Navigates Like Me: Understanding How People Evaluate Human-Like AI in Video Games

Stephanie Milani
smilani@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Arthur Juliani
Microsoft Research
New York, New York, USA

Ida Momennejad
Microsoft Research
New York, New York, USA

Raluca Georgescu
Microsoft Research
Cambridge, United Kingdom

Jaroslaw Rzpecki
Monumo
Cambridge, United Kingdom

Alison Shaw
Ninja Theory
Cambridge, United Kingdom

Gavin Costello
Ninja Theory
Cambridge, United Kingdom

Fei Fang
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Sam Devlin
Microsoft Research
Cambridge, United Kingdom

Katja Hofmann
Microsoft Research
Cambridge, United Kingdom

ABSTRACT

We aim to understand how people assess human likeness in navigation produced by people and artificially intelligent (AI) agents in a video game. To this end, we propose a novel AI agent with the goal of generating more human-like behavior. We collect hundreds of crowd-sourced assessments comparing the human-likeness of navigation behavior generated by our agent and baseline AI agents with human-generated behavior. Our proposed agent passes a Turing Test, while the baseline agents do not. By passing a Turing Test, we mean that human judges could not quantitatively distinguish between videos of a person and an AI agent navigating. To understand what people believe constitutes human-like navigation, we extensively analyze the justifications of these assessments. This work provides insights into the characteristics that people consider human-like in the context of goal-directed video game navigation, which is a key step for further improving human interactions with AI agents.

CCS CONCEPTS

• **Applied computing** → **Computer games**; • **Computing methodologies** → *Reinforcement learning*; • **Human-centered computing** → Empirical studies in HCI.

KEYWORDS

human subject study, believable AI, games, navigation

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581348

ACM Reference Format:
Stephanie Milani, Arthur Juliani, Ida Momennejad, Raluca Georgescu, Jaroslaw Rzpecki, Alison Shaw, Gavin Costello, Fei Fang, Sam Devlin, and Katja Hofmann. 2023. Navigates Like Me: Understanding How People Evaluate Human-Like AI in Video Games. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581348

1 INTRODUCTION

Games are considered one of the oldest forms of human social interaction [49, 79]. Throughout history, people have played games as important cultural and social bonding events [45], as teaching and learning tools [15, 44], and for enjoyment [41]. Today, video games have emerged as a popular form of structured play, inviting players to immerse themselves in captivating virtual worlds. This immersion is vital to making these games enjoyable [11, 53].

To enhance this immersive experience, game designers focus on creating believable non-player characters (NPCs) that can interact with players in diverse ways. A crucial part of believability is *human-likeness* — that is, the ability of the NPC to behave as a person would. Because many player-NPC interactions are critical to the game, it is important that the NPCs behave believably to maintain immersion [14, 35, 85]. Indeed, video game players find playing against more human-like agents more enjoyable [72]. Traditionally, game developers have designed NPCs to follow a predetermined set of actions. However, this approach can be both time-consuming and challenging, which has motivated designers to turn to artificial intelligence (AI) for assistance with NPC design. Due to this shift, there is a need for research into understanding what people perceive as human-like in AI agents.

At the same time, AI researchers have identified achieving complex human-like behavior as a critical milestone [8, 21, 71] towards developing agents that can flexibly collaborate with people in

shared human-AI environments [9, 25, 84] and various robotics applications [64]. This goal is not satisfied by agents demonstrating a high proficiency level at the assigned task. For example, AI-powered vehicles must behave sufficiently human-like for human drivers to interpret, anticipate, and act in their presence [27]. As a result, understanding the behaviors that contribute to people's perceptions of human likeness is a foundational first step towards achieving general human-like behavior of artificial agents.

In this work, we contribute to the objective of developing human-like agents by identifying and understanding what constitutes human-like behaviors in a video game. To scope our study, we focus on a 3D video game where agents must navigate from one point to another. This form of navigation is pervasive in many video games, making it a key area of interest for game developers [3, 18]: in embodied games, players must move from place to place to accomplish their goals or explore the world. More generally, it is considered fundamental to embodied biological intelligence [26, 59], making it of interest to cognitive scientists [56, 62] and researchers interested in intelligent behavior [87]. It has also been a key area of interest in HCI [80] due to how people (or robots) navigate in real, augmented, or entirely virtual spaces.

To study navigation in video games, we leverage the recently-proposed Human Navigation Turing Test (HNTT) [16], in which human judges indicate which of two videos demonstrates more human-like behavior. The judges then justify their decision and indicate their certainty about their choice. In that work, the authors compared the accuracy of the human-likeness assessments to random chance but did not instantiate a statistical test to definitively conclude whether an agent passed the HNTT. According to their assessment, both studied AI agents did not pass the HNTT. As a result, producing an agent that passes the HNTT is still an open challenge. To this end, we design a novel agent to pass the HNTT. To assist with our design of a human-like agent, we inspect the resulting behavior of the two baseline agents from prior work [16]. With these insights, we design our novel agent — the *reward-shaping* agent — using simple and intuitive techniques.

We then conduct a behavioral study on Amazon Mechanical Turk (MTurk) of the HNTT to investigate the behavior of our agent and the baselines. To determine whether agents pass the HNTT, we propose a firm criterion: a statistical test that determines whether human judges distinguish between human and agent behavior at a level that is *significantly different* from chance. We then validate the conclusion of previous work: the two baseline agents are not sufficiently human-like because they do not pass the HNTT. In contrast, human judges cannot reliably distinguish between the behavior of our *reward-shaping* agent from one controlled by a person. To our knowledge, this agent is the first to pass the HNTT.

To understand these assessments, we analyze the free-form responses to determine which characteristics people believe are representative of human and AI navigation behavior. We annotate the responses with codes that summarize the provided rationale. Using these annotations, we find that there are key differences between how people characterize human-like and non-human-like behavior. Specifically, we find that people utilize the same high-level characteristics when describing human-like and non-human-like behavior, but the presence or absence of these characteristics strongly informs their judgments.

Based on the findings of our analysis, we summarize considerations when developing and evaluating the human likeness of AI agents. For example, considering the end use of the agent is critical for defining what is meant by human like and designing a study accordingly. In summary, we make the following contributions.

(1) We contribute a novel *reward-shaping* agent that exhibits more human-like navigation behavior.
(2) We conduct a behavioral study to assess: a) whether people reliably distinguish the behavior produced by the AI agents from that generated by people and b) what characteristics people believe are indicative of human-like behavior.
(3) We conduct an extensive analysis of the resulting data. We propose a firm criterion to determine whether an agent passes the HNTT and find that only our *reward-shaping* agent passes the HNTT according to this metric. We analyze the free-form responses to determine the characteristics that people believe are representative of human-like behavior.
(4) Based on our findings, we propose concrete suggestions for developing and evaluating human-like AI.

2 RELATED WORK

Researchers have taken various approaches to address the challenge of developing believable AI agents in games, including learning from demonstrations [33, 38, 52], reinforcement learning [5, 20, 52, 89], and more [54, 77]. We focus on *reinforcement learning* [73] because it provides a generally-applicable set of algorithms for learning to control agents in settings including (but not limited to) modern game environments [3, 24, 43, 75, 81]. It also offers significant benefits as an approach for generating navigation behavior [3]. In particular, the use of reinforcement learning may enable more complex navigation abilities (such as grappling or teleportation) and alleviate game designers from the labor-intensive procedure of the most popular alternative method to produce this behavior [51].

In reinforcement learning, an agent learns to accomplish a task by maximizing a reward, or score, that tells the agent how well it is performing. Although agents learn effective navigation by maximizing this reward, they make no consideration for the *style* with which they act [3]. If these approaches are to be adopted in commercial game development, practitioners have firmly asserted that controlling style is essential [34]. As an extreme example, reinforcement learning approaches that have recently defeated world champion human players at modern games demonstrated unusual behaviors [30] that made collaborative play between human and AI in mixed teams far less successful [6]. Simply maximizing the task-specific reward signal is unlikely to produce human-like agents.

Reward shaping [58, 86] is a simple yet powerful technique that allows practitioners to clearly specify the desired agent behavior. This approach involves crafting a reward signal that provides dense feedback to the agent. It is an intuitive way for those without a machine learning background to control the agent's behavior by specifying objectives instead of dedicating time to optimizing unintuitive hyperparameters. Additionally, reward shaping can be used with any reinforcement learning algorithm, making it possible to swap in and out the underlying algorithm as needed. We utilize reward shaping to generate more human-like behavior.

Figure 1: Navigation task as observed by study participants (screenshot, left), and detail of the mini map of the game level (right). Agents spawn on the island outside of the main map, which is shown in the bottom portion of the mini map on the right. They must jump to the main area and navigate to the goal location. The light blue containers in the left screenshot represent the goal location.

There is no standard set of metrics for evaluating human-like AI. One paradigm involves measuring human similarity with proxy metrics for human judgments. Some work measures the task performance of the AI [77, 89], but this metric is an insufficient proxy for human similarity. Other work assesses how well the AI agent can predict the following human action [33] or align its behavior with people [54, 76], but these metrics do not include actual human evaluations. They do not assess whether people can accurately distinguish the AI player from the human one, which is vital for assessing human likeness in games [22] and beyond [88].

Studies with human evaluations tend to be small-scale surveys to understand the opinions regarding human-likeness [20, 52]. They often offer only a preliminary investigation into the specific characteristics that inform these beliefs and typically do not include a form of Turing test [39], a well-established framework for addressing these problems [22]. Work that uses a Turing test often does not investigate the behaviors or provide concrete metrics [55], or it focuses on assessing the full spectrum of game behaviors [5, 20, 52]. Due to the complexity of these games and the resulting behaviors, providing concrete recommendations to game designers is challenging. In contrast, we focus on a specific but widely-used behavior: point-to-point navigation. To perform our assessment, we utilize the setup of the recently-proposed Human Navigation Turing Test [16]; however, we propose and perform a deeper evaluation of human assessments of AI and human behavior.

3 BACKGROUND AND PRELIMINARIES

We utilize the navigation task from previous work [16] and instantiate in the same modern AAA video game for our experiments. We first describe the game in more detail, then provide an overview of the navigation task.

3.1 The Video Game

To enable the reuse of agent and human-generated videos in our study, we choose the same game as previous work. This game is a multiplayer online combat game that features 13 customizable characters, each with special abilities. The game is commonly compared with other popular team-based action games, such as Overwatch and DotA. Players compete against one another in two teams of four. The game has two game modes. One mode requires capturing and defending specific locations (called objectives) on the map, while the other involves collecting items called cells and deposit them to active platforms on the map. The game's team-based mechanics, objective balancing, and character customization offer a distinct multiplayer experience, making it an excellent choice for studying both AI behavior and human-AI interactions.

