Using Exploratory Agents to Evaluate Game Environments

Bobby Khaleque[1,*,†], Mike Cook[2,**,‡] and Jeremy Gow[3,***,§]

[1]Queen Mary University of London
[2]Kings College London
[3]Queen Mary University of London

Abstract

Designing engaging game environments, particularly for exploration, requires understanding diversity in exploratory behaviour. This study introduces a novel framework to model exploration via exploratory agents. This framework advances beyond prior work like PathOS+. Our agents employ multiple theoretically-grounded motivations and complex behaviour characterisation to distinguish between levels designed to encourage meaningful exploration and those that do not. Validated through human studies (40 participants, 14 level pairs), our coverage metric reliably reflects engagement differences, while combining coverage with novelty captures spatial-structural variations. This work establishes a richer foundation for AI-driven design tools that advance both exploration modelling and player engagement analysis.

Keywords

Game AI, Game Design, AI Agents, Exploration, Exploratory Behaviour

1. Introduction

Game designers create vast quantities of levels, yet conventional approaches fail to adequately model the diversity of spatial exploration. We introduce a framework modelling exploration with AI agents that advances beyond prior work such as PathOS+ [1] and Cook's work [2], providing designers with evaluative tools that identify levels that model more complex exploration. Our approach models exploration through a richer lens featuring multiple motivations and behaviour characterisation. Our approach models exploration through motivations inspired by level design theory from Totten [3].

We created 50 3D levels (25 "more engaging" and 25 "less engaging") using Unity to validate our agent framework. These levels operationalise Totten's principles. Our "more engaging" levels featured meaningful object distribution at vantage points, structured layouts, and clear paths. Meanwhile, our "less engaging" levels feature objects clustered haphazardly (such as along boundaries). These were explicitly designed to test whether our agents could detect the differences between these levels.

We validated perceptions of how engaging a level was through human evaluations (40 participants, 14 level pairs), confirming participants consistently distinguished types of level (83% agreement).

1.1. Exploratory Agents

Exploratory agents model the diversity of exploratory behaviour through theoretically-grounded motivations and advanced characterization. While our initial framework [4] simulated open-ended exploration, it lacked the richness needed for practical design applications. The current work makes advances beyond both our prior approach and systems like PathOS.

Our agents now operate with defined start/end points while maintaining rich exploratory diversity. We also integrate Totten's level design principles [3] to create multiple theoretically-grounded motivations. We replace simplistic context steering with our utility system to enable complex goal evaluation and action selection, which allows us to capture nuanced exploration beyond current methods.

This framework represents an improvement in modelling exploration. Taking exploratory agents from reactive simulators into potential diagnostic tools that quantify engagement through metrics like coverage, novelty, and inspection.

Joint AIIDE Workshop on Experimental Artificial Intelligence in Games and Intelligent Narrative Technologies, November 10-11, 2025, Edmonton, Canada.

2. Background

2.1. Modelling Exploration

Prior research has established valuable frameworks for understanding exploratory behaviour: Si's player study [5] made significant contributions by identifying distinct behavioural archetypes (wanderers, seers, pathers, targeters) through carefully designed experimental conditions. Their constrained tasks (e.g., 3-minute map surveys) provided important methodological rigor that revealed how task parameters shape exploration patterns.

Pathak's curiosity-driven agents [6] pioneered the use of intrinsic motivation as a computational driver for exploration, demonstrating how prediction-error minimisation can generate meaningful exploration in reward-sparse environments.

These foundational studies provide important insights into how exploration can be modelled in video games. Our work builds upon this existing work by expanding models of exploration to encompass a broader motivational spectrum as identified in psychological literature. Where prior research has examined specific exploration dimensions in isolation, our framework integrates these perspectives to model how inspective, diversive, and affective motivations interact within complex game environments. This allows us to preserve the methodological strengths of constrained experimental approaches while capturing the emergent richness of exploratory behaviour in more naturalistic settings.

2.2. Agent Frameworks for Design

Recent years have seen valuable innovations in AI-assisted design tools that model exploratory behaviour. PathOS [7] pioneered agent-based navigation prediction, demonstrating how simulated paths could reduce the burden of play testing. Its focus on accessibility and generalisability established important benchmarks for practical design tools. PathOS+ [1] significantly advanced this paradigm by introducing concrete data features (mass tagging, directional arrows) to reduce subjectivity in expert evaluations. This represents a commendable step toward bridging AI analysis and human design intuition. The ultimate goal is to reduce subjectivity in expert evaluations and improve the overall usability testing process. Also, Guerrero-Romero et al [8] present agents to assist with game design and testing. These include: Map Explorer, Novelty Explorer and Curious Agent. The map explorer focuses on spatial exploration by covering as much of the reachable areas of the game map as possible. It provides information such as the number of different positions visited, the total percentage of the map explored, and the game ticks required to complete the exploration. The novelty explorer emphasises exploring different game states rather than just physical positions. It aims to traverse as many unique game states as possible, which is related to the concept of novelty appraisal in intrinsically motivated agents. This approach helps in reducing uncertainty by connecting learning processes with count-based exploration.

The curious agent is designed to interact with game elements, prioritising those that have not been interacted with before. Providing data on the number of elements that are interacted with, actions triggered during interactions, and the game ticks required for these interactions. These agents collectively provide valuable insights into spatial exploration and interaction with objects, aiding game designers in evaluating and refining game mechanics and parameters. However, they are not designed to give a model of exploratory behaviour and assess how suitable a level might be for exploration.

These systems established important groundwork for operationalising exploration in design contexts. Our approach extends this foundation in modelling exploration as *motivationally influenced behaviour* rather than purely reactive wayfinding. While PathOS and PathOS+ effectively capture navigation patterns, and Guerrero-Romero's agents quantify atomic exploration aspects, our framework integrates Totten's architectural principles to simulate how exploratory motivations manifest in diverse behavioural signatures. This allows us to preserve the practical benefits of prior systems while adding deeper theoretical grounding in why players explore, enabling richer engagement predictions that complement existing tools.

In [2] Cook attempts to generate exploratory 3D spaces ("walking simulators") which reveals critical limitations in vision-only approaches, in that they failed to capture affective qualities, while content

generation suffered from semantic incoherence. This highlights a persistent gap: existing tools model exploration as *reactive wayfinding* rather than *motivationally influenced behaviour*.

2.3. Theoretical Foundation for Exploration

Totten's architectural principles [3] provide the missing theoretical basis for exploration modelling. His principles identifies distinct engagement drivers.

