How Humans Perceive Human-like Behavior in Video Game Navigation

Evelyn Zuniga*†
Microsoft Research, Cambridge
Cambridge, United Kingdom
t-ezuniga@microsoft.com

Stephanie Milani*†
Carnegie Mellon University
Pittsburgh, USA
smilani@cs.cmu.edu

Guy Leroy*
Microsoft Research, Cambridge
Cambridge, United Kingdom
t-gleroy@microsoft.com

Jaroslaw Rzepecki†
Monumo
Cambridge, United Kingdom

Raluca Geogescu
Microsoft Research, Cambridge
Cambridge, United Kingdom

Ida Momennejad
Microsoft Research, New York
New York City, USA

Dave Bignell
Microsoft Research, Cambridge
Cambridge, United Kingdom

Mingfei Sun
Microsoft Research, Cambridge, and University of Oxford
Cambridge and Oxford, United Kingdom

Alison Shaw
Ninja Theory
Cambridge, United Kingdom

Gavin Costello
Ninja Theory
Cambridge, United Kingdom

Mikhail Jacob†
Resolution Games
Stockholm, Sweden

Sam Devlin
Microsoft Research, Cambridge
Cambridge, United Kingdom

Katja Hofmann
Microsoft Research, Cambridge
Cambridge, United Kingdom

ABSTRACT

The goal of this paper is to understand how people assess human-likeness in human- and AI-generated behavior. To this end, we present a qualitative study of hundreds of crowd-sourced assessments of human-likeness of behavior in a 3D video game navigation task. In particular, we focus on an AI agent that has passed a Turing Test, in the sense that human judges were not able to reliably distinguish between videos of a human and AI agent navigating on a quantitative level. Our insights shine a light on the characteristics that people consider as human-like. Understanding these characteristics is a key first step for improving AI agents in the future.

*Authors contributed equally to this research.
†Work done while at Microsoft Research, Cambridge.

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04…$15.00
https://doi.org/10.1145/3491101.3519735

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; • **Computing methodologies** → **Cognitive science**.

KEYWORDS

Human-AI Interaction, Human-subject Study, Believable AI

ACM Reference Format:
Evelyn Zuniga, Stephanie Milani, Guy Leroy, Jaroslaw Rzepecki, Raluca Geogescu, Ida Momennejad, Dave Bignell, Mingfei Sun, Alison Shaw, Gavin Costello, Mikhail Jacob, Sam Devlin, and Katja Hofmann. 2022. How Humans Perceive Human-like Behavior in Video Game Navigation. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3491101.3519735

1 INTRODUCTION

A core goal of artificial intelligence research is to build artificially intelligent (AI) agents (computational characters or entities) capable of learning and displaying complex human-like behavior [4]. Achieving human-like behavior is a milestone towards future agents that can flexibly collaborate with people in shared human-AI environments. Without the ability to perform tasks in a human-like or human-compatible manner, achieving high skill alone is likely insufficient. For example, in a shared human-AI driving setting, AI drivers must behave sufficiently human-like for human

drivers to interpret, anticipate, and act in their presence [13]. Furthermore, researchers have identified human-AI compatibility as a key milestone towards a wide range of robotics applications [22]. In a nearer-term future application, achieving human-like behavior of artificial agents in video games promises new engaging game experiences to delight billions of players [7]. An open challenge in developing such human-like agents is understanding what characteristics are considered human-like. Understanding the behaviors that contribute to people's perception of human likeness is a foundational first step towards achieving general human-like behavior of artificial agents in human settings.

In this work, we contribute to the goal of identifying human-like behaviors. Specifically, we focus on a 3D video game navigation setting. We leverage the recently-proposed Human Navigation Turing Test (HNTT) [9], in which participants distinguished between human and AI-generated navigation behavior in a video game and explained their decisions. We run a behavioral study on Amazon Mechanical Turk (MTurk) and analyze a data set of 426 free-form responses to the HNTT to gain insights into the behaviors that human judges perceive as characteristic of AI or people.

We find clear differences in the way that human judges perceive human and AI behavior. We also find preliminary indicators that expectations about AI behavior may be more stable than those of human behavior. We find more nuance when considering the concept of "human play": in particular, human judges struggle to identify what constitutes human play when presented with a more human-like AI.

2 RELATED WORK

This work focuses on studying human-like behavior in the context of navigation. Navigation is a well-studied task in biological settings [8] and fundamental to biological intelligence embodied within the real world [12, 19]. Navigation is also a crucial in many video games and a key area of interest for game developers [1]. Prior work has focused on enabling robots to demonstrate this ability. The problem of Simultaneous Localization and Mapping (SLAM) was established formally by Borthwick and Durrant-Whyte [3], building on an earlier line of work [10]. When applied to simulated environments (e.g., video games) the problem is simplified, as we can assume the precise location of the agent and a complete map of the environment is known. These simplifications enable the use of A* search on nav-meshes as a common approach to navigation in 3D games for over two decades [25]. These approaches provide a range of algorithms specific to the challenges of 3D navigation. However, we desire methods that can be applied to a broader set of sequential decision-making problems.