Underlying the game is the crucial mechanic of goal-directed navigation: players must move from one location to another to collect powerups or cells, go to drop-off platforms when they are active, and engage in combat with other players. As a result, navigation between points represents an abstraction of the most common task in the game. To allow us to concentrate on characteristics specific to navigation, we utilized a simplified version of the game that excludes other complex mechanics and objectives.

3.2 The Navigation Task

We instantiate the navigation task in the same way as prior work: a single avatar must navigate to a target location. The left screenshot of Figure 1 shows this location, indicated by the three blue containers. Navigating to a goal is a subtask of the main game, in which players must balance navigating to target locations to collect cells or boost health while warding off other players.

Before the player moves, the navigation target spawns uniformly at random in one of 16 possible locations, denoted by the green crosses in the right-hand image of Figure 1. Then, the player spawns on an island outside the main map (shown in the bottom portion of the mini-map) and must jump to the map's main area using the available jump areas. Once the player is in the central region, they can move to the target location.

The HNTT asks human judges to identify which of two navigation behaviors more closely resemble how people navigate in *reality*. This phrasing aims to capture how *convincing* an agent is [48], in contrast to another interpretation of the Turing test: whether a human or AI agent *controls* an entity. We chose this phrasing because we want to create *convincing* NPCs that contribute to an immersive game experience. In contrast, we do not wish to deceive the player into thinking that an agent is controlled by a person when it is not.

3.3 The Baseline Agents

Previous work [16] conducted their study with two agent types: a *symbolic* and a *hybrid* agent. When presented with the two agents, participants accurately detected human players above chance, meaning that people did not perceive their behavior as sufficiently human-like. We utilize these agents as baselines in our experiments, so we describe their essential details.

To progress toward the goal location, the agents take actions from a prespecified set (called an action space). This action space consists of 8 possible actions: do nothing, move forward, and move left and right (30, 45, and 90 degrees on each side). To facilitate training, the agents receive a dense reward signal to encourage successful navigation to the goal. It consists of the following terms: a -0.01 per-step penalty to encourage the agent to efficiently reach the goal, a -1 one-time penalty for dying because the agent may fall off the map, an incremental reward for approaching the goal, and a +1 reward for reaching the goal. We observed that this reward signal only includes terms to encourage successfully reaching the goal as quickly as possible.

The main difference between these two agents is the observations that they take as input. The *symbolic* agent receives only a semantic, low-dimensional representation as input; the *hybrid* agent also receives an image input. For more details about the baseline agents, we refer an interested reader to Appendix A.1 and Devlin et al. [16].

4 BUILDING A MORE HUMAN-LIKE AI

To help design our *reward-shaping* agent, we analyze the *hybrid* and *symbolic* agents to find characteristics that may have influenced the previous judgments of human likeness. Based on this analysis, we introduce a novel agent for the HNTT: the *reward-shaping* agent.

4.1 Designing our *Reward-Shaping* Agent

This agent extends the *hybrid* agent with two critical changes to promote learning of human-like behavior. Specifically, we introduce additional terms to the reward signal and expand the action space available to the agent. To test whether our contributions result in differences in perceptions of human likeness, we fix all other components of our *reward-shaping* agent to be the same as the *hybrid* agent.

Because the *symbolic* and *hybrid* agents previously exhibited non-human-like behavior, we inspected examples of their generated navigation and isolated three classes of problematic behavior. Agents would:

P1. Wildly swing camera angles or make sudden turns,
P2. Frequently collide with walls, and
P3. Sometimes move more slowly than expected.

To correct these behaviors, we utilize reward shaping [63] by including terms corresponding to desired or undesired behavior. We introduce the following terms. First, we include a camera angle difference penalty for swift camera angle changes over a set 0.15 difference threshold value to combat **P1**. Second, we introduce a penalty of -0.05 for any wall collisions to address **P2**. Third, to address **P3**, we provide a penalty of -0.01 if the distance traveled between steps is lower than an environment-specific threshold value of 220 map units. We choose these values in line with previous training rewards and expert assessments of the relative importance of each of the components.

To encourage smoother control and avoid abrupt turns, we utilize an approach similar to action-space shaping [37] by introducing additional available actions to the agent. Intuitively, we anticipate that the introduction of finer-grain controls will yield more fluid navigation. We extend the action space to 14 actions from the previous 8. In addition to the 'do nothing' and 'move forward' actions, we include 6 degrees of turning left and right, rather than the 3 used by the baselines. The updated list of turning degrees for this agent is: 18, 36, 45, 54, 72, and 90 on each side.

Taken together, these two components comprise the novel aspects of the *reward-shaping* agent. We design this agent in a relatively *agnostic* way to make it more accessible to those without expertise in deep reinforcement learning. Consequently, these two components can be applied to any state-of-the-art deep reinforcement learning algorithm. Depending on the underlying algorithm, the specific values, particularly those used for each term of the reward signal, may need to be set differently. However, we believe that adjusting these values is more intuitive than specifying complex parameters that are specific to a particular algorithm.

4.2 Producing High-Quality Navigation

We train all agents to achieve a similar level of performance on the navigation task (see Appendix A.2 for the details of our training setup) to ensure that task skill is not responsible for the perceived differences in human likeness. We measure task proficiency using the number of steps needed to reach the goal. Each step corresponds to around 5 seconds of real-time play. Figure 2 confirms that the agent models are indeed representative of state-of-the-art techniques for learning navigation in complex, 3D games.

The *reward-shaping* and *hybrid* agents exhibit higher variance during training than the *symbolic* agent. Because these agents must also learn from pixels, their learning task is more challenging than the *symbolic* agent (that only takes in symbolic input). As a result, we expect higher variance during training as the agent learns this more complex task. Importantly, all agents learn to reliably reach the goal, indicated by the performance near the end of training. A skilled agent now takes approximately 60 steps to complete the task (about 12 seconds of real-time play). This result ensures that

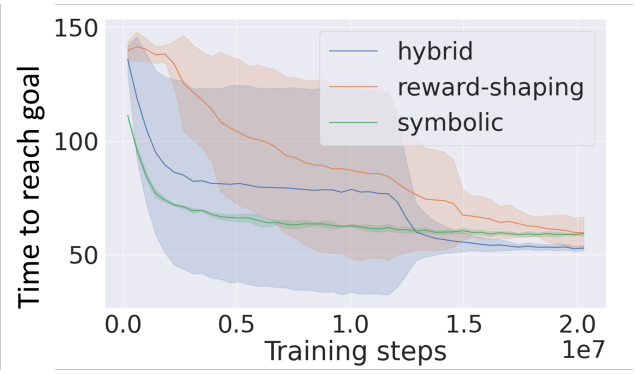


Figure 2: Hybrid, symbolic, and reward-shaping agents successfully learn to navigate. This plot shows the average amount of time needed to solve the task (y-axis) as a function of the amount of time taken to train the agent. The shaded area shows the standard deviation. For reward-shaping, N=3; for hybrid and symbolic, N=4). All curves are smoothed with a rolling window of 200. Importantly, on average, all agents converge to solve the task in around 60 steps (around 12 seconds of in-game time). In contrast, agents start out needing around 140 steps (around 28 seconds of in-game time) on average to solve the task. The starting performance on this task is similar to how long an agent taking random actions would take to solve it. The main takeaways are that performance differences are not responsible for perceived differences in human likeness, and standard metrics of task performance are insufficient to assess human likeness.

differences in the human-likeness of assessments are not due to differences in the ability of the agents to solve the task.

5 EXPERIMENTAL DESIGN

To understand what characteristics people believe are indicative of human likeness, we conducted a behavioral study with human participants. Our setup closely follows prior work [16]; however, we introduce important extensions, including collecting assessments from a greater number of participants using a crowd-sourcing platform (MTurk) and additional data for a more thorough analysis. For completeness, we detail the full study design here.

5.1 Experimental Task

We asked each human participant to act as a judge by completing a survey consisting of 6 HNTT trials. In each HNTT trial, the judge was presented with two side-by-side video stimuli of people or agents completing the navigation task. After watching these videos, the judge answered three questions to indicate which video they believed navigated more like a human would in the real world, a justification of their response, and an indication of their certainty. More specifically, participants answered the following questions:

(1) **Which video navigates more like a human would in the real world?** The judge clicked the button underneath the video that they believed navigated more like a human would. This decision was a forced binary choice.
(2) **Why do you think this is the case? Please provide details specific to the video.** The judge answered this question as a free-form response in the box below the question.
(3) **How certain are you of your choice?** The judge answered this question on a 5-point Likert scale, with choices ranging from extremely certain to extremely uncertain.

To mitigate subject learning effects from sequentially viewing multiple videos, we did not reveal to the judges which of the videos was AI-generated. In other words, participants completed each task and, in the end, did not know which videos were human-generated.

Study	Number of participants	Number of trials
Human vs. *hybrid*	50	6
Human vs. *symbolic*	50	6
Human vs. *reward-shaping*	92	6

Table 1: Conditions tested in each study and the number of trials per condition. Importantly, note that the human vs. *hybrid* and human vs. *symbolic* studies are replications of prior work [16] to validate the switch to a crowd-sourcing platform.

5.2 Experimental Procedure

We completed 3 studies; each study pitted a human-controlled agent against a different AI agent. Within each study, all judges viewed the same 6 trials. The trials were presented in a randomized order per judge. Within each trial, the ordering of the two videos was randomized, such that the human-generated video could not be inferred by presentation order. Table 1 outlines the conditions tested in each study.

Each participant first read through an introduction page with the required task instructions (see Appendix B for the full text). They then completed a consent form and read through a background page with brief details about the video game. They answered a series of questions to assess their comprehension of the task and familiarity with video games. Finally, participants engaged in the 6 HNTT trials. Figure 3 shows screenshots of the comprehension and familiarity questions (a) and an example HNTT trial (b).

5.3 Navigation Video Generation and Sampling

A key part of the study is the videos that were shown to the human judges. For the human-generated navigation data and videos, we use the publicly-available sample published by previous work [16].[1] We sampled human videos from the 40 published under their "study 1" protocol. To generate the AI navigation data, we select each agent's most recently saved version. Then, we instantiate a new session and deploy the agent in the game 100 times, producing 100 total

[1]Data use under MSR-LA license. License details can be found in the original authors' GitHub https://github.com/microsoft/NTT. This link includes *all* data used in this study, including from our *reward-shaping* agent.