The concepts of narrative stages and meditative spaces are described by Totten as "reward spaces". These spaces are distinguishable from the rest of the level in terms of lighting, music or spatial characteristics. A narrative stage is a space in the environment which tells a story, an exposition through the game environment. They are spaces distanced from gameplay and have a strong narrative tie to the game. Meditative spaces are smaller, low-intensity moments of game spaces which help with game pacing.

Totten articulates that game environments can be organised in a linear, branching, or interconnected manner. For example, labyrinths afford a trajectory that can function as a linear odyssey. In this sense, paths can be used between locations in landscapes to guide players to interesting locations.

Totten further mentions the use of framing. Framing is when foreground elements are used to surround the view of something important. These highlight objects that would be worth investigating and potentially make the level more engaging for exploration. Refuges are "comfortably enclosed dark spaces". They may be perceived as comfortable as it would be accommodating to the abilities of the in game character (not too small or not too large). For example, a small section of trees in a large open space might be considered a refuge. Unlike prior computational approaches, Totten offers a unified theory linking spatial design to exploratory behaviour informing our motivation architecture.

Cook [2] independently recognises the importance of these elements (particularly framing devices and landmarks) but lacks the theoretical framework to operationalize them computationally. This theoretical gap explains their struggle to generate coherent exploratory spaces despite using advanced techniques.

Existing tools fundamentally lack theoretically-grounded motivational diversity, integrated behaviour characterisation across exploratory dimensions, and predictive models of engagement quality. Our work bridges these gaps by introducing a framework that synthesises Totten's principles with utility-based decision-making, enabling the first agent system capable of evaluating exploratory engagement through operationalised metrics.

3. Exploratory Agent Framework

We introduce an agent to explore these levels and derive metrics that might distinguish them as an extension of our previous work. Other works such as Pathak et al.'s [6] techniques are different from ours because our agent is not meant to be general in the sense that it would explore many different environments using intrinsic motivations. Our agent is given motivations and it will explore in different ways in different environments. Our framework uniquely models exploratory behaviour, attempts to evaluate how well a game level supports exploration, building directly upon our prior work in this domain. By leveraging exploratory agents tailored for this purpose, we address an unexplored niche level evaluation, providing insights for game designers aiming to enhance engagement through exploration. Our agent is utility-based and operates as follows:

- **Environment Scanning:** The agent periodically scans its surroundings within a defined field of view (FOV) and vision length (up to 230 Unity units of 350x350 level, covering 2/3rds of the map).
- **Goal Generation:** The agent considers various "modules" that evaluate nearby objects and areas.
- **Score Assignment:** Each identified goal is assigned a score (0 to 1), influenced by object significance, spatial relationships, and distance from the agent.

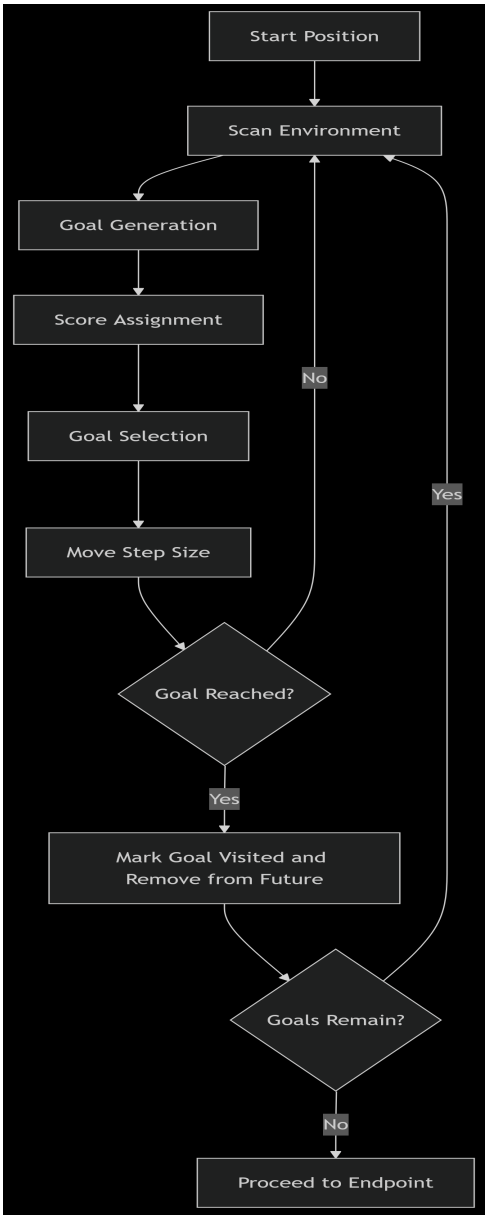


Figure 1: A flowchart showing how the agent framework functions

- **Goal Selection and Navigation:** The agent starts from a designated start point (which is a common practice in level design) and sequentially visits the highest-scoring goals. Once goals are visited (the agent comes within 10 units of the goals position), that object is discounted from future calculations. When no more goals remain, the agent proceeds to the designated endpoint. Conflicting scores are resolved through simply picking the closest goal, if the conflicting goals are the same distance, then a random one is picked between the two.

Several parameters in the system can be adjusted, including Field of View (defines the angular breadth of the agent's vision), Length of View (maximum distance for object detection (230 units in Unity)), Step Size (distance travelled before rescanning and recalculating goals) and Goals (maximum number of goals stored at any time). The figure below shows a visual representation of the framework.

3.1. Modules

Modules are used to form goals using the list of objects by applying certain criteria. Most modules, excluding Anticipation and Refuges, use a form of knowledge modelling via manually tagging entities

in Unity. Tags help identify entity types for scripting. For example, a tree might be tagged as “Tree” to identify tree types. This makes it easier to write scripts that interact with specific objects based on their assigned tag, without needing to manually reference each object individually. We have attempted to model the phenomena mentioned by Totten [3] in our agent implementation.

While our prototype requires manual Unity tagging for object identification, this is not fundamental to the framework. Future implementations could use; procedural tagging during PCG, computer vision approaches or semantic object recognition.

