain puzzles more difficult than others. Generally speaking, puzzles or puzzle sequences that induce more constraints or require more restructuring are more difficult. As we designed these puzzles, we have begun developing a list of dimensions or potential metrics to analyze a puzzle's complexity.

D1. **Number of steps** in optimal or near-optimal solution: If the subject[5] were told the exact procedure to solve the problem, how many actions would they need to perform. This serves as a lower bound but glosses over important

[5]We use the term "subject" to broadly capture AI, human, and non-human animal problem-solvers.

questions of how the subject can arrive at the plan, how easy it is for the subject to execute the plan, and the like.

D2. **Number of first-order discoveries** that will need to be made through direct interaction with the environment. Not all objects or environmental states are perceivable by the subject initially. The subject might need to perform a specific series of actions to uncover a required object, property, or affordance. The subject might need to use one object in conjunction with another to acquire a new ability.

D3. **Number of higher-order discoveries** that build on prior discoveries and their distance from perceivable aspects of the environment.

D4. **Likelihood of prior knowledge constraining** problem representation. Certain problem elements and objects might come burdened with semantic significance, their visual appearance may trigger certain associations that induce problematic problem representations.

D5. **Size and fidelity of the action repertoire** available to the subject. A large number of action possibilities will decrease the likelihood of finding the "right action".

D6. **Number of solution paths**. Similar to the above dimension, this too relates to how difficult it will be for the subject to find the correct solution.

D7. **Number of new objects** that need to be constructed from available resources, and the number of such objects that, in theory, can be constructed.

The ESCAPE framework together with the set of puzzles allows us to explore human performance on different puzzle types and begin to unpack how our cognitive processes are influenced by the puzzle dimensions.

Empirical Case Studies

We presented 13 puzzles (12 from Figure 2 and 1 from Figure 3) to 50 subjects in an online study run via Amazon MTurk. The puzzles were designed by us using Puzzlescript, an online puzzle-making tool[6]. Of the 50 subjects 28 identified as male, 21 as female and 1 as other. There were 42 Caucasian/White, 4 Black or African American and 4 Asian subjects. 49 subjects identified as not having any colorblindness and 1 subject was with deuteranomaly. The average age of subjects was 36.8 (std dev. 12.3), ranging from 21-74 years.

Method and Data

Following an introduction and consent page, each subject was presented with a series of 13 puzzles, one at a time.[7] Every puzzle had an instruction "go to the green square" along with a description of available actions: arrow keys (up, down, left, right), "Z" to undo the last move (with unlimited undo's), and "R" key for resetting the puzzle to the start state. Each subject was allowed five minutes to complete the task. They were also shown a countdown timer. If they either completed the task (i.e., reached the green square with their white square) or were timed out, they would be shown the next puzzle and the timer would restart. After twelve puzzles, the subjects were divided into two conditions A and B. In condition A, they were given a difficult but solvable puzzle and in condition B, they were given an impossible puzzle – one that is by design unsolvable, even if it might not seem that way at first. After attempting each of the 13 puzzles, the subjects were given a questionnaire asking about how they felt while solving the puzzles, how difficult certain puzzles were, and whether they could solve additional puzzles. In this paper, we will not be able to discuss all aspects of the experiment or its results. Instead, we will focus on a few interesting phenomena to highlight the potential value of the proposed experimental paradigm. We have made all the (anonymized) data publicly available together with the code to facilitate reproducibility and transparency.[8]

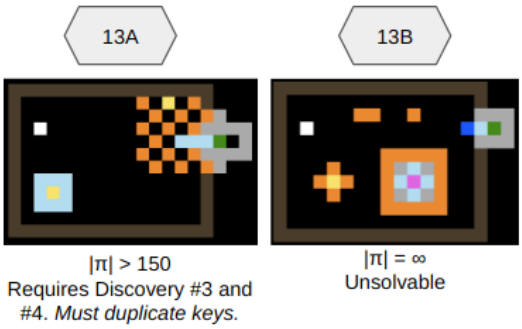

Figure 3: Thirteenth puzzle for each of the two conditions.