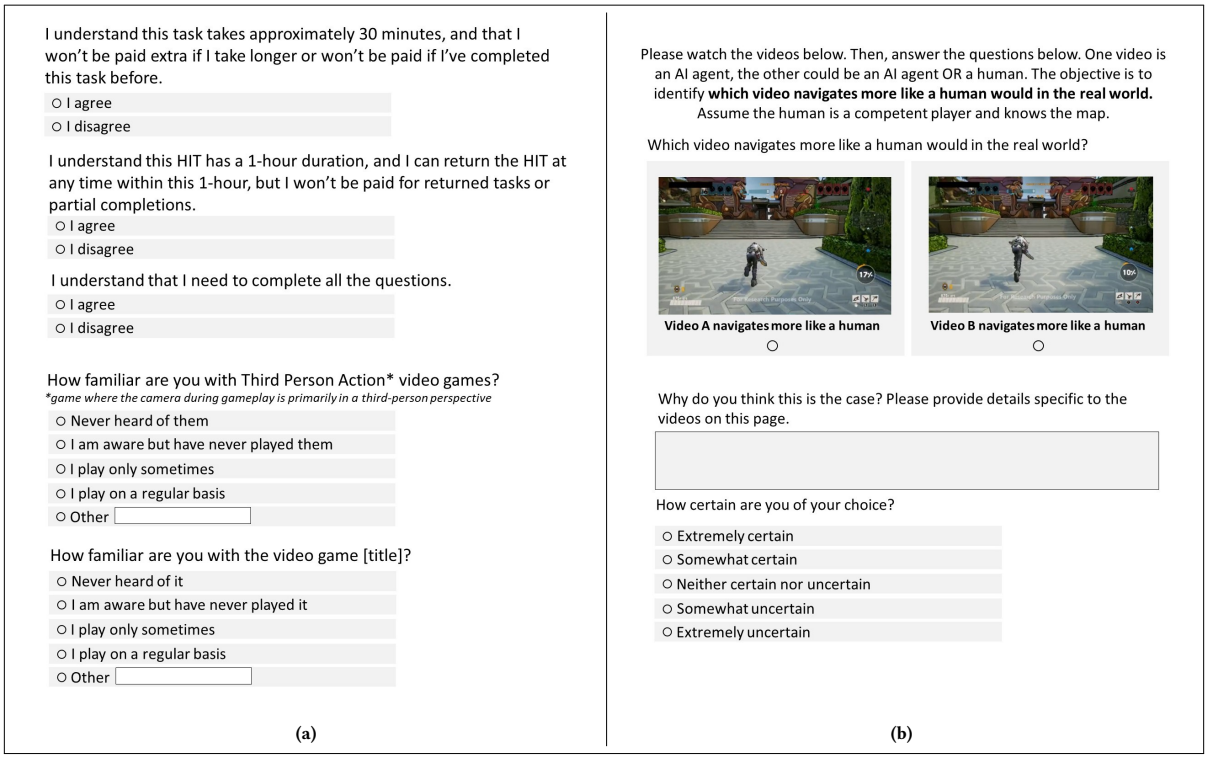


Figure 3: Screenshots of HNTT survey questions. The screenshot in (a) shows the comprehension and familiarity questions (asked once per participant). We gauge the participant's familiarity with the time the task will take, understanding of task completion, familiarity with third-person action video games, and familiarity with the video game used in the survey. The screenshot in (b) depicts one HNTT trial. We ask participants to choose their response to the human likeness question, justify it, and indicate their level of certainty.

navigation videos per agent. To produce the video stimuli used in the study, we sample the recordings uniformly at random.

We implemented several measures to standardize the videos and ensure that any measurement noise applied to all conditions. First, we checked that any changes in light applied similarly across conditions. Second, we designed the timing of the stimuli to ensure that participants had sufficient time to engage in and provide meaningful responses in all trials. As a result, we did not use videos that were too long and excluded videos shorter than 10 seconds (before post-processing) because they were deemed too short to assess navigation quality in pilot studies. Third, because the goal locations may differ depending on the game-controlled initialization, we matched the goal locations of the human videos with the AI agent videos. Consequently, we used different human videos for different studies. Fourth, we applied the post-processing steps from prior work [16]), including masking identifying information, adding a "For Research Purposes Only" watermark, and cutting out the last few seconds of the human videos. We implemented the last change to correct an effect of the data collection process, where the human players manually ended their recording, adding a few seconds at the end of the videos.

5.4 Other Experimental Control

The MTurk crowd-sourcing platform [60] is widely used for data collection and research due to its scalability, as long as researchers implement appropriate steps for quality control [32]. Here, we detail the study inclusion criteria that we implemented for quality control.

We set the following MTurk requirements for survey participation: location is United States, age is 18 or older, and language is English. We did not collect demographic information or any other personally identifiable information. To target more experienced MTurk Workers, we set the following Human Intelligence Task (HIT) qualifications: HIT Approval Rate greater than 98%, Number of HITs Approved greater than 500, and a qualification to prevent repeat responses. To incentivize quality, we included a bonus payment for each high-quality response. We reviewed the free-form answers to find low-quality or suspected bot responses; for example, we excluded from analysis responses with high instances of typos, copy/pasted answers, or nonsensical wording. We paid all participants who completed the task for the HIT, even if their response was identified as low-quality. The low-quality responses did not receive the bonus payment. We paid on average 15 USD per hour. We obtained approval for our studies from our Institutional Review Board (IRB) and informed consent from each participant.

We included details of the study and a description of any potential participant risks in the consent form.

6 ANALYSIS

Our primary objective is to evaluate the human-likeness of the agents using both quantitative and qualitative measures. To quantify the ability of the human judges to distinguish between the human-like and non-human-like agents, we analyze their accuracy scores and self-reported uncertainty. To identify the factors that influence their perceptions of human likeness, we adopt a qualitative approach. We construct and use codes to summarize the reasons cited in the open-ended responses and compare the frequency of these codes across different settings.

6.1 Assessing Human-Likeness

We first aim to identify which agents pass the HNTT according to our proposed criterion. Because existing work demonstrates differences in assessment ability depending on expertise, we seek to identify whether this phenomenon holds in our setting. We finally seek to investigate the relationship between self-reported uncertainty and accuracy when assessing the agents. We instantiate the following research questions:

RQ 1. Which agents are judged as being human-like?
RQ 2. Do the judges exhibit greater accuracy in assessing human likeness as a function of their experience with games?
RQ 3. What is the relationship between the accuracy of human judges and their self-reported uncertainty?

To answer **RQ 1**, we propose a firm criterion for deciding whether an agent is sufficiently human-like, formalizing the question: *are human assessors unable to distinguish between agent and human behavior*? We implement this criterion as a statistical test that determines whether human judges distinguish between human and agent behavior at a level significantly different from chance. We instantiate this test by computing the 95% confidence interval for the median of the human-agent comparisons using bootstrap sampling (a non-parametric approach). If the 95% confidence interval includes 0.5 (chance-level agreement), then the agent passes the HNTT.

For both **RQ 2** and **RQ 3**, we compare our variables of interest with *accuracy*. We define accuracy to mean that the participant identified that the human-generated behavior was more human-like than the AI-generated behavior. To answer **RQ 2**, we compare accuracy to the self-reported familiarity of the participants with action games in general and the specific game in the study. To answer **RQ 3**, we examine the self-reported uncertainty of the judges and its relationship to accuracy.

6.2 Assessing Human-Like Characteristics

To analyze the *characteristics* that correspond to assessments of human likeness, we instantiate the following research questions:

RQ 4. Are there key differences between how people characterize human-like and non-human-like behavior? Does this differ when the agent does or does not pass the HNTT?
RQ 5. What is the relationship between the characteristics that people use to assess human likeness and their ability to accurately assess it?

We selected a sub-sample of the responses from the *hybrid* agent and the *reward-shaping* agent studies for analysis. We chose these studies to enable comparison between an agent that does not pass the HNTT with one that does (see Section 7.1). We first randomly sampled a set of 55 responses to compute the initial agreement, called the *agreement sample*. We filtered this sample to 53 after removing responses that were ambiguous or could not be categorized by any of our codes. We then constructed the sample for analysis by randomly sub-sampling three free-form responses per judge for each study. To minimize bias, we shuffled responses before sampling. We removed responses that were ambiguous or could not be categorized by any of our codes, resulting in a dataset of 395 responses for our analyis of human-like characteristics.

We followed a pair-coding approach to annotate the data. The annotator with more familiarity with the data proposed an initial list of codes derived from previous work [91]. Following established notation [4], we have a set of I items (or responses), labeled as at least one of the K categories by $C = 2$ coders. We decompose each label as more or less human-like H = (`more, less`) and quantify its direction D = (`more, less`), when applicable. For example, if we label item i as *smoothness of movement*, we note whether the judge considered the behavior human-like and whether they noted it as being more + or less − smooth.

The two annotators then convened to discuss the meaning of the codes and jointly code a set of 5 responses. Table 2 illustrates an example of a coded response. After that, the two annotators separately coded the agreement sample with the initial set of codes. Optionally, the annotators could label responses as *other* and provide specific examples to enable revisions of the codes if other themes emerged. The two annotators iteratively reconvened to discuss disagreements and refine the codes. After multiple rounds of discussion, independent coding, and disagreement resolution, the annotators fixed the set of codes (Table 3) and their inclusion criteria to label the full sample.

Because we aim to design human-like AI agents, we want to identify codes that could be utilized by AI designers. For that reason, when deciding on codes, we prioritize codes that refer to specific behaviors over more general ones. For example, a collision avoidance behavior could be coded as goal-directed; however, we code it only as collision avoidance. This protocol promotes the independence of categories while prioritizing specific, lower-level behaviors to use in designing agents. When coding, the annotators first consider whether the response could be categorized as a lower-level code, then move to more general codes if needed. Appendix C contains more details about this process.