The modules used in this framework are:

- **Landmarks** - Detecting objects tagged as “Landmark”. Similar to how anonymous citation [4] models large object detection. Takes an object tagged as “Landmark” and compares it against the largest “Landmark” object it has seen so far. If no “Landmark” object has been seen so far, the largest “Landmark” object is set to the first “Landmark” object the agent has observed. A value of between 0 and 1 is returned, this value represents the percentage of how large the observed object is compared to the largest object observed. 1 is returned if the observed object is the largest one seen so far. Totten [3] mentions landmarks acting as eye candy, they should be able to be seen from multiple points of the level, as such we thought size was an appropriate scoring factor.
- **Narrative Stages** - Detecting objects tagged as “Narrative Stage”. Takes an object tagged as “Narrative Stage” and returns a score of between 0 and 1. This score is determined by how many other “Narrative Stage” objects are close to the observed one (within 10 units). The idea behind this is that the more of these objects there are in a certain vicinity the more exposition or environmental storytelling there could be from these objects. This module comes from the idea of “Narrative Stages” by Totten. These are expositions told through game environments, typically they are game spaces distanced from gameplay (e.g. a space that has no enemies) and have a strong narrative tie to the game.
- **Meditative Spaces** - Detecting objects tagged “Meditative Space”. Take an object tagged as “Meditative Space” and returns a score between 1 and 0. The score is determined by the lack of objects (within 50 units) around the meditative space, these other objects do not have to be tagged as “Meditative Space”. If a meditative space has 0 objects around it, the score is 1, for each object found around the space a penalty of -0.1 is given to the score. Totten described meditative spaces as smaller, more low intensity moments of game spaces. Having a lot of objects around these types of spaces would increase the intensity of the meditative space.
- **Framing** - Detecting objects between tagged “Framer” objects. Framing is using foreground elements to surround the view of something important as described by Totten. Takes an object tagged as “Framer” and checks if there is another framer to the right or to the left of the object, within camera screen space. Any object that is between the 2 framer objects is considered to be a framed object, no matter the tag. The score is determined by the angle of the framed object given the agent’s position. If the angle is 90 degrees, a score of 1 is given, the farther away from 90 degrees the framed objects angle is relative to the agent (to a maximum of 120 degrees or a minimal of 60 degrees) a penalty is applied of -0.33 per 10 degrees (so the minimum score is 0).
- **Paths** - Detecting tagged “Path” objects. Takes an object tagged as “Path” and assigns a score of 1. Paths can be used between locations in landscapes to guide players to interesting locations [3]. So, if the agent encounters a path, it assumes that there will always be an interesting location at the end of it, therefore the maximum score is given at all times when a path object(s) is encountered.
- **Anticipation** - Making an estimation of the area behind the object. This is essentially the same metric as “Anticipation Detection” used in anonymous citation [4]. An object, of any tag (excluding “Framer”, “Meditative Space”, “narrative stage” or “Landmark”), is taken and the penumbra of the object is calculated, given our agent is the light source. A minimum penumbra of 10000 is required, otherwise the score is 0, this is to make sure very small objects (such as blades of grass) are not investigated. If an object exceeds the maximum penumbra of 100000 then the max score of 1 is given.

- **Refuges** - Detecting enclosed spaces. Takes an object of any tag (excluding "Framer", "Meditative Space", "Narrative Stage" or "Landmark") and checks for any objects near it (within 10 units). If there are no objects within this distance a score of 0 is given. For every object found near the object being observed 0.1 is added to the score, until a max of 1 is reached. Refuges are described by Totten as enclosed dark spaces. We measure a refuge by how enclosed a space is by the amounts of objects surrounding it.

3.2. Baseline agents

A significant challenge in evaluating our exploratory agent framework is the lack of existing baselines specifically designed to model exploratory behaviour and assess how well game levels support exploration.

We constructed a **naive agent** and a **random agent** as baselines to represent simple alternatives to our framework. These baselines, while not directly comparable to our agent in complexity or purpose, serve to highlight the unique capabilities of our exploratory framework. The Naive Agent moves directly from start to end without any exploration. The random agent moves increments in a random direction (within -135 to +135 degrees), but with a bias towards the end point. This agent is different enough from both the naive and the exploratory agent to serve as a non-deterministic baseline. Furthermore, we engage in a comparison with PathOS+, but we thought it wouldn't serve as a good baseline. This is described section 4.2.1.

4. Designing Levels and Human Judgement Comparisons

We created 50 levels in a 3D game environment: 25 were designed to be "more engaging" and 25 were designed to be "less engaging" . We designed our more/less engaging levels according to theories of level design according to Totten [3] and our own judgement. Levels were 3D spaces with objects placed according to certain design principles. We designed these levels in the Unity game engine [1].

More engaging levels featured spatial arrangements where objects were distributed in a manner designed to encourage exploration. Objects were positioned at meaningful vantage points, key intersections, and visually distinct landmarks. The design relied on theories of level design for architecture, where structured layouts, navigable paths, and visually coherent object placements might support exploratory behaviour.

Less engaging versions were derived by taking the more engaging levels and putting the objects to one side, clustering them in less structured ways (e.g. aligning them along a riverbank or haphazardly bunching them together). These manipulated levels break the meaningful structure and reduce opportunities for interesting exploration. Intuitively, this reduces the level's cognitive affordances for discovery and reduces navigational complexity.

When objects, landmarks, and points of interest are distributed in a meaningful, balanced fashion, they create patterns that might support exploration and navigation. Conversely, when distribution is poor, such as objects clustered haphazardly against a boundary, or strewn without logical connections, this reduces the environment's legibility and disrupts intuitive way finding cues. Without clear, enticing focal points and structured levels that guide the observer's attention, we hypothesise that the environment becomes visually and cognitively less stimulating.

We then conducted an evaluation against human judgements to validate our assumption that the crafted "more engaging" levels are indeed perceived as such by humans.

4.1. Methodology

We selected 14 pairs of levels (side-by-side comparisons), each pair containing one more engaging and one less engaging variant. A total of 40 participants (all over 16 years old, who played 3D video games

[1] https://unity.com/

at least once a month) viewed top-down snapshots of these pairs and were asked to select which level in each pair seemed more engaging or if they were equally as engaging. We deliberately recruited gamers who played 3D video games at least once a month to ensure evaluators possessed knowledge to discern exploration centric design features. This study was conducted over 1 week with participants recruited through Reddit, X and email. We gained ethical approval from Anonymous institution with approval number (anonymised for review) to conduct this user study.

Top-down views remove potential distractions from camera angles or first-person perspectives, which might bias perception based on aesthetics or immersion. This perspective ensures that the focus remains on spatial arrangement and object distribution (specifically how a designer might view a level). We selected 14 pairs of levels to balance the statistical power needed for reliable conclusions with the cognitive load on the participants. This number ensures enough data points for meaningful analysis while minimising participant fatigue, which could negatively affect response quality. Drawing inspiration from Nielsen and Landauer's [9] principle that the probability of discovering usability issues decreases after the fifth user due to overlapping findings, we adapted this idea to level-pair evaluations. Testing an excessive number of pairs in one session might lead to diminishing returns, as participants could experience cognitive overload or reduced attention, potentially biasing their responses. Although this adaptation differs in context, the principle of balancing robustness with participant capacity remains relevant.