Results (Summary Statistics)

As seen in Table 1, the completion rate for the puzzles drops significantly for levels 10, 11, and 12. Up until puzzle level 9, very few discoveries are combined and even those made are not required. Interestingly, for these levels, there are significantly more "undos" or physical backtracking by the subjects. We also observe large increases in elapsed time in solving these more difficult puzzles. Finally, it is worth noting that the idle time is also increased in these puzzles, which represents the time that the user is not pressing any keys (i.e., either thinking and engaging mental processes or having walked away from the puzzle). We also observed strong performance jumps (higher completion rates, lower elapsed time, undos and restarts) in puzzles 5, 6, 7 and 8. Interestingly, these were where discoveries needed to be made, suggesting that either the subjects successfully made these discoveries or they bypassed making these discoveries by walking through the fake wall (brown) that might have been discovered in prior levels. To better understand the types of hypotheses that can be raised and tested in this framework, we will next take a closer look at two subjects.

[6]https://www.puzzlescript.net/

[7]Participants were not explicitly told that there would be unsolvable puzzles. But that puzzles will vary in difficulty.

[8]https://github.com/vasanthsarathy/puzzle-games-as-insight-problems.

puzzle	completion rate	no. of restarts	no. of undos	elapsed time (s)	idle time (s)
1	0.98/ 0.14	0.36/ 2.01	0.24/ 1.04	12.41/ 39.29	6.6968/ 7.57
2	1.0/ 0.0	0.56/ 3.96	0.68/ 4.81	5.97/ 4.52	4.8192/ 4.4
3	0.94/ 0.24	0.36/ 0.8	2.34/ 15.4	50.98/ 67.96	31.885/ 25.94
4	0.94/ 0.24	0.06/ 0.42	2.48/ 15.87	37.64/ 70.21	19.91/ 26.38
5	1.0/ 0.0	0.0/ 0.0	0.0/ 0.0	5.71/ 2.69	4.1954/ 2.2
6	1.0/ 0.0	0.0/ 0.0	0.0/ 0.0	10.39/ 6.76	7.6452/ 6.26
7	1.0/ 0.0	0.04/ 0.2	0.0/ 0.0	8.94/ 7.19	6.902/ 6.83
8	1.0/ 0.0	0.12/ 0.48	0.44/ 1.77	15.62/ 20.23	13.4598/ 19.89
9	0.98/ 0.14	0.2/ 0.93	1.56/ 9.94	34.64/ 60.61	23.8296/ 41.65
10	0.86/ 0.35	0.56/ 2.68	15.84/ 78.45	67.47/ 104.55	29.9206/ 53.84
11	0.64/ 0.48	1.8/ 2.95	13.16/ 52.92	163.15/ 117.44	64.705/ 68.57
12	0.24/ 0.43	5.04/ 4.86	33.5/ 102.99	268.73/ 52.77	86.4944/ 79.67

Table 1: Average/standard deviation for various metrics for all 50 subjects across both conditions.

Qualitative Analysis - The Tale of Two Subjects

Subject 62 and 31 were selected for further analysis. Both are Caucasian females without any colorblindness impairments. Subject 62 (46 y.o) considers themselves to be more experienced in puzzle game solving than Subject 31 (29 y.o).

Different Approaches to Problem-Solving

We visually analyzed the video playback of the performance of these two subjects for all puzzle levels[9]. Subject 62 proceeded systematically through each level, making each of discoveries 2 and 3. Subject 62 did not explore the possibility of the brown walls being fake until Puzzle 10. They also did not appear to make discovery 4. They did not solve Puzzle 10 and Puzzle 13A within the allotted time of 300 seconds. However, they solved all other puzzles, including 11 and 12. In contrast, Subject 31 made discovery 1 (fake walls) early in Puzzle 2, and used it in subsequent Puzzles to walk through walls. As a result, they did not make discoveries 2, 3, or 4 through Puzzle 10. In Puzzle 11, they eventually made discovery 2, allowing them to solve the level. However, they failed to solve Puzzles 12 and 13A.