The annotators achieved an overall average inter-annotator agreement of 0.84 on the agreement sample. We calculate inter-annotator agreement with binary Cohen's kappa κ [13] over K, D, and H, as previously defined. See Table 4 for more details. After fixing the list of codes, the annotators divided the data sub-sample such that there was overlap on 25% of the data (99 items). We report Cohen's kappa in Table 4 for the overlapping sample to ensure that

Response	Free-Form Response	More Human-Like	Less Human-Like
B	The character in Video B runs in straight lines and goes to where he needs to be going. The character in Video A is running in circles,into objects, etc.	Smoothness of movement +; Goal directed +	Collision avoidance −; Goal directed −

Table 2: Example coded response to the question, "Which video navigates more like a human would in the real world?". The leftmost column indicates that this judge believed Video B to exhibit the more human-like behavior. The highlighted text illustrates the annotation process. The judge identifies that the more human-like character runs in straight lines (more human-like code: smoothness of movement +) and navigates to the goal (more human-like code: goal directed +), while the character that they believe is less human-like runs in circles (less human-like code: goal directed −) and into objects (less human-like code: collision avoidance −).

Annotation Code	Shorthand	Definition	Key Words and Phrases	Example Snippet
Smoothness of movement	smooth	The quality of the agent's navigation or camera movement	Smooth, jerky, straight, swerve, steady, fluid	Movements are way more smooth
Goal directed	goal	How goal-directed the agent's behavior seems	Intention, focus, knew where to go	Deliberate camera movements
Collision avoidance	avoidance	Whether the agent avoids collisions	Collide, avoid, crash runs into obstacle	Runs into a box
Environment receptivity	receptivity	Whether the agent understands and/or properly interacts with the environment	Explore, stay on path, collect power-ups	Ignores all the health/mana/etc
Intuition	intuition	The judge cannot pinpoint behaviors	Natural, feeling, seems to be	Just a feeling
Self-reference	self-reference	Relationship to the judge's own movement or play	Like I play	[Like] how I navigate with that ... view

Table 3: Annotation code definitions. The codes used to label the free-form responses are presented in the leftmost column. The middle-left column shows the corresponding shorthand for the codes, used later in the paper. In the middle column, a brief definition of each code is presented. The middle-right column lists the keywords and phrases that the annotators used to determine if a response could be labeled as containing a particular code. An example snippet of a response that would be labeled with that code is provided in the rightmost column. Although the included examples are fairly clear, the free-form responses often contain more ambiguous content.

our understanding of the codes did not overfit the specific examples in the agreement sample.

We provide a more detailed discussion of the annotation codes and inclusion criteria. Table 3 includes these definitions and phrases that helped us identify the presence of each code. For each code, we provide a supporting example to give the reader a sense of what common responses may look like. *Smoothness of movement* refers to the quality of the agent's navigation or camera movement. This code considers both immediate jerky actions and temporally-extended zig-zagging behavior. *Goal directed* refers to how intentional the agent's behavior appears. We include descriptions of behavior that pertain to a perceived goal, even if that goal is not the primary one. We include the code *collision avoidance* because it is a long-standing area of research in the robotics community [67]. This code refers to intentional behavior to redirect from a potential crash. *Environment receptivity* aims to capture the agent's relationship with the game environment, its contextual understanding, and adherence to norms. In a real-world setting, this might look like a person walking on a path instead of the grass or crossing the street when permitted by a pedestrian signal. Any responses that refer to non-specific feeling that a behavior was more human-like are categorized as *intuition*. We include this code to capture instances where participants can identify what they believe is more human-like behavior but struggle to express it. Finally, we include *self-reference* as a code to capture when judges relate the agent's behavior to their own play.

During the iterative coding process, the annotators assessed the likely causes of disagreements. After resolving mistakes and other easy-to-resolve issues, the annotators determined that the remaining disagreements arose from individual differences in interpreting ambiguous natural-language responses. This cause means that neither annotator can be treated as more correct for disagreement resolution. The annotators, therefore, decided on the following disagreement resolution scheme. When a disagreement arises in at least one label for an item annotated by both annotators, we randomly choose an annotator to treat as correct and use their labels.

7 RESULTS

We first present the results from our analysis described in Section 6.1; in particular, we demonstrate that our reward-shaping agent passes the HNTT while other agents do not. We then present the results from our analysis described in Section 6.1 by highlighting characteristic behaviors and key differences in how human judges perceive AI vs human players. We find that people tend to utilize

Annotation Codes	Direction	Cohen's κ Agreement Sample	Cohen's κ Overlapping Sample
Smoothness of movement	More +	0.90	0.79
	Less -	0.64	0.64
Goal directed	More +	0.82	0.78
	Less -	0.82	0.63
Collision avoidance	More +	1.00	1.00
	Less -	0.64	0.73
Environment receptivity	More +	0.82	0.93
	Less -	0.73	0.67
Intuition		1.00	0.87
Self-reference		1.00	1.00
	Average	**0.84**	**0.78**

Table 4: Per-code Cohen's κ score. The two annotators achieved an average Cohen's κ score of 0.84 over all of the codes for the *agreement sample*. According to Cohen's suggested interpretation, we achieve at least moderate agreement on each category and achieve almost-perfect agreement on 7 of the 10 categories when annotating the agreement sample. When annotating the *overlapping sample*, the two annotators achieved an average Cohen's κ score of 0.78 over all of the codes. According to Cohen's suggested interpretation, we achieve at least substantial agreement on each category. There was only a small overall decrease in agreement between these two settings, indicating that our coding process is fairly general.

Agent	Median Accuracy (IQR) [95% CI]
symbolic	0.83 (0.67 – 1.00) [0.67, 1.00]
hybrid	0.83 (0.67 – 1.00) [0.83, 1.00]
reward-shaping	0.50 (0.33 – 0.67) [0.50, 0.50]
Agent	Median Uncertainty (IQR)
symbolic	2.17 (1.67 – 2.42)
hybrid	1.92 (1.33 – 2.25)
reward-shaping	2.17 (1.75 – 2.67)

Table 5: Full summary statistics of accuracy and uncertainty. We show the median accuracy (IQR=Q1-Q3) for each agent, reported as non-parametric measures of central tendency and spread; we report 95% confidence interval and median uncertainty (IQR=Q1-Q3) of the human-agent comparisons for each agent. Only the *reward-shaping* agent passes the HNTT according to our proposed metric.

similar high-level characteristics when characterizing human-like behavior. However, their beliefs about AI capabilities may inform whether they think AI agents more or less strongly exhibit these characteristics.

7.1 Analysis of Human Likeness

Only the *reward-shaping* agent passes the HNTT. Table 5 shows the full summary statistics, which are computed over the full dataset from our survey. Each bootstrap calculation is run over 10000 iterations. The *symbolic* and *hybrid* baseline agents do not pass the HNTT according to our criterion. The judges had median accuracies of 0.83 (*symbolic* agent, 95% CI=[0.67, 1.0]) and 0.83 (*hybrid* agent, 95% CI=[0.83, 1.0]), indicating that they distinguish the agents from humans significantly higher than chance level. In contrast, our *reward shaping* agent passes this test of human-likeness: the median accuracy has a 95% confidence interval that includes 0.5 (chance-level agreement). This result suggests that the judges cannot consistently differentiate between the *reward shaping* agent and the human player (*reward shaping* agent, median accuracy=0.50, 95% CI=[0.50, 0.50]).

Because the sample sizes of the trials differ (50 samples for the human vs. *hybrid* and human vs. *symbolic* conditions; 92 samples for the human vs. *reward-shaping* condition), we validate our results by subsampling the data for the *reward-shaping* agent to 50 samples, then run the bootstrap sampling procedure 100 times. We find that the computed CI always contains 0.5, or chance-level agreement, in each run of the bootstrap. The average median accuracy is 0.50, with a variance of 0.00; the averaged CI is [0.44, 0.63], with a variance of 0.01 for the lower bound and 0.00 for the upper bound. We, therefore, answer our **RQ 1**: the *reward-shaping* agent is the only agent that is judged as human-like according to this proposed metric.

There is no relationship between game familiarity and ability to accurately assess the human likeness of the AI agents. For each study, we perform a multiple linear regression analysis to test whether specific game familiarity and general game familiarity significantly predicted accuracy in assessing human-likeness. There is no relationship between either of the self-reported familiarities and accuracy for all agents. For the *symbolic* agent, the fitted regression model was:

$$\texttt{accuracy} = 0.68 - 0.01(\texttt{specific game familiarity}) + 0.03(\texttt{general game familiarity}).$$

The overall regression was not statistically significant ($R^2 = 0.01$, $F(2, 47) = 0.21$, $p = 0.814$). Decomposing the results further, neither specific game familiarity ($\beta = -0.01$, $p = 0.878$) nor general game familiarity ($\beta = 0.03$, $p = 0.525$) predicted accuracy.

For the *hybrid* agent, the fitted regression model was:

$$\texttt{accuracy} = 0.67 - 0.03(\texttt{specific game familiarity}) + 0.06(\texttt{general game familiarity}).$$

The overall regression was not statistically significant ($R^2 = 0.06$, $F(2, 47) = 0.26$, $p = 0.261$). We found that specific game familiarity did not significantly predict accuracy ($\beta = -0.03$, $p = 0.377$). General game familiarity also did not significantly predict accuracy ($\beta = 0.06$, $p = 0.109$).

Turning our attention to the *reward-shaping* agent, the fitted regression model was:

$$\texttt{accuracy} = 0.40 - 0.02(\texttt{specific game familiarity}) + 0.04(\texttt{general game familiarity}).$$

The overall regression was again not statistically significant ($R^2 = 0.02$, $F(2, 89) = 0.94$, $p = 0.393$). This result holds for both specific game familiarity ($\beta = -0.02$, $p = 0.522$) and general game familiarity ($\beta = 0.04$, $p = 0.191$).

These findings suggest that game familiarity is generally *not* predictive of accuracy for this specific task, answering **RQ 2**. In contrast, previous findings have demonstrated a relationship between the ability to assess human likeness and familiarity with the domain of study. We suspect that this result differs because we are studying a relatively simple setting, in which most people have strong priors about what constitutes human likeness. Navigating by walking or running is an activity that most people either perform or observe daily, meaning we will likely have a strong internal sense of human-like movement — even if we are not familiar with games that require navigation. In contrast, we hypothesize that game familiarity would be predictive of accuracy in the full game setting, implicating the importance of assessing human likeness in more complex settings as an important next step.

Human judges exhibit less false confidence in their assessments of the *reward-shaping* agent. We assess the median uncertainties of participants; lower values correspond to more certainty and higher values correspond to less certainty. Table 5 depicts the results of this analysis. Participants reported similar levels of uncertainty when assessing our *reward-shaping* agent (median=2.17, IQR=(1.75-2.67)) and the symbolic agent (median=2.17, IQR=(1.67-2.42)). In comparison, participants reported higher certainty when assessing the *hybrid* agent (median=1.92, IQR=(1.33-2.25)).