In the absence of directly comparable studies within this field, we can draw parallels from broader usability testing and game user research methodologies to justify our study design. For example, [10] discusses the determination of sample sizes in usability studies, emphasising the balance between statistical validity and practical constraints. Although our study involves 14 pairs of levels and 40 participants, exceeding these typical sample sizes, this approach enhances the robustness of our findings by providing a more comprehensive data set for analysis.

To mitigate potential biases from order effects, the presentation order of level pairs was randomised for each participant using Qualtrics [2] own randomisation method, ensuring no fixed sequence influenced the responses.

4.2. Questionnaire Results

To evaluate the reliability of our exploratory agent metrics, we employed multiple statistical methods. Inter-rater agreement was assessed using Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) and Intraclass Correlation (ICC), alongside raw agreement percentages to provide interpretable context. 83% of responses across all evaluator pairs correctly identified the intended "more engaging" level as "more engaging to explore.". While Cohen's Kappa adjusts for chance agreement, its utility was diminished here due to dataset bias. High prevalence of one category (e.g., our survey's intentional skew toward a preferred engagement level) and rater bias can skew marginal distributions, artificially depressing kappa values despite strong raw agreement. For example, near-unanimous agreement in skewed data can paradoxically yield low kappa scores, misrepresenting true consensus [11] address this, we prioritized PABAK (which corrects for prevalence and bias) [12] and ICC (suitable for continuous/non-uniform data) [13]. These yielded statistically significant results. PABAK showed moderate agreement with a value of 0.437 and a p-value of 0.005. ICC also showed moderate agreement with a value of 0.5 and a p-value of 0.00007. The preference for the intended "more engaging" levels, combined with statistically reliable agreement, validates our survey design and metric choices.

4.2.1. Comparison Against PathOS+

While other agent frameworks, such as those used in PathOS or curiosity-driven exploration, provide benchmarks for player navigation or potential motivations for exploration, they do not specifically attempt to model exploratory behaviour and evaluate how well a level might support exploration. However, PathOS+ does provide a model of exploration, via a curiosity metric. Nonetheless, we decided

[2] https://www.qualtrics.com/

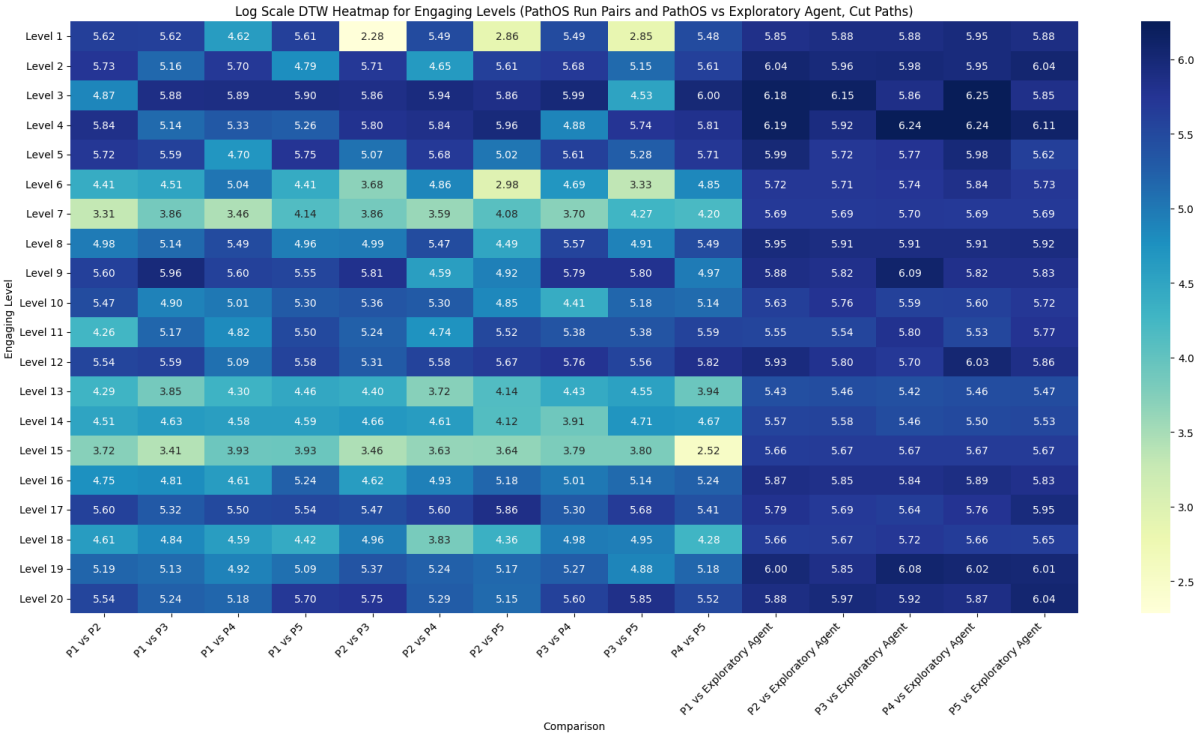


Figure 2: Log scale DTW heatmap of Path OS runs vs each other and the Exploratory Agent for the Engaging levels

to compare our exploratory agent to PathOS+. Specifically, the curiosity model, which Stahlke [14] describes in their thesis as 'represents a player's drive to explore the game world and see all it has to offer, uncovering secrets and traversing as much of the map as possible. Curious players are also interested in uncovering more information about a game's lore and narrative, and are always on the lookout for new content.'

PathOS also includes it's own markup system, similar to the tagging system we use in our agent framework, where individual objects can be "marked up" with several tags such as "Final goal", "Point of interest", "Mandatory Goal" and etc.

After appropriately marking up our levels we then ran PathOS plus on our engaging and less engaging levels. Due to the stochastic nature of PathOS's curisoity model, we decided to run it on each level 5 times and compare it to our exploratory agent's path trajectory using Dynamic Time Warping (DTW) [15] as well as using DTW to compare PathOS trajectories between runs. Figures 2 and 3 show these for each engaging and less engaging level.

PathOS+ runs were found to be significantly more consistent in the less engaging levels than in the engaging levels: Mean difference: 0.42 (4.92 - 4.50), where $p < 0.001$. PathOS+ vs exploratory agent divergence is significantly lower in less engaging levels: Mean difference: 0.16 (5.82 - 5.66), where $p < 0.001$.