Some Qualitative Takeaways

Two-sides of Fake Walls: Solving Puzzle 10 requires walking through the fake wall directly to the goal and ignoring all other objects. Our results (see Table 2) show that not only did Subject 62 not solve Puzzle 10, but they also had significantly more restarts and spent more time idling in this puzzle than Subject 31. Subject 62 was likely constrained by not having made discovery 1. Subject 62 was likely also distracted by items and objects that they previously used, believing those to be necessary here. This is supported by post-trial survey responses in which they said they were attempting to *"extract the yellow block."*. Even when they lifted the fake wall constraint, they were constrained by the other objects.

Subject 31 on the other hand faced none of these challenges as they directly walked through the fake wall and reached the goal, suggesting that the fake wall was not a constraint. That said, it seems that at one point in time, Subject 31 did have to restructure their mental representation – they state in the post-trial survey that they *"figured out that you can leave the bounds of the brown box,"* suggesting that they initially conceived of it as a box or wall. Our proposed ESCAPE framework enables deeper dives into such behaviors with the potential for novel experiments to explore different theoretical frameworks of representation restructuring. Do all subjects who break the wall constraint go on to use it as Subject 31 did? What meaning do we impart on visual cues (e.g., if the brown was replaced with blue, would we build a water metaphor)? How do stories and metaphors help us make sense of our perceptions and what biases do those impart?

[9]We have built a tool to playback the performance of any subject, which we will also make available in our github repo

		Subject 62	Subject 31
Puzzle 10	**time (s)**	300	4.9
	restarts	6	0
	idle (s)	215.9	1.9
Puzzle 11	**time (s)**	37.3	168.7
	restarts	0	8
	idle (s)	N/A	155.7
Puzzle 12	**time (s)**	278.8	300
	restarts	6	8
	idle (s)	N/A	121.2

Table 2: Comparative analysis of Subject 62 and Subject 31. Subject 31 discovered the fake wall constraint (discovery 1) and used it early allowing them to solve Puzzle 10 easily, but struggled with 11. Subject 62 on the other hand did not make this discovery until they attempted Puzzle 10, which made it difficult for them, however, they solved Puzzle 11 easily because they picked up other discoveries along the way.

Overloading Semantics: Solving Puzzle 11 required breaking the wall constraint (discovery 1) as well as utilizing the knowledge that the orange square when pushed on top of the pink square makes the blue square disappear (discovery 2). Our results show that for Puzzle 11, Subject 31 had more restarts and spent significantly more time idling in this level than Subject 62. This is likely because Subject 31 had not made discovery 2 previously and had to do so in this puzzle. By the end of puzzle-solving both subjects 62 and 31 had discovered the mechanic associated with discovery 2.

But, unexpectedly, in the post-trial survey, both Subjects

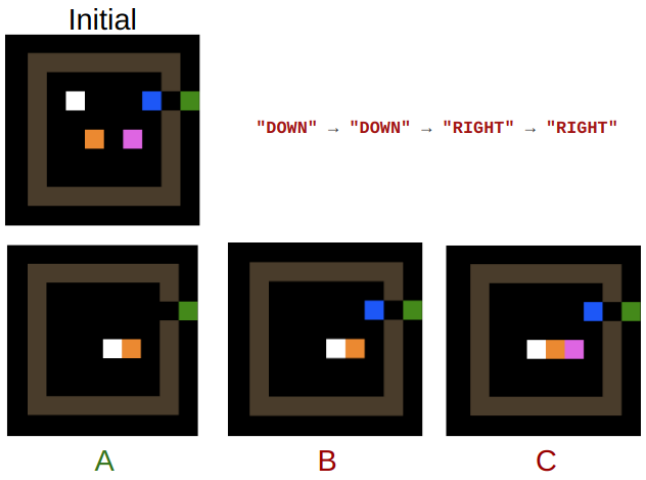

Figure 4: In a post-trial survey we showed subjects this initial state (top left) and a series of actions and asked them to let us know which of A, B or C is the correct final state. Here, "A" is the correct answer as it requires the use of discovery 2, namely pushing the orange square over the pink square makes the blue square disappear.