People felt less confident about their assessments of the *symbolic* and *reward-shaping* agents compared to the *hybrid* agent; however, participants more accurately detected human-generated behavior in the presence of the *symbolic* and *hybrid* agents. This result is surprising because it suggests that self-reported uncertainty and accurate assessments are not necessarily correlated. In other words, participants may exhibit false confidence in their ability to assess the human likeness of agents. We believe that participants may have been less certain about their assessments of the *symbolic* agent due to differences in the lengths of the videos: on average, the videos of the symbolic agents were 8.3 seconds long, whereas the hybrid agent videos were 15.3 seconds long. Participants may have not had enough time with the agent to accurately assess it. Taken together, the accuracy and uncertainty results indicate that, when presented with behavior from the *reward-shaping* agent, participants exhibited less false confidence in their assessment ability compared to when they were presented with behavior generated by the *hybrid* agent. This result answers **RQ 3**.

7.2 Analysis of Human-Like Characteristics

In all plots, we use the shorthand version of the codes, noted in Table 3, along with the + and − notation. The + and − notation indicate the degree, or direction, of the code. For example, smooth + indicates that the participant referenced more smooth movement, and smooth − indicates that the participant referenced less smooth movement.

Human judges rely on similar high-level characteristics when assessing human-like behavior. Figure 4 shows the codes that participants use to describe human-like and non-human-like behavior. We investigate the relative number of times a code was used

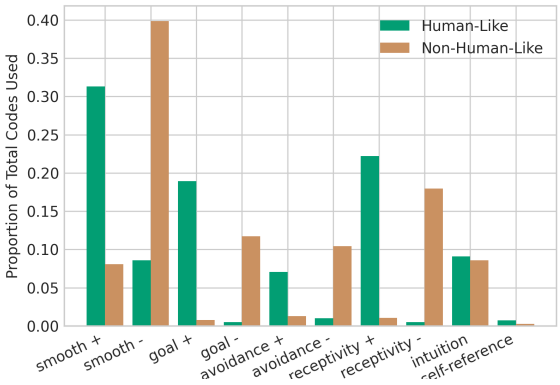
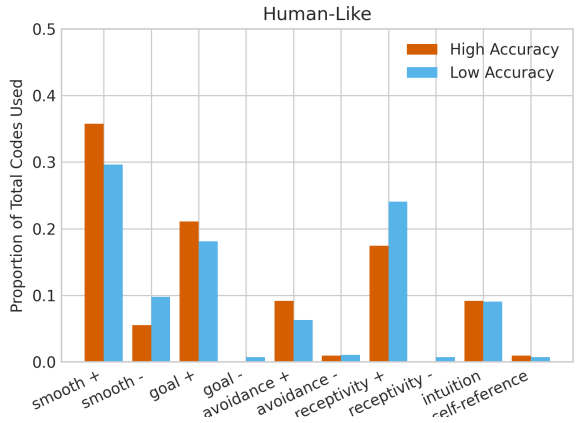


Figure 4: Codes used to describe human-like and non-human-like behavior. We compare the proportion of codes used to describe human-like and non-human-like behavior by human judges in their assessment of human likeness. People more frequently characterize human-like behavior as being more smooth, receptive and responsive to the environment, and goal-directed. In contrast, participants more frequently describe non-human-like behavior as being less smooth, receptive and responsive to the environment, and goal-directed.

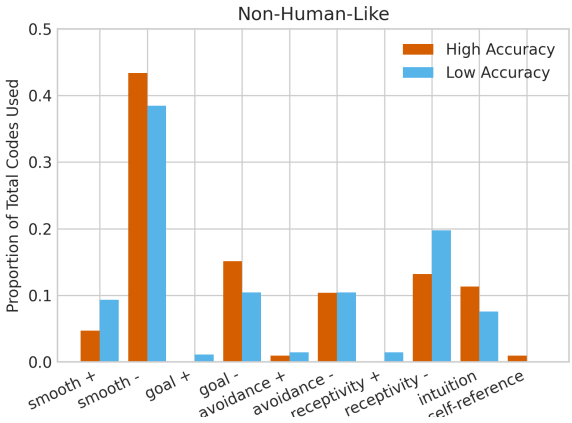
compared to all codes used to describe either human-like or non-human-like behavior (human-like and non-human-like code proportions should sum to 1). Judges tend to rely on similar high-level characteristics when characterizing human-like behavior. Overall, they most often reference the following high-level codes: smoothness of movement, environment receptivity, and goal-directedness. When we decompose the responses based on whether the behavior was assessed as human-like or not, we find that people more frequently characterize human-like behavior as more smooth, receptive and responsive to the environment, and goal-directed. In contrast, participants more frequently describe non-human-like behavior as being less smooth, receptive and responsive to the environment, and goal-directed. They rely on intuition and self-reference to a similar degree when describing human-like and non-human-like behavior.

We investigated these responses based on agent type but did not find a difference between the resulting proportions of codes. This result supports the assertion that people may have relatively stable beliefs what constitutes human-like behavior. Therefore, the rationale is only sometimes useful: in other words, looking for the jerkier agent only makes sense if the AI has not been designed to be less jerky than the person. We, therefore, conclude that, although people rely on different specific characteristics to determine human likeness, the general characteristics are relatively stable across different AI agents, which answers **RQ 4**.

Human judges that more accurately assess human likeness exhibit different beliefs about characteristics than human judges that less accurately assess human likeness, despite relying on the high-level characteristics to similar degrees. We divide the participants into two groups: high-accuracy



(a) Codes used to describe human-like behavior by low- and high-accuracy judges. We compare the proportion of codes used to describe human-like behavior by human judges in their assessment of human likeness.



(b) Codes used to describe non-human-like behavior by low- and high-accuracy judges. We compare the proportion of codes used to describe non-human-like behavior by human judges in their assessment of human likeness.

Figure 5: Codes used to describe human-like and non-human-like behavior, further decomposed by high- and low-accuracy judges. We compare the proportions of codes that are used to describe human-like behavior (left) and non-human-like behavior (right). We further decompose these codes by high- and low-accuracy judges to determine whether individuals who are more accurate rely on different features to rationalize their decisions. Interestingly, we see that the judges rely on similar characteristics to different degrees.

(greater than 80% of responses indicating the more human-like agents aligned with the human-generated video) and low accuracy (less than or equal to 80% of responses indicating the more human-like agents aligned with the human-generated video). We examine which codes are more frequently used to describe human-likeness by the participants in each group. Figure 5 shows this decomposition. Although high- and low-accuracy judges generally rely on similar characteristics, they do so to different degrees. For example, both types of judges refer to the high-level code of smoothness of movement in 40% of their codes when describing human-like behavior. Similarly, they both refer to the high-level code of smoothness of movement in around 49% of their codes when describing behavior that they do not perceive as human like. These results indicate that there is no difference in their tendency to rely on this characteristic to explain behavior. However, high-accuracy judges more commonly describe smooth motion when describing human-like behavior. In contrast, low-accuracy judges more often mention smoothness as a characteristic of behavior that is not human-like. This result further supports the idea that people's beliefs about AI capabilities may inform their assessments. In our case, the low-accuracy participants seem to share a similar belief that an AI agent is more capable than a person (by producing more smooth or "perfect" navigation).

Low-accuracy judges more often describe human-like agents as exhibiting more receptivity and responsiveness to their surroundings. A similar pattern emerges with less receptivity to justify non-human-likeness. This result indicates that low-accuracy participants may incorrectly attribute behaviors to interacting with the environment. As an example, a human judge that incorrectly identified Video A as being more human-like claims,

> Video A takes the more obvious route to the finish while B takes the longest possible one. A human generally would take the easiest route.

Interestingly, both high- and low-accuracy judges utilize intuition and self-reference to a similar level of frequency when assessing human-like behavior. In combination with the previous results showing that assessments of human likeness are influenced by stereotyped beliefs about AI capabilities, this finding suggests that some participants have better intuition because it aligns with the actual capabilities of AI agents.

8 DISCUSSION AND FUTURE DIRECTIONS

Although conducted in a limited scope, our findings should assist with future work on designing and evaluating human-like agents.

8.1 Limitations

Our study specifically evaluates the human-likeness of third-person perspective point-to-point navigation behavior in agents. Although this type of navigation is present in many settings, like pedestrian navigation in driving simulator [83] there are many other forms of navigation that exist in both real-world and virtual environments. Each of these types presents unique challenges and requires different strategies for designing human-like behavior. Although our study does not address all types of navigation, it provides a valuable starting point for evaluating the human-likeness of agents in one specific type of navigation. The codes that we identify are general enough to provide a starting point for researchers to analyze different forms of navigation. For instance, collision avoidance is a general characteristic that is persistent in many domains featuring

diverse types of navigation, like driving and running. Future work should consider expanding these evaluations to provide a more comprehensive understanding of how to design agents that behave in a more human-like manner.

Additionally, the analysis of the free-form responses revealed that there were different interpretations of the human likeness question. Some judges related the movement directly to human navigation in the real world. One judge said,

> In real life a human would almost certainly not jump down as far as the character in video A did without severely hurting themselves.

However, others related the movement to how human players would *control* an agent in a video game. Another judge mentioned,

> ... in Video A, the player bumps into a wall briefly before readjusting. This is something humans do when they get distracted and look away for a moment.

To investigate this disagreement, we annotated the agreement sample with which interpretation of the question the subject answered: real-world human navigation, video game navigation, and unclear. The two annotators had a high agreement for this annotation (Cohen's kappa: $\kappa = 0.94$). It was largely unclear which question the subjects were answering (40 out of 53 responses). However 11 responses referred to *video-game* navigation, while only 2 responses were clearly about real-world human navigation. We suspect that including the video game familiarity questions primed subjects to believe that the question was about video-game-specific navigation, rather than general human-like navigation. In future studies, we recommend that the study designers clarify which question is asked of participants by including an additional question that asks the participants the other interpretation of the question to provide an obvious contrast or describing the situation in which they would like the participants to envision themselves.

8.2 Designing and Evaluating Human-Like AI

Our study revealed that only the agent designed to display more human-like behavior passed our test of human likeness, highlighting the importance of explicitly incorporating these objectives when designing agents. However, determining what exactly constitutes human likeness requires careful consideration from designers. This assertion is further supported by the different interpretations of the human likeness question by the human judges. One interpretation of human likeness is acting as if the agent is controlled by a person, while the other refers to exhibiting more realistic behaviors. Both perspectives can be useful in different contexts.