A two-sampled t-test was used to calculate statistical significance. While its stochasticity produces variable paths in specific levels (e.g., DTW spread >3.5 in Level 19), this variability shows no correlation with exploratory alignment (all agent DTW >5.3). PathOS's design is optimised for QA testing, leading to undirected randomness that conflates coverage artifacts with intentional exploration. By contrast, our agent models goal-directed discovery through contextual environmental cues (e.g. landmark prioritisation, engagement triggers), demonstrating behaviourally plausible exploration that isn't apparent in PathOS plus. The inverse relationship between PathOS's internal variability and exploratory alignment (e.g., Level 8: highest variability [5.44] yet worst alignment [6.09]) suggests its fundamental mismatch with modelling exploratory behaviour.

Furthermore, PathOS+ exhibits high inter-run variability (DTW spread) in Levels 12, 17, and 20, yet

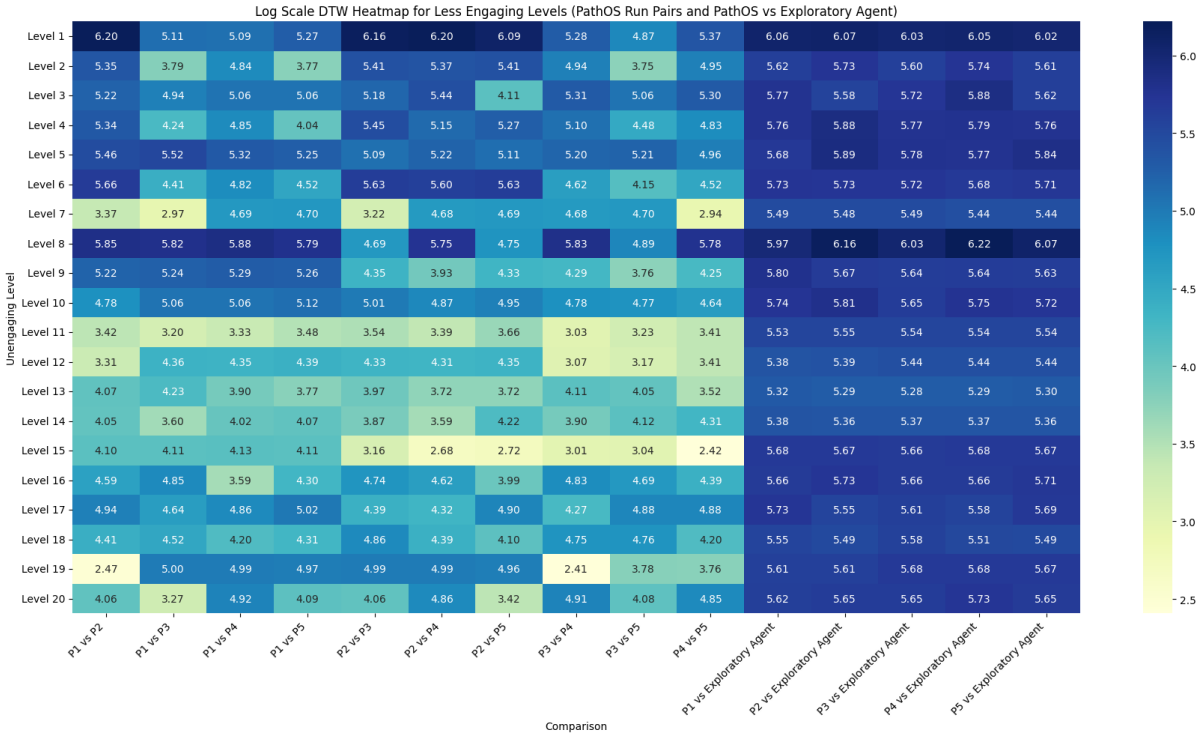


Figure 3: Log scale DTW heatmap of Path OS runs vs each other and the Exploratory Agent for the Less Engaging levels

this randomness reflects undirected coverage rather than intentional exploration. Its paths demonstrate repetition, going back and forth between objects in the level (Level 12), chaotic branching into empty spaces (Levels 17/20), and goal revisitation without purposeful discovery, this further suggests the hallmarks of QA stress-testing rather than modelling exploration. Conversely, our exploratory agent maintains goal-directed efficiency, avoiding non-informative areas while systematically prioritising objectives. This divergence is most pronounced in less engaging levels (1,4,8) where PathOS's randomness becomes extreme, yet persists even in engaging contexts (Levels 6,7,15) where PathOS exhibits unusual determinism. The consistent DTW >5.3 between PathOS and our agent further suggests its fundamental misalignment with exploratory principles: its coverage-maximisation exploration heuristic generates variability that inversely correlates with true exploration quality (e.g., Level 8's highest randomness [DTW 5.44] coincides with worst alignment [6.09]).

PathOS+ provides excellent navigation stress-testing, while our agent specializes in motivation-driven exploration. This complementary relationship is evidenced by, PathOS+'s superior coverage of edge cases and our agent's alignment with designed engagement cues.

5. Experiments

We ran experiments on a set of 40 levels (20 engaging, 20 less engaging), holding out 10 levels (5 engaging, 5 less engaging) as a test set. Giving an 80-20 train to test split.

5.1. Evaluation Metrics

We measured coverage (percentage of the map's area visited), inspection (the number of unique or special objects visited by the agent) and novelty (our own custom measure to reflect the diversity of uniqueness of observed objects).

Coverage was measured by splitting the level into 10x10 cells and measuring what percentage of the cells were visited by the agent(s). This is similar to the technique used by [4] although, our resolution

is higher.

Inspection consisted of taking all types of unique/special objects. E.g. Objects tagged as "narrative stage", "meditative space", as well as objects not tagged as anything unique but only counting them as one object (so if an agent visits one "tree" object it has effectively visited them all) and measuring if the agent had come within 10 units of these special objects. The inspection score is the percentage of these unique objects the agent has "visited".

Novelty Quantifies engagement based on object-type exposure, with penalties for repetition and peripheral visibility. N_t represents the novelty score at time t for a given type of object. S_t represent the total novelty score at time t. Δt represents the time interval, where $\Delta t = 1$ seconds. r represents the rate of novelty score recovery, where $r = 0.03$ per second. M represents the minimum novelty score an object type can recover to, where $M = 0.1$. P represents the penalty applied to the novelty score when an object type is seen, where $P = 0.1$. PS represents the peripheral score penalty applied to the novelty score when an object type is seen, where $PS = 0.75 * N_t$ $PS = 0.5 * N_t$ v_t represents the visibility flag at time t, where $v_t = 1$ if the object type is seen and $v_t = 0$ otherwise.