Reinforcement learning (RL) [26] provides a more generally applicable set of algorithms for learning to control agents in settings including (but not limited to) 3D navigation problems in modern game environments [1]. RL trains an agent to perform a task by learning to maximize a (usually) hand-crafted reward, or score, that tells the agent how well it is doing on that task. Crafting this reward is referred to as *reward shaping* [27]. Although agents can learn effective navigation, they make no consideration for the *style* with which they act [1]. If these approaches are to be adopted in commercial game development, practitioners have firmly asserted that controlling style is essential [16]. As an extreme example, RL approaches that have recently defeated world champion human players at modern games demonstrated unusual behaviors that made collaborative play between human and AI in mixed teams far less successful [2].

Prior approaches [1, 2] are evaluated by quantitative measures and not on how humans may perceive them. We argue (and it has recently been demonstrated [24]) that for an agent to collaborate well with a human, how that agent's behavior is perceived by humans is of critical importance. Early efforts to measure how people perceive agent behavior in video games [14] struggled to differentiate between human and agent play styles. More generally, evaluating human perception of artificially intelligent systems varies across disciplines and no central benchmark exists today. In robotics, promising quantitative approaches have been proposed, such as the human-likeness index for robot path evaluation [11]. However, this index does not account for human perception of the computed paths. Kahn et al. [17] proposed a set of psychological benchmarks for humanoid robot evaluation but did not validate them in an experimental setting. Closer to our work, Kim et al. [18] studied players' evaluations of human-likeness of AI bots in StarCraft using a scoring framework and evaluation of short-text responses. The paper analyzes top-performing AI bots but does not consider AI *designed* to exhibit human-like behaviors. More recently, Devlin et al. [9] proposed the Human Navigation Turing Test (HNTT) as an open challenge in which to learn human-like behavior. We leverage this experimental setup, but propose and perform a deeper qualitative evaluation of human assessments of AI and human behavior.

3 METHOD

Here we discuss our methods, starting from a high-level overview of the relevant components. We focus on an established 3D navigation task, and summarize this task, as well as collection of human player data on this task in Section 3.1. Contrasted with human navigation data is our agent-generated navigation data; we detail the agent architectures and data collection protocol in Supplemental Research Methods B.1. Next, we turn to the key focus of the present study: collecting human assessments of human-likeness and free form justifications of these assessments (Section 3.2). Finally, we detail our data analysis approach in Sections 3.3 (quantitative analysis) and 3.4 (qualitative analysis).

3.1 Navigation Task and Data

We focus on the navigation task introduced by Devlin et al. [9], illustrated in Figure 1. A human player or AI agent completing the task starts in the area visible at the bottom of the mini-map. They are tasked to navigate to one of the 16 goal locations, selected at random at the start of each trial. Human players experience the task from a third-person perspective, as shown in Figure 1. They also have access to the mini map indicating the current player position and goal location. In the third-person view, the goal location is visible as light blue containers once in view (e.g. at the back of the screenshot).

3.1.1 Human Navigation Data. Building on the established navigation task [9] allows us to use the human navigation data and

Figure 1: Navigation task as experienced by human players (screenshot, left), and detail of the mini map of the game level (right). Screenshot is not representative of the final game visuals.

videos from their work.[1] The human player data was collected from individuals who were familiar with the game and the map. For additional details on sampling human player data and post-processing steps, see Supplemental Research Methods A.1 and Devlin et al. [9].

3.1.2 Agent Navigation Data. Our method asks human judges for relative comparisons between pairs of human and AI navigation behavior. We consider three AI agents, designed to reflect state-of-the-art machine learning approaches to navigation and provide a spectrum of human-likeness. Our agents closely align with the *symbolic* and *hybrid* agents proposed by Devlin et al. [9]. These agents are differentiated based on how they observe the navigation environment; see Supplemental Research Methods A.2 for more details. These agents successfully learned to navigate; however, human judges could identify their learned behavior as decidedly non-human [9]. For our study, we use the *symbolic* and *hybrid* agents; however, we augment these agents based on insights into behaviors that they exhibited that people likely identified as being non-human. To even more closely align with these expectations, we design and introduce a novel *reward shaping* agent.

We visually inspected the learned navigation behavior of the *symbolic* and *hybrid* baseline agents and isolated three classes of problematic behavior. Agents would: (1) swing camera angles wildly or make sudden turns, (2) collide frequently with walls, and (3) sometimes move more slowly than ideal. To correct these behaviors, we use reward shaping, a simple yet effective approach to achieving desired qualities in agent behavior [21]. We design the reward signal as follows. To combat the problematic behavior, we introduce: (1) a camera angle difference penalty for swift camera angle changes over a set 0.15 difference threshold value, (2) a 0.05 penalty for any wall collisions, and (3) a penalty of 0.01 if the distance traveled between steps is lower than a set 220 threshold value specific to the environment.