When designers seek to automate parts of the development process, such as playtesting, it is more important to create agents that appear to be human-controlled. In automated playtesting of games [23, 68], AI agents that act like real users would enable video game designers to expedite the iterative development process while also alleviating the burden of game players to extensively evaluate new content. Users could provide feedback only after obvious bugs, like those related to movement, have been corrected, which may enhance their enjoyment of the feedback process. In *shared autonomy* [2], developing agents that behave like that user would enable a more seamless integration of semi-autonomous control with user inputs. For example, we observed that the judges called out strafing as an example of what a human would do in a video game. Strafing is a tactical, sideways maneuver that would not be performed by a person navigating in the real world. Incorporating these game-specific movements would likely increase the perception of the agent being human-controlled, especially by expert players. The creation of such agents would enable players who experience disruptions, like network issues, to still play cloud games [57]. When the system detects a disruption, it can take control and begin emulating human-like behavior. When the user can take back control, they can do so seamlessly. This can also be included as an option for players who desire in-game assistance for other reasons, such as mobility issues. Conversely, when the objective is immersion, producing more realistic navigation is essential.

In our study, we focused on producing more realistic navigation. To that end, we identified a set of high-level characteristics, such as smoothness of movement, that the judges relied on to assess human likeness. As a result, game AI designers can first focus on adjusting these characteristics. As we demonstrate with our *reward-shaping* agent, these characteristics may be targeted using simple techniques and assessed with an *automated* Turing test [16]. After handling the most frequently mentioned characteristics, designers can then focus on more fine-grained details, such as agents not walking in puddles, to reflect more real-world navigation.

Furthermore, we employed a *third-person* Turing test where participants watched videos of the agents navigating. Although the ability to pause, rewind, and replay the videos provided a means of interrogation, it was based solely on observation, and lacked the intervention-based approach of a typical Turing test. Intervention-based approaches could include changing the camera perspective, adversarially interrupting the AI agent's intended path, and more. These forms of interaction may yield different insights.

There are some downsides, however, to deploying a more interactive test, particularly at scale. Recruiting a human evaluator and a human player to interact requires their simultaneous availability for real-time feedback. One solution is in-person studies, which can be challenging to scale and deploy. For instance, at the time of this study, we could not run in-person studies due to the ongoing global pandemic or distribute our proprietary game build to remote participants. Future work could take advantage of advances in game streaming, which may enable interactive remote studies with proprietary game builds. This solution can also incorporate previous work on simultaneous recruitment of participants [7, 82]. However, constructing the architecture to incorporate these different technologies may require significant engineering effort.

Importantly, previous work has demonstrated that the inclusion of more direct ways to interrogate the agent by embodying the player and agent in the same virtual space can lead to limited insight [78]. Indeed, work that included an in-game assessment of the human or bot introduced the side effect of an additional game mechanic causing some players to prioritize either gameplay or on the believability assessment [28]. This division of attention yields unreliable results, leading to other researchers adopting third-person variants of these assessments [5, 16, 70]. As a result, we believe that the following pipeline could be useful for evaluating human-like agents. Designers can initially deploy a third-person Turing test to evaluate the human likeness of specific behaviors. The resulting characteristics can then be used to design a set of

agents that exhibit different behavior that depends on the most common beliefs of the participants. For example, agents could move more smoothly if the participant believes smooth movement to be a feature of human likeness. Players could then choose the characters that they want to interact with in the game, which would enable them to tailor the game to their own subjective experience and enjoyment. This approach may offer more reliable insights into the effectiveness of the agent's design without sacrificing the integrity of the assessment process. It could also empower game players by enabling them to exert control over their experience.

8.3 Toward More General Human-Like Agents

Although the specific agent created for our study may not generalize to different games, this is a common and open challenge in the field of AI [46, 61]. Instead, we offer suggestions for using feedback from designers and players (e.g., through user research) to train human-like agents more efficiently and effectively.

The differences in how more or less skilled human judges characterize human likeness suggests that different people have different interpretations of what constitutes human-like behavior. This supports the idea that the believability of NPCs in games is highly subject to the prior beliefs and expectations of the players. This finding aligns with the fundamental principle of *familiarity* [29] that centers the real-world personal experience and knowledge of the user and implicates the importance of *player-centered* design and customization [74]. Rather than producing monolithic human-like agents, we should strive to understand the beliefs of the player and tailor their experiences accordingly.

When moving to more complex settings, an additional difficulty is introduced. The evaluations of human likeness become even more subjective, varying based on individual differences and cultural factors [50]. This result underscores the importance of involving diverse groups of people in the evaluation of AI agents to obtain a more comprehensive understanding of how people perceive these agents. In the context of games, this could look like utilizing participatory design methods [65] to involve game players in the design of the AI agents themselves. With the consent of the players, we could use techniques in the area of learning from human feedback [10, 31, 90], which provide additional channels for people to communicate what they want from AI agents. With these techniques, players can provide training data to the agents in the form of preferences over paired demonstrations generated by the agent, demonstrations of the desired behavior, and more. This can help to ensure that the AI agents are designed with the needs and preferences of diverse groups in mind.

This approach can also be used to help reduce the burden on video game designers: in complex domains, it is often challenging to specify reward signals by hand [12, 42, 47]. In part, this difficulty stems from the complexity of the desired behavior: as we have shown, human-like behavior is multi-faceted and necessitates optimizing over multiple objectives. Furthermore, it is sometimes challenging to write down exactly what we mean when specifying a task. For instance, how do we construct a reward signal that captures the task of *build a house in a video game in the same style as surrounding houses?* [69]. When designing a reward signal for this task, we would need to encode what counts as a house, what components are most important to emulate in the style, and which structures count as houses. A person can quickly understand the intention of this instruction, but it is challenging to make explicit this implicit understanding.

As a result, an exciting avenue for future work involves developing more effective techniques for learning from people, evaluating user experiences of these techniques, and incorporating them into a flexible, user-friendly tool. This tool can also help extend this work to more general game settings. To more easily enable this line of work, assessments of human likeness could be incorporated into commonly-used game engines, like Unity [36]. This tool would enable game developers to easily evaluate the human likeness of their AI agents using metrics and benchmarks that have been validated in previous research. Additionally, this tool could contain libraries of pretrained *human-like* AI agents, which developers could use as a starting point for their own work. For example, developers could utilize a pretrained human-like navigation agent to perform navigation but develop their own algorithm to use for different tasks. Using this tool could save developers time and effort by enabling them to quickly and easily create more believable and engaging agents to enhance the player experience.

9 CONCLUSION

In this work, we aimed to understand how people assess human likeness in human- and AI-generated behavior in the context of navigation in a 3D video game. Toward this goal, we designed and implemented a novel AI agent to produce human-like navigation behavior. We deployed a large-scale study of human-generated navigation behavior with three AI agents, including our novel *reward-shaping* agent. We find that our proposed agent passes a Turing test, while the other two agents do not. We further investigated the justifications people provided when assessing these agents and found that people rely on similar higher-level characteristics when determining human similarity. In this context, we suspect that differences in the accuracy of assessing these agents are based more on fixed beliefs about the capabilities of AI systems rather than familiarity with the assessment domain of games. We conclude by discussing the limitations of the work, suggesting concrete design considerations for video game designers, and identifying a few critical areas for future research.

By highlighting design considerations and challenges, we hope that this paper will serve as a call for work that integrates perspectives and techniques from the HCI and AI communities. Building more general human-like agents requires careful design of both the agents and the evaluation protocol. Developing tools that can be incorporated into games and other settings enables quick iterations of these designs and the incorporation of these different techniques. At the highest level, we hope researchers can develop and evaluate agents that exhibit human-like behavior that improves human interaction with AI agents.

ACKNOWLEDGMENTS

We would like to thank Evelyn Zuniga, Guy Leroy, Mikhail Jacob, Mingfei Sun, and Dave Bignell for their contributions and feedback to an earlier study in this project. We would also like to thank Cecily Morrison, Youngseog Chung, and Max Meijer for their helpful

comments and feedback. We additionally thank the anonymous CHI reviewers for their detailed comments; the paper is significantly improved thanks to their suggestions. Co-author Fang is supported in part by NSF grant IIS-2046640 (CAREER).

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

[2] Peter Aigner and Brenan McCarragher. 1997. Human integration into robot control utilising potential fields. In *Proceedings of International Conference on Robotics and Automation*, Vol. 1. IEEE, 291–296.

[3] Eloi Alonso, Ubisoft La Forge, Maxim Peter, David Goumard, and Joshua Romoff. 2020. Deep Reinforcement Learning for Navigation in AAA Video Games. In *Challenges of Real-World Reinforcement Learning NeurIPS Workshop*. NeurIPS, Montreal, Canada, 1–13.

[4] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics* 34, 4 (2008), 555–596.

[5] Christian Arzate Cruz and Jorge Adolfo Ramirez Uresti. 2018. HRLB2: A Reinforcement Learning Based Framework for Believable Bots. *Applied Sciences* 8, 12 (2018), 2453.

[6] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* 1 (2019), 1–66.

[7] Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 33–42.

[8] Rodney A Brooks, Cynthia Breazeal, Robert Irie, Charles C Kemp, Matthew Marjanovic, Brian Scassellati, and Matthew M Williamson. 1998. Alternative essences of intelligence. *AAAI/IAAI* 1998 (1998), 961–968.

[9] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems* 32 (2019).

[10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*. 4299–4307.

[11] Georgios Christou. 2014. The interplay between immersion and appeal in video games. *Computers in human behavior* 32 (2014), 92–100.

[12] Jack Clark and Dario Amodei. 2016. Faulty reward functions in the wild. https://blog.openai.com/faulty-reward-functions

[13] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. https://doi.org/10.1177/001316446002000104 arXiv:https://doi.org/10.1177/001316446002000104

[14] David Conroy, Peta Wyeth, and Daniel Johnson. 2011. Modeling player-like behavior for game AI design. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*. ACM, New York, NY, 1–8.

[15] Sara De Freitas. 2018. Are games effective learning tools? A review of educational games. *Journal of Educational Technology & Society* 21, 2 (2018), 74–84.

[16] Sam Devlin, Raluca Georgescu, Ida Momennejad, Jaroslaw Rzepecki, Evelyn Zuniga, Gavin Costello, Guy Leroy, Ali Shaw, and Katja Hofmann. 2021. Navigation Turing Test (NTT): Learning to Evaluate Human-Like Navigation. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Virtual, 1–10.