- **Initialisation:** At first encounter, $N_0 = M = 0.1$.
- **Update Rules:**
  - If unseen ($v_t = 0$) and not "new":

$$N_{t+1} = \min(N_t + r\Delta t, M)$$

  - If seen ($v_t = 1$) and "new":

$$N_{t+1} = N_t - P$$

  - If seen ($v_t = 1$) and not "new":

$$N_{t+1} = N_t + r\Delta t$$

- **Peripheral Penalty (PS):**

$$PS = \begin{cases} 0.75 & (\pm 30^\circ\text{–}45^\circ \text{ from camera centre}) \\ 0.5 & (>\pm 45^\circ) \end{cases}$$

  Applied as $N_t \leftarrow PS \times N_t$ when calculating S_t.

Worked Example **Timeline:** Agent encounters a "Grass" type.

- **t=0:** First seen (central view, PS=1): N_0=0.1, S_0=0.1.
- $t = 1$**:** Seen again (peripheral, $PS = 0.75$):

$$N_1 = N_0 - P = 0.1 - 0.1 = 0.0 \quad \text{(marked "not new")}$$
$$S_1 = 0.0 \times 0.75 = 0.0$$

- $t = 2$**:** Not seen: $N_2 = \min(0.0 + 0.03 \times 1, 0.1) = 0.03$.
- $t = 3$**:** Seen peripherally ($PS = 0.75$):

$$N_3 = N_2 + r\Delta t = 0.03 + 0.03 = 0.06 \quad \text{(not "new")}$$
$$S_3 = 0.06 \times 0.75 = 0.045$$

This novelty measure ensures that with the recovery rate r that novelty rebounds if objects are infrequently seen. Also, the peripheral discounting reflects attention bias to the centre that humans prioritise [16]. Also, visual novelty detection frameworks often weight central regions more heavily due to foveal resolution advantages [17]. We hypothesise that higher novelty scores indicate diverse, salient exploration.

5.2. Results

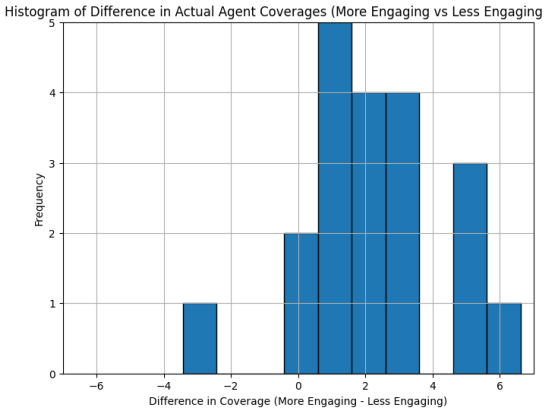
Our exploratory agent showed the greatest sensitivity to level differences in terms of coverage. Statistical tests (t-tests with Bonferroni correction) confirmed that coverage was significantly different between the more engaging and less engaging levels. Inspection and novelty did not show statistically significant differences, but their trends were still informative. The mean of Novelty alone fails to distinguish between sustained exploration (consistent exposure to new stimuli) and sporadic novelty spikes (infrequent, unpredictable discoveries). Statistical manipulations (mean, standard deviation and max) of novelty address this. The standard deviation quantifies the variability in novelty over time. A high standard deviation indicates erratic exploration (e.g. alternating between new and familiar regions), while low standard deviation suggests stable engagement. The negative skew in novelty standard deviation for the agent implies that "more engaging" levels elicit consistent novelty, avoiding monotony without chaotic spikes. Max novelty identifies peak "surprise" moments. High max novelty correlates with memorable, unexpected discoveries that might anchor player attention. While max novelty lacked significance post-Bonferroni, its inclusion guards against over-smoothing engagement signals (e.g., penalising levels with rare but impactful discoveries).

Coverage in particular shows a much more positive skew for differences, favouring the engaging levels, as shown in figure 4, compared to the naive and random agents which show a normal distribution. The same differences cannot be seen as strongly in inspection or standard deviation of novelty for all agents. However, the standard deviation of novelty for the exploratory agent does show a negative skew (though not as strong as the positive skew shown in coverage for the actual agent), indicating there is potential in using standard deviation of novelty as a differentiator for our more and less engaging levels. To assess the differences between the more engaging and less engaging levels, we employed t-tests to compare our evaluation metrics, coverage, inspection, and novelty (max, mean and standard deviation), across the two level categories for all three agents (our exploratory agent, random, and naive). The t-test is a statistical method for evaluating whether there are significant differences between two independent groups, which aligns with the experimental setup of contrasting more engaging and less engaging levels. This approach allows us to determine whether the performance metrics of our exploratory agent effectively capture the structural differences in the level design. The results provided valuable insight into how well our agent aligns with our hypothesis that certain metrics, particularly coverage, are sensitive to level design features.

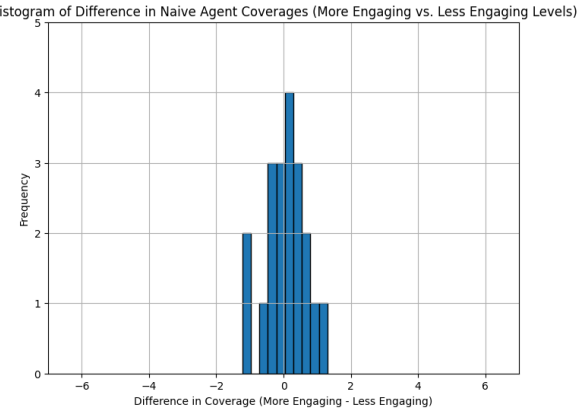
Given that multiple t-tests were conducted across different metrics and agent types, the Bonferroni correction was applied to mitigate the increased risk of Type I error. Without such adjustments, the likelihood of falsely rejecting the null hypothesis increases with the number of statistical comparisons. By dividing the significance threshold (commonly set at 0.05) by the number of tests, the Bonferroni correction ensures a stricter criterion for significance, enhancing the robustness of the conclusions. This precaution was especially important in validating our exploratory agent's superior performance compared to the random and naive baselines. Because we performed 15 t-tests the adjusted significance level was 0.003.