62 and 31 were unable to characterize the mechanic of discovery 2 when asked (see Figure 4 for the survey question). One possible reason for this is that the function of the "blue square" was overloaded. Depending on its history, one could have a blue square from walking over a yellow square (akin to picking up a key leading you to the state of holding a key) or one could have a blue square from pushing a blue square onto an orange square (turning it blue). Depending on this history, the blue square can be removed differently – in some instances by pushing an orange square over a pink square (like in the early puzzles) and in other instances by pushing the blue square against a light blue square (like in the later puzzles). The framework opens many questions of human problem-solving which go beyond imposing simple constraints to ones in which there are potentially interesting overlaps in semantics and history of objects and their use.

Solvability and Unsolvability: Puzzle 12 required both discoveries 3 and 4, which means subjects should have made an otherwise non-required discovery back in Puzzle 8, which neither did. Subject 62 solved this puzzle, but Subject 31 did not. There could be many reasons for this – for example, Subject 62 could have been a more experienced puzzle solver (from their self-assessment). Post-trial survey reveals that Subject 62 was not even sure if they solved it in time and even if they did *"it was just luck when [they] finally got it."* Subject 31 knew they were unable to solve it but expressed a possible solution idea indicating they thought it was solvable: *"I know I have to get a dark blue box using the yellow and that would get through the light blue somehow,"* suggesting while they had made discovery 4, they could not execute on it in time. Both subjects appear to have believed this level was possible. We also asked them in a post-trial survey how difficult they thought Puzzle 13B (the impossible puzzle from the other condition) would be on a scale of 0 (very easy) to 7 (impossible). Subject 62 believed the puzzle to be a 6, whereas Subject 31 thought it would be easier 3. A more thorough analysis would be needed to explain this difference – however, some possible reasons might include Subject 62 with their prior puzzle-solving experience and their potentially better understanding of our underlying mechanics was in a better position to evaluate 13B on its face.

Discussion and Conclusion

There is a wealth of useful hints about the necessary ingredients for a comprehensive theory of model human creative problem solving and computational models based on it buried in the above experimental data, much more than we could possibly address in a short conference publication. However, even the above analysis of the problem solving steps of two subjects already reveals important principles for computational process models. For example, they must allow for context-driven variability in exploratory vs. goal-oriented behavior and explain when restarts and undos and backtracking behaviors are initiated and how. They must account for idle time without any overt behaviors (presumably spent on different cognitive tasks such as analyzing impasses and planning the next actions) and for representational shifts caused by discoveries (both internally and with respect to overt behaviors), leading to different levels of understanding of the problem. And stepping back from the model, a theory of creative problem solving ought to be able to predict whether a given computational model will be able to solve a given puzzle based on the necessary discoveries and representational shifts it needs to effect, and thus also predict whether a solution might be impossible for a given model.

We believe that the proposed ESCAPE paradigm is particularly well-suited for develping process models in that a sequence of stimuli in multiple puzzles can be specifically designed to evoke different parts of the cognitive processes (constraint formation, impasse detection, restructuring, and search) by allowing precise control along various puzzle dimensions (action space size, solution size, number of discoveries, prior knowledge biases). The paradigm allows for a range of measures from behavioral (actions), to task-related (completion rate, timing), as well as neural and cognitive (eye gaze and fMRI). The framework allows us to hone in specifically on what aspect of a domain representation (objects, predicates, action preconditions, effects, goals etc.) are restructured by the human and used in problem-solving. We can do this by tracking the behaviors and inferring underlying formally described domain representations (MDP, PDDLs etc.) to infer state and goal descriptions that best explain the behaviors. This was not possible with classical creative and insight puzzles in the past. In sum, we believe ESCAPE provides a fertile ground with which to design a process model of insight and creative-problem solving in humans, and use the model to advance AI solvers.