[17] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. 2017. OpenAI Baselines. https://github.com/openai/baselines.

[18] Hippolyte Dubois, Patrick Le Callet, and Antoine Coutrot. 2021. Visualizing navigation difficulties in video game experiences. In *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 77–80.

[19] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. 2019. Implementation matters in deep RL: A case study on ppo and trpo. In *International conference on learning representations*.

[20] Nobuto Fujii, Yuichi Sato, Hironori Wakama, Koji Kazai, and Haruhiro Katayose. 2013. Evaluating human-like behaviors of video-game agents autonomously acquired with biological constraints. In *International Conference on Advances in Computer Entertainment Technology*. Springer, 61–76.

[21] Yolanda Gil and Bart Selman. 2019. A 20-year community roadmap for artificial intelligence research in the US. *arXiv preprint arXiv:1908.02624* (2019).

[22] Astrid Glende. 2004. Agent design to pass computer games. In *Proceedings of the 42nd annual Southeast regional conference*. 414–415.

[23] Stefan Freyr Gudmundsson, Philipp Eisen, Erik Poromaa, Alex Nodet, Sami Purmonen, Bartlomiej Kozakowski, Richard Meurling, and Lele Cao. 2018. Human-like playtesting with deep learning. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8.

[24] William H Guss, Cayden Codel, Katja Hofmann, Brandon Houghton, Noboru Kuno, Stephanie Milani, Sharada Mohanty, Diego Perez Liebana, Ruslan Salakhutdinov, Nicholay Topin, et al. 2019. NeurIPS 2019 competition: the MineRL competition on sample efficient reinforcement learning using human priors. *arXiv preprint arXiv:1904.10079* (2019).

[25] Rotem D Guttman, Jessica Hammer, Erik Harpstead, and Carol J Smith. 2021. Play for Real (ism)-Using Games to Predict Human-AI interactions in the Real World. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1–17.

[26] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. 2005. Microstructure of a spatial map in the entorhinal cortex. *Nature* 436, 7052 (2005), 801–806.

[27] Simon Hecker, Dengxin Dai, Alexander Liniger, Martin Hahner, and Luc Van Gool. 2020. Learning accurate and human-like driving using semantic maps and attention. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Piscataway, New Jersey, 2346–2353.

[28] Philip Hingston. 2010. A new design for a turing test for bots. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*. IEEE, Piscataway, New Jersey, 345–350.

[29] Vita Hinze-Hoare. 2007. The review and analysis of human computer interaction (HCI) principles. *arXiv preprint arXiv:0707.3638* (2007).

[30] Sean D Holcomb, William K Porter, Shaun V Ault, Guifen Mao, and Jin Wang. 2018. Overview on deepmind and its alphago zero ai. In *Proceedings of the 2018 international conference on big data and education*. 67–71.

[31] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–35.

[32] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, New York City, New York, 64–67.

[33] Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. 2022. Modeling strong and human-like gameplay with KL-regularized search. In *International Conference on Machine Learning*. PMLR, 9695–9728.

[34] Mikhail Jacob, Sam Devlin, and Katja Hofmann. 2020. "It's Unwieldy and It Takes a Lot of Time"—Challenges and Opportunities for Creating Agents in Commercial Games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 16. AAAI Press, Palo Alto, CA, 88–94.

[35] Magnus Johansson. 2013. Do non-player characters dream of electric sheep. *A thesis about players, NPCs, immersion and believability (Doctoral dissertation, Department of Computer and Systems Sciences, Stockholm University)* (2013).

[36] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, et al. 2018. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627* (2018).

[37] Anssi Kanervisto, Christian Scheller, and Ville Hautamäki. 2020. Action space shaping in deep reinforcement learning. In *2020 IEEE Conference on Games (CoG)*. IEEE, 479–486.

[38] Igor V Karpov, Jacob Schrum, and Risto Miikkulainen. 2013. Believable bot navigation via playback of human traces. In *Believable bots*. Springer, 151–170.

[39] Man-Je Kim, Kyung-Joong Kim, Seungjun Kim, and Anind K. Dey. 2018. Performance Evaluation Gaps in a Real-Time Strategy Game Between Human and Artificial Intelligence Players. *IEEE Access* 6 (2018), 13575–13586. https://doi.org/10.1109/ACCESS.2018.2800016

[40] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[41] Raph Koster. 2013. *Theory of fun for game design.* " O'Reilly Media, Inc.".

[42] Victoria Krakovna. 2018. Specification gaming examples in AI. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

[43] Guillaume Lample and Devendra Singh Chaplot. 2017. Playing FPS games with deep reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[44] Joni Lämsä, Raija Hämäläinen, Mikko Aro, Raine Koskimaa, and Sanna-Mari Äyrämö. 2018. Games for enhancing basic reading and maths skills: A systematic review of educational game design in supporting learning by people with learning disabilities. *British Journal of Educational Technology* 49, 4 (2018), 596–607.

[45] P Lavega et al. 2004. Traditional games and education to learn to create bonds. To create bonds to learn. *Studies in Physical Culture and Tourism* 11, 1 (2004), 9–32.

[46] Kuang-Huei Lee, Ofir Nachum, Mengjiao Yang, Lisa Lee, Daniel Freeman, Winnie Xu, Sergio Guadarrama, Ian Fischer, Eric Jang, Henryk Michalewski, et al. 2022.

Multi-Game Decision Transformers. *arXiv preprint arXiv:2205.15241* (2022).
[47] Joel Lehman, Jeff Clune, and Dusan Misevic. 2018. The surprising creativity of digital evolution. In *Artificial Life Conference Proceedings*. MIT Press, 55–56.
[48] Daniel Livingstone. 2006. Turing's test and believable AI in games. *Computers in Entertainment (CIE)* 4, 1 (2006), 6–es.
[49] Rossella Lorenzi. 2013. Oldest known gaming tokens dug up in Bronze Age Turkish graves. *NBC News* (2013). https://www.nbcnews.com/sciencemain/oldest-known-gaming-tokens-dug-bronze-age-turkish-graves-6c10920354
[50] Brian Mac Namee. 2004. Proactive persistent agents-using situational intelligence to create support characters in character-centric computer games. (2004).
[51] Colt McAnlis and James Stewart. 2008. Intrinsic detail in navigation mesh generation. *AI Game Programming Wisdom* 4 (2008), 95–112.
[52] Matheus RF Mendonça, Heder S Bernardino, and Raul F Neto. 2015. Simulating human behavior in fighting games using reinforcement learning and artificial neural networks. In *2015 14th Brazilian symposium on computer games and digital entertainment (SBGames)*. IEEE, 152–159.
[53] Lazaros Michailidis, Emili Balaguer-Ballester, and Xun He. 2018. Flow and immersion in video games: The aftermath of a conceptual challenge. *Frontiers in psychology* 9 (2018), 1682.
[54] Maximiliano Miranda, Antonio A Sánchez-Ruiz, and Federico Peinado. 2016. A Neuroevolution Approach to Imitating Human-Like Play in Ms. Pac-Man Video Game.. In *CoSECivi*. 113–124.
[55] Maxim Mozgovoy and Iskander Umarov. 2011. Behavior capture: Building believable and effective AI agents for video games. *International Journal of Arts & Sciences* 4, 20 (2011), 243.
[56] Kara Murias, Kathy Kwok, Adrian Gil Castillejo, Irene Liu, and Giuseppe Iaria. 2016. The effects of video game use on performance in a virtual navigation task. *Computers in Human Behavior* 58 (2016), 398–406.
[57] Mohamed Musbah, Matthew Mitchell Dixon, Geoffrey Jacoby Gordon, Mahmoud Adada, Soroush Mehri, Andrew James McNamara, and Jonathan David Morrison. U.S. Patent 11213746, Oct. 2019. Providing automated user input to an application during a disruption.
[58] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, Vol. 99. 278–287.
[59] John O'Keefe and Lynn Nadel. 1978. *The hippocampus as a cognitive map*. Oxford: Clarendon Press, New York City, New York.
[60] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419.
[61] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).
[62] Anthony E Richardson, Morgan E Powers, and Lauren G Bousquet. 2011. Video game experience predicts virtual, but not real navigation performance. *Computers in Human Behavior* 27, 1 (2011), 552–560.
[63] Ariel Rosenfeld, Moshe Cohen, Matthew E Taylor, and Sarit Kraus. 2018. Leveraging human knowledge in tabular reinforcement learning: A study of human subjects. *The Knowledge Engineering Review* 33 (2018), 1–26.
[64] Matthias Scheutz, Paul Schermerhorn, James Kramer, and David Anderson. 2007. First steps toward natural human-like HRI. *Autonomous Robots* 22, 4 (2007), 411–423.
[65] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.
[66] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* 1 (2017), 12 pages.
[67] Peter Seiler, Bongsob Song, and J Karl Hedrick. 1998. Development of a collision avoidance system. *SAE transactions* (1998), 1334–1340.
[68] Alessandro Sestini, Linus Gisslén, Joakim Bergdahl, Konrad Tollmar, and Andrew D Bagdanov. 2022. Automated Gameplay Testing and Validation with Curiosity-Conditioned Proximal Trajectories. *IEEE Transactions on Games* (2022).
[69] Rohin Shah, Cody Wild, Steven H Wang, Neel Alex, Brandon Houghton, William Guss, Sharada Mohanty, Anssi Kanervisto, Stephanie Milani, Nicholay Topin, et al. 2021. The MineRL BASALT competition on learning from human feedback. *arXiv preprint arXiv:2107.01969* (2021).
[70] Noor Shaker, Julian Togelius, Georgios N Yannakakis, Likith Poovanna, Vinay S Ethiraj, Stefan J Johansson, Robert G Reynolds, Leonard K Heether, Tom Schumann, and Marcus Gallagher. 2013. The turing test track of the 2012 mario ai championship: entries and evaluation. In *2013 IEEE Conference on Computational Inteligence in Games (CIG)*. IEEE, 1–8.
[71] Aaron Sloman. 1999. What sort of architecture is required for a human-like agent? In *Foundations of rational agency*. Springer, 35–52.
[72] Bhuman Soni and Philip Hingston. 2008. Bots trained to play like a human are more fun. (2008).
[73] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press, 1 Broadway, Cambridge, MA.
[74] Jonathan Sykes and Melissa Federoff. 2006. Player-centred game design. In *CHI'06 extended abstracts on Human factors in computing systems*. 1731–1734.
[75] István Szita. 2012. Reinforcement learning in games. In *Reinforcement learning*. Springer, 539–577.
[76] Fabien Tencé and Cédric Buche. 2010. Automatable evaluation method oriented toward behaviour believability for video games. *arXiv preprint arXiv:1009.0501* (2010).
[77] Christian Thurau, Christian Bauckhage, and Gerhard Sagerer. 2004. Learning human-like movement behavior for computer games. In *Proc. Int. Conf. on the Simulation of Adaptive Behavior*. 315–323.
[78] Julian Togelius, Georgios N Yannakakis, Sergey Karakovskiy, and Noor Shaker. 2013. Assessing believability. In *Believable bots*. Springer, 215–230.
[79] Edward B Tylor. 1879. THE HISTORY OF GAMES. *Fortnightly* 25, 149 (1879), 735–747.
[80] Teija Vainio. 2010. A Review of the Navigation HCI Research During the 2000's. *international Journal of Interactive Mobile Technologies (iJIM)* 4, 3 (2010).
[81] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.
[82] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326.
[83] Marcel Walch, Julian Frommel, Katja Rogers, Felix Schüssel, Philipp Hock, David Dobbelstein, and Michael Weber. 2017. Evaluating VR driving simulation from a player experience perspective. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2982–2989.
[84] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–6.
[85] Henrik Warpefelt. 2013. *Mind the gap: Exploring the social capability of non-player characters*. Ph. D. Dissertation.
[86] Eric Wiewiora. 2010. *Reward Shaping*. Springer US, Boston, MA, 863–865. https://doi.org/10.1007/978-0-387-30164-8_731
[87] Erik Wijmans, Manolis Savva, Irfan Essa, Stefan Lee, Ari S Morcos, and Dhruv Batra. 2023. Emergence of Maps in the Memories of Blind Navigation Agents. *arXiv preprint arXiv:2301.13261* (2023).
[88] Anthony Zador, Blake Richards, Bence Ölveczky, Sean Escola, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. 2022. Toward next-generation artificial intelligence: Catalyzing the neuroai revolution. *arXiv preprint arXiv:2210.08340* (2022).
[89] Yunqi Zhao, Igor Borovikov, Jason Rupert, Caedmon Somers, and Ahmad Beirami. 2019. On multi-agent learning in team sports games. *arXiv preprint arXiv:1906.10124* (2019).
[90] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. 2010. Modeling interaction via the principle of maximum causal entropy. (2010).
[91] Evelyn Zuniga, Stephanie Milani, Guy Leroy, Jaroslaw Rzepecki, Raluca Georgescu, Ida Momennejad, Dave Bignell, Mingfei Sun, Alison Shaw, Gavin Costello, et al. 2022. How Humans Perceive Human-like Behavior in Video Game Navigation. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–11.