The poorer performance of the random and naive agents, as anticipated, underscores the strength of the exploratory agent framework. The naive agent's linear trajectory from start to end neglects any consideration of environmental features, resulting in minimal coverage and interaction with the level's elements. The random agent, though less deterministic, fails to prioritise meaningful goals, leading to inefficient and aimless exploration. In contrast, the exploratory agent incorporates a utility-based framework with defined goals and scoring systems, enabling it to navigate and interact with the environment in a structured and purposeful manner. This design ensures that the exploratory agent aligns more closely with the qualities that characterise engaging levels, particularly through its high coverage and nuanced sensitivity to level design features.

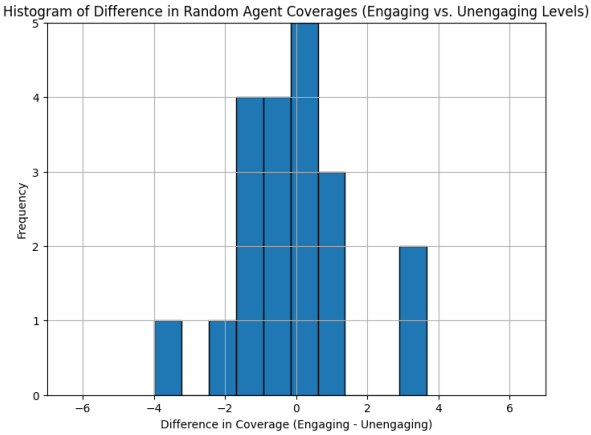
These results support the validity of the exploratory agent's design and its metrics, further emphasising its utility as a tool for evaluating levels.



(a) Histogram showing the difference in coverage between our more engaging and less engaging levels for our exploratory agent



(b) Histogram showing the difference in coverage between our more engaging and less engaging levels for our naive agent



(c) Histogram showing the difference in coverage between our more engaging and less engaging levels for our random agent

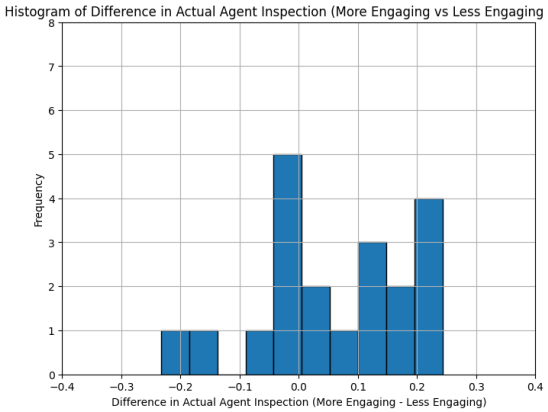
Figure 4: Histograms showing coverage differences for all agents

6. Discussion and Future Work

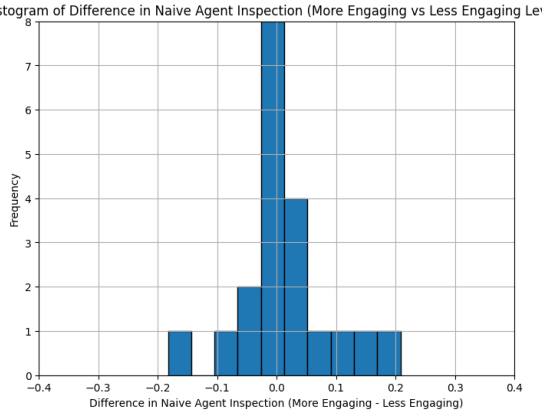
Our study establishes a framework that advances exploratory agent design beyond prior systems like PathOS. While coverage emerged as a robust indicator of engagement, the synergistic power of combined metrics, particularly coverage and novelty, demonstrates how quality modelling captures nuances in exploratory behaviour that single metric approaches miss. This represents a paradigm shift, where prior systems like PathOS+ lacked nuance in modelling exploratory behaviour, our framework evaluates how spatial-structural variations inspire exploration. The ability of coverage to differentiate between our "less engaging" and "more engaging" levels provides compelling evidence that agent-derived metrics can predict human engagement perceptions. This addresses a core limitation noted by designers using PathOS+.

The weaker performance of inspection and novelty metrics opens an avenue for future work, our current object-centric evaluation overlooks how narrative elements (e.g. Totten's "meditative spaces") might inspire affective exploration. This suggests that our motivational architecture requires further diversification to fully capture experiential quality dimensions.

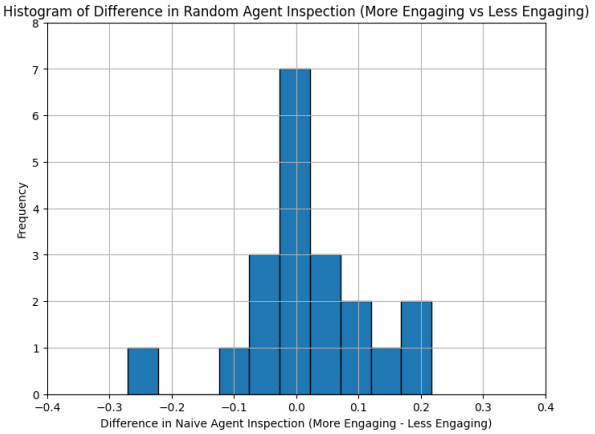
Future work could involve integrating Totten's narrative staging and refuge concepts to model affective exploration, creating agents that respond to environmental storytelling. We also plan to implement our agent framework in procedural content generation (PCG) pipelines to assist in the



(a) Histogram showing the difference in inspection between our more engaging and less engaging levels for our exploratory agent



(b) Histogram showing the difference in inspection between our more engaging and less engaging levels for our naive agent



(c) Histogram showing the difference in inspection between our more engaging and less engaging levels for our random agent

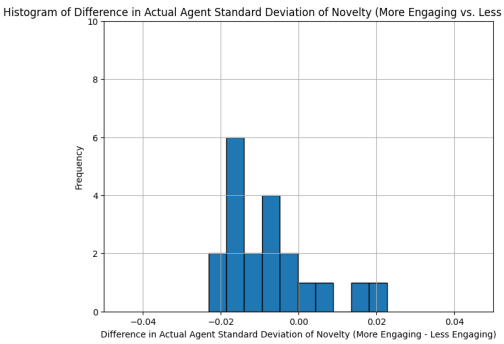
Figure 5: Histograms showing inspection differences for all agents

creation of levels that support exploration. Incorporating additional motivational drivers to capture the full spectrum of exploratory behaviour identified in the psychological literature is also another area of investigation.