References

Bamford, C. (2021). Griddly: A platform for ai research in games. *Software Impacts*, *8*, 100066.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., ... others (2023). Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Boot, W. R. (2015). Video games as tools to achieve insight into cognitive processes. *Frontiers in Psychology*, *6*.

Danek, A. H., Wiley, J., & Öllinger, M. (2016, February). Solving classical insight problems without aha! experience: 9 dot, 8 coin, and matchstick arithmetic problems. *The Journal of Problem Solving*, *9*(1). Retrieved from https://docs.lib.purdue.edu/jps/vol9/iss1/4 doi: 10.7771/1932-6246.1183

Donchin, E. (1995, June). Video games as research tools: The space fortress game. *Behavior Research Methods, Instruments, & Computers*, *27*(2), 217–223. doi: 10.3758/BF03204735

Dubey, R., Agrawal, P., Pathak, D., Griffiths, T. L., & Efros, A. A. (2018, July). Investigating human priors for playing video games. (arXiv:1802.10217). Retrieved from http://arxiv.org/abs/1802.10217 (arXiv:1802.10217 [cs])

Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: a unified theory and a connectionist model. *Psychological review*, *117*(3), 994.

Kachergis, G., & Austerweil, J. L. (2023). Video games as a path to a contextualized cognitive science, or how to move beyond 20 questions with nature. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). Retrieved from https://escholarship.org/uc/item/6695r26v

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*.

Langley, P., & Jones, R. (1988). A computational model of scientific insight. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (p. 177-201). New York, NY: Cambridge University Press.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: a process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Maier, N. R. (1931). Reasoning in humans. ii. the solution of a problem and its appearance in consciousness. *Journal of comparative Psychology*, *12*(2), 181.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* Henry Holt and Co., Inc.

Naeini, S., Saqur, R., Saeidi, M., Giorgi, J., & Taati, B. (2023). Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. *arXiv preprint arXiv:2306.11167*.

Oellinger, M., Fedor, A., Brodt, S., & Szathmary, E. (2017). Insight into the ten-penny problem: guiding search by constraints and maximization. *Psychological Research*, *81*(5), 925–938.

Oellinger, M., Jones, G., & Knoblich, G. (2014). The dynamics of search, impasse, and representational change provide a coherent explanation of difficulty in the nine-dot problem. *Psychological research*, *78*(2), 266–275.

Ohlsson, S. (1984). Restructuring revisited: Ii. an information processing theory of restructuring and insight. *Scandinavian journal of psychology*, *25*(2), 117–129.

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

Renz, J. (2015). Aibirds: The angry birds artificial intelligence competition. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 29).

Sarathy, V. (2018). Real world problem-solving. *Frontiers in human neuroscience*, *12*, 261.

Sarathy, V., & Scheutz, M. (2018). Macgyver problems: Ai challenges for testing resourcefulness and creativity. *Advances in Cognitive Systems*, *6*, 31–44.

Steed, R., Panda, S., Kobren, A., & Wick, M. (2022). Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 3524–3542).

Tian, Y., Ravichander, A., Qin, L., Bras, R. L., Marjieh, R., Peng, N., ... Brahman, F. (2023). Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Tulver, K., Kaup, K. K., Laukkonen, R., & Aru, J. (2023, April). Restructuring insight: An integrative review of insight in problem-solving, meditation, psychotherapy, delusions and psychedelics. *Consciousness and Cognition*, *110*, 103494. doi: 10.1016/j.concog.2023.103494

Xie, Y., Xie, T., Lin, M., Wei, W., Li, C., Kong, B., ... Li, Z. (2023). Olagpt: Empowering llms with human-like problem-solving abilities. *arXiv preprint arXiv:2305.16334*.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Zhang, X.-M., Shen, Z., Luo, X., Su, C., & Wang, J. (2009). Learning from video game: A study of video game play on problem-solving. In R. Huang, Q. Yang, J. Pei, J. Gama, X. Meng, & X. Li (Eds.), *Advanced data mining and applications* (p. 772–779). Berlin, Heidelberg: Springer.