Hyperparameter	Value
Batch size	2048
Dropout rate	0.1
Learning rate	2.5e-4
Optimizer	Adam [40]
Gamma	0.99
Lambda	0.95
Clip range	0.2
Gradient norm clipping coefficient	0.5
Entropy coefficient c_2	0.0
Value function coefficient c_1	0.5
Minibatches per update	4
Training epochs per update	4
Replay buffer size	5 x batch size

Table 6: Hyperparameters for training the *symbolic*, *hybrid*, and *reward-shaping* agents. We train all of the agents with the PPO algorithm [66]. For additional detail on what these hyperparameters correspond to, we encourage an interested reader to refer to the original PPO paper. We provide these hyperparameter values for reproducibility.

A DETAILS ABOUT REINFORCEMENT LEARNING AGENTS

In this section, we first provide information about the baseline agents used in our study. We then provide training details for all agents.

A.1 Baseline Agents

The *hybrid* and *reward-shaping* agents receive an additional input of a 32x32 cropped depth buffer visual input which the *symbolic* agent did not. The visuals present a third-person view of the agent in the environment. To process this additional visual channel, the *hybrid* and *reward-shaping* agents are equipped with a convolutional neural network which learns to extract high level visual features which are then concatenated with a representation of the symbolic inputs.

A.2 Training Details

We provide important details about our training setup here.

We train all three agents with PPO [66], a popular deep reinforcement learning algorithm. We choose PPO for training our agents for a few reasons. This algorithm is commonly used because it is found to be empirically robust and effective in a wide range of tasks [19]. We train each of the three agent architectures for 15 hours, the equivalent of 10 million training timesteps, on at least 3 different random seeds. We trained all agents using Tensorflow 2.3 [1] and the OpenAI Baselines PPO2 implementation [17] with a distributed sampler. For a full list of training hyperparameters used in all agent versions, please refer to Table 6. We found this set to perform best on preliminary experiments.

To effectively train agents in a complex video game setting, we use a distributed approach leveraging an in-house sample collection framework and Azure cloud resources. Training samples are collected from a scaleset of 20 low priority GPU virtual machines (Azure NV6), each running 3 video game instances. The samples are then sent to one training head node, a CPU-only Azure E32s memory-optimized virtual machine.

B BEHAVIORAL STUDY DETAILS

In this section, we include additional details about our MTurk behavioral study.

B.1 Full Instructions

We detail the full instructions included in the MTurk study here.

> We are conducting a survey on navigation in video games for a research project. Please read the **Description** and **Requirements**, and then select the link below to complete the survey. At the end of the survey, you will receive a code to paste into the box below to receive credit for taking our survey.
>
> **Description**:
>
> - **Overview**: The survey is anonymous and includes a required consent form, comprehension check, some background info, and 6 video sections with 3 questions each. All questions are marked *required.
> - **Time required**: about **30 minutes**.
> - **Compensation**: you will receive a fixed compensation of **$6.50** for completing the task, with potential for a **$1 bonus** for a high-quality response. For example, copy/pasting answers, or responses that are not specific to the videos on each page, will not get the bonus.
> - The MTurk HIT has a 1-hour duration. It will **not** allow you to submit after 1-hour has passed (*remember to submit or return HITs within 1-hour so you don't time out!*)
> - If you start the task but change your mind, you may terminate your participation at any time and **return the HIT** within 1-hour, but you will **not** be paid for returned HITs or partial completions.
>
> **Requirements**:
>
> - You must complete all the questions.
> - You must not have previously completed a HIT called "Navigation Turing Test (NTT)". Repeat participants are ineligible and will not be paid.
> - You cannot participate from tablets or mobile phones.
>
> **Make sure to leave this window open as you complete the survey.** When you are finished, you will return to this page to paste the code into the box.

B.2 Data Collected

In addition to the consent, familiarity, and HNTT questions, we collected the following data: timing data broken down by page and the order in which the trials were presented to each participant.

C CODING DETAILS

We include the coding guide agreed upon and used by both annotators when annotating their responses. Figure 6 shows a screenshot of the guide.

Goals of Codes

- Independence
- Separation

General Protocol

- Ideally, we want nice separation between codes. We also don't want the codes to be highly dependent. For example, we don't want every instance of goal directed to also be coded as collision avoidance. Therefore, if something can be coded in two ways, it is better to code it as the more *specific* instantiation.
- If we think collision avoidance, only code as collision avoidance and not environment receptivity.
- If we think smoothness, only code as smoothness and not mechanistic. This is true even if the response calls out stereotyped bot behavior.

Codes

1. **Smoothness of movement:** refers to the quality of the agent's navigation or camera movement. This code considers both more immediate jerky actions and more temporally-extended zig-zagging behavior.
    a. Key words:
        i. Smooth (+)
        ii. Janky (-)
        iii. Jerky (-)
        iv. Straight (+)
        v. Swerve (-)
        vi. Moving camera (-)
        vii. Steady camera (+)
        viii. Fluid (+)
        ix. Rigid (-)
        x. Zig-zag (-)
        xi. Clunky (-)
2. **Goal directed** looks at how intentional the agent's behavior seems to be. Participants describing behavior that pertains to a perceived goal, even if that goal is not the main one in the video, is included in this code. For example, if they refer to intentional exploration, we code it as goal directed.
    a. Key words:
        i. "Knew where they were going" (+)
        ii. Moving with intention (+)
        iii. Navigating with purpose (+)
        iv. Moving with focus (+)
        v. Making decisions (+)
3. **Collision avoidance** can be considered an instantiation of goal-directed behavior, but we set up a separate code because we believe it is an important feature in its own right. To preserve the independence of codes, behaviors that are coded as collision avoidance should not be coded as any other code --- except when e.g., the participant calls out another goal-directed behavior in the same response. An example of collision avoidance would be an agent trying not to "crash" into other objects. An example of a response that would not be considered collision avoidance is if an agent tries to collect a power-up.
    a. Key words:
        i. Collide (-)
        ii. Avoid (+)
        iii. Crash (-)
        iv. Run into obstacle (-)
4. **Environment receptivity** aims to capture the agent's relationship with the game environment. In a real-world setting, this might look like a person walking on a path instead of the grass or being responsive to the environment (such as a walk sign). Dynamic. The focus here is on the *intention* of the behavior. This code also aims to capture abiding by *norms* or (potentially unspoken) conventions that people may only have a sense for.
    a. Keywords:
        i. Explore (+)
        ii. Stay on the path (+)
        iii. Collecting power-ups (+)
        iv. Already knows everything about the environment (-)
        v. Take shortcut (+)
        vi. Seeing through walls (-)
        vii. Took shortcut (+)
5. **Non-mechanistic** is a more nebulous code that tries to capture pre-conceived notions of human imperfection. An example of mechanistic behavior is one identified as being too perfect. This code also captures any references to mistakes or error correction.
    a. Keywords:
        i. Too perfect (-)
        ii. Makes mistakes (+)
        iii. Precise (-)
        iv. Micro-corrections (-)
        v. Random (+)
        vi. Overcorrect (-)
6. **Intuition** refers to feelings that a behavior was more human-like, without sufficient specific justification. We include this code to capture instances where participants can identify what they believe is more human-like behavior but struggle to express it.
    a. Keywords:
        i. Natural
        ii. Feeling
        iii. Seems to be
        iv. Normal
        v. Realistic
7. **Self-reference** is a code to capture when judges relate the agent behavior to their own play. This could look like participants mentioning that they would feel ill if they played this way or mentioning that they typically collect all power-ups when playing.
    a. Keywords:
        i. "Like how I play"

Discard

1. If the participant seems to randomly assign human-like or not. This may look like saying, "I don't know" or "I can't tell". In contrast, if they say that the response is based off of feeling, we categorize it as intuition.
2. If the free-form response is in conflict with the actual response.

Figure 6: Coding guide used by the annotators. The guide includes a description of the codes, the general protocol, and the discard guidelines.