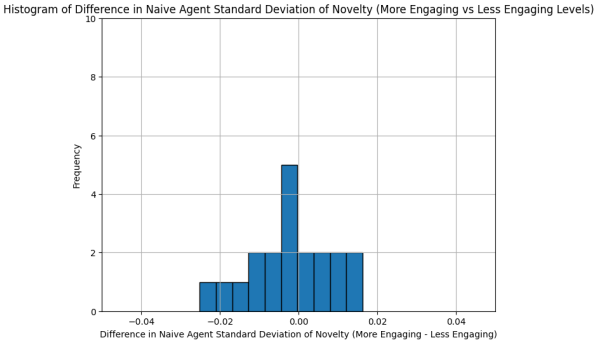
By addressing these directions, our framework could evolve into a tool for game designers, providing theoretically-grounded insights throughout the design process while advancing computational models of player experience. The foundation established here demonstrates significant potential for bridging architectural theory with AI-driven game design tools.

7. Conclusion

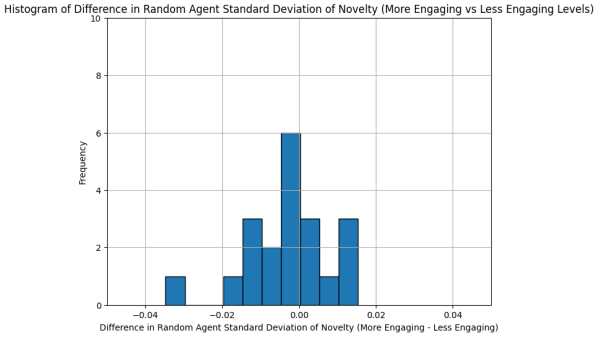
We demonstrated that a utility-based exploratory agent can differentiate the levels deemed more engaging from those deemed less engaging. By capturing differences in spatial layouts and object distributions, the agent's coverage metric, and to some extent combined metrics, can serve as proxies of engagement. This study lays the foundation for integrating exploratory agents into the design process, where they can serve as automated evaluators in designer workflows, filtering out less engaging content and driving the iterative improvement of game levels.



(a) Histogram showing the difference in novelty standard deviation between our more engaging and less engaging levels for our exploratory agent



(b) Histogram showing the difference in novelty standard deviation between our more engaging and less engaging levels for our naive agent



(c) Histogram showing the difference in novelty standard deviation between our more engaging and less engaging levels for our random agent

Figure 6: Histograms showing novelty standard deviation differences for all agents

Acknowledgments

This work was supported by the EPSRC Centre for Doctoral Training in Intelligent Games & Games Intelligence (iGGi) EP/S022325/1. We thank the reviewers for their constructive feedback.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. N. Nova, S. C. F. Sansalone, P. Mirza-Babaei, Pathos+: A new realm in expert evaluation, in: Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play, CHI PLAY '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 122–127. URL: https://doi.org/10.1145/3450337.3483495. doi:10.1145/3450337.3483495.

[2] M. Cook, The road less travelled: Trying and failing to generate walking simulators, CoRR abs/2104.10789 (2021). URL: https://arxiv.org/abs/2104.10789. arXiv:2104.10789.

[3] C. Totten, An Architectural Approach to Level Design, CRC Press, usa, 2018. URL: https://books.google.co.uk/books?id=WH06DwAAQBAJ.

[4] A. Author, Anonymous title, Anonymous, 2024.

[5] C. Si, Believable Exploration: Investigating Human Exploration Behavior to Inform the Design of Believable Agents in Video Games, Ph.D. thesis, UNIVERSITY OF TECHNOLOGY SYDNEY, 2017.

[6] D. Pathak, P. Agrawal, A. A. Efros, T. Darrell, Curiosity-driven exploration by self-supervised prediction, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, Sydney, Australia, 2017, pp. 2778–2787. URL: https://proceedings.mlr.press/v70/pathak17a.html.

[7] S. Stahlke, A. Nova, P. Mirza-Babaei, Artificial players in the design process: Developing an automated testing tool for game level and world design, in: Proceedings of the Annual Symposium on Computer-Human Interaction in Play, CHI PLAY '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 267–280. URL: https://doi.org/10.1145/3410404.3414249. doi:10.1145/3410404.3414249.

[8] C. Guerrero-Romero, S. M. Lucas, D. Perez-Liebana, Using a team of general ai algorithms to assist game design and testing, in: 2018 IEEE Conference on Computational Intelligence and Games (CIG), IEEE, The Netherlands, 2018, pp. 1–8. doi:10.1109/CIG.2018.8490417.

[9] J. Nielsen, T. K. Landauer, A mathematical model of the finding of usability problems, in: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, CHI '93, Association for Computing Machinery, New York, NY, USA, 1993, p. 206–213. URL: https://doi.org/10.1145/169059.169166. doi:10.1145/169059.169166.

[10] R. Macefield, How to specify the participant group size for usability studies: a practitioner's guide, J. Usability Studies 5 (2009) 34–45.

[11] A. R. Feinstein, D. V. Cicchetti, High agreement but low kappa: I. the problems of two paradoxes, Journal of Clinical Epidemiology 43 (1990) 543–549. URL: https://www.sciencedirect.com/science/article/pii/089543569090158L. doi:https://doi.org/10.1016/0895-4356(90)90158-L.

[12] T. Byrt, J. Bishop, J. B. Carlin, Bias, prevalence and kappa, Journal of Clinical Epidemiology 46 (1993) 423–429. URL: https://www.sciencedirect.com/science/article/pii/089543569390018V. doi:https://doi.org/10.1016/0895-4356(93)90018-V.

[13] K. Hallgren, Computing inter-rater reliability for observational data: An overview and tutorial, Tutorials in Quantitative Methods for Psychology 8 (2012) 23–34. doi:10.20982/tqmp.08.1.p023.

[14] S. N. Stahlke, Synthesizing play: exploring the use of artificial intelligence to evaluate game user experience, 2020. URL: https://ontariotechu.scholaris.ca/items/c5dc89c9-9d21-4140-bfbd-0d5ebb4b3472/full.

[15] M. Müller, Dynamic Time Warping, in: Information Retrieval for Music and Motion, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 69–84. URL: https://doi.org/10.1007/978-3-540-74048-3_4. doi:10.1007/978-3-540-74048-3_4.

[16] B. W. Tatler, The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, Journal of Vision 7 (2007) 1–17. doi:10.1167/7.14.4.

[17] L. Itti, C. Koch, Computational modelling of visual attention, Nature Reviews Neuroscience 2 (2001) 194–203. doi:10.1038/35058500.