Our novel *reward shaping* agent learns to navigate using reward explicitly designed to reduce human-perceived differences between human and AI agent behavior. This agent extends the baseline *hybrid* agent with additional reward components and a finer-grained action space. We introduce the reward components to explicitly encourage learning human-like behavior and the finer-grained action space to enable smoother control and avoid abrupt turns when combined with reward shaping. We extend the action space to 14 discrete actions, providing 3 additional degrees of turning left/right, compared to the baseline agents. The updated list of degrees for this agent is (±0.2, ±0.4, ±0.5, ±0.6, ±0.8 and ±1.0). Human judges cannot reliably distinguish this agent from human behavior using a Navigation Turing Test (which we show in Section 4), making the qualitative insights on perceived differences particularly valuable.

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, and confirm training has converged by inspecting training curves (more details in Supplemental Research Methods B.1). For each agent we select the latest training model checkpoint. Finally, to generate the agent navigation data, we roll out 100 navigation runs, then sample from these to match the goal locations of the human data (see Supplemental Research Methods A.1).

[1]Data use under MSR-LA license. License details can be found in the original authors' GitHub: https://github.com/microsoft/NTT.

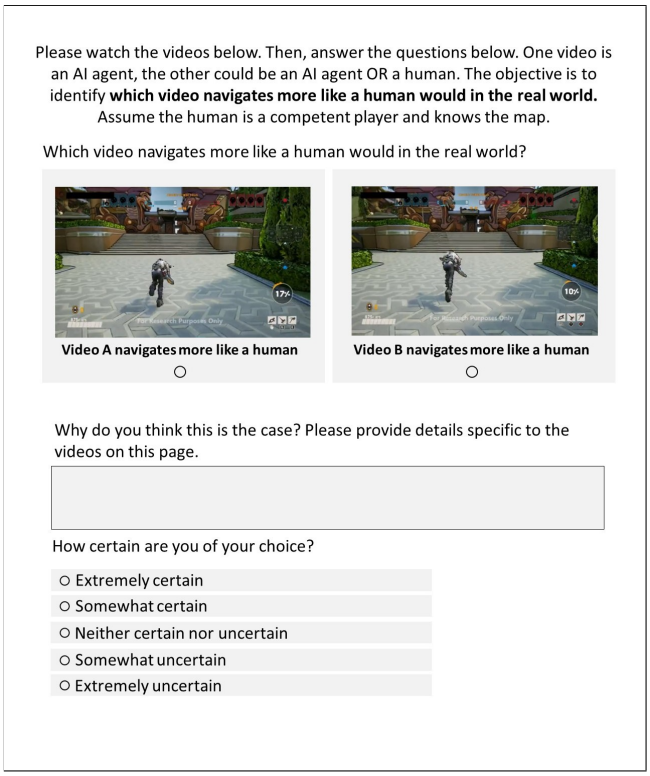

Figure 2: Example of one HNTT trial. Screenshots are not representative of the final game play or visuals.

3.2 Data Collection: Assessing Human-Likeness

Our data collection approach follows the setup of the Human Navigation Turing Test (HNTT) behavioral study design [9]. Each human judge is asked to complete a survey comprising of 10 Turing Test trials (Figure 2). In each trial, the judge watches two side-by-side video stimuli of humans or AI agents completing the navigation task and answer 3 questions. First, a binary choice: "Which video

navigates more like a human would in the real world?". Then, a free-form response: "Why do you think this is the case? Please provide details specific to the videos on this page." Finally, a multiple choice: "How certain are you of your choice?", where the options range from "extremely uncertain" to "extremely certain" on a 5-point Likert scale. Supplemental Research Methods B.4 contains screenshots of the survey questions.

We completed 3 studies with a within-subject design, where all human judges viewed the same 10 trials per study, and the trials were presented in a randomized order per judge. In each survey, 6 trials were human-vs-agent comparisons, and 4 trials were agent-vs-agent comparisons. We exclude the agent-vs-agent trials in this work. Study 1 tested human vs. *hybrid* agent (50 subjects), Study 2 tested human vs. *symbolic* agent (50 subjects), and Study 3 tested human vs. *reward shaping* agent (92 subjects) (total: 192 completed surveys).

The earlier study [9] relied on a smaller pool of locally-recruited assessors; however, our judgments were collected on the Amazon Mechanical Turk (MTurk) crowd-sourcing platform [20]. MTurk is widely used for data collection as it provides the benefit of scalability, as long as appropriate steps for quality control are implemented [15]. The MTurk participant requirements were: location is United States, age is 18 or older, 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 in each response to distinguish high-quality versus low-quality or suspected bot responses; for example, responses with high instances of typos, copy/pasted answers, or nonsensical wording were identified as low-quality and excluded from analysis. 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. Details of the study, as well as description of any potential participant risks, were included in the consent form. We include the full text of the MTurk HIT instructions in Supplemental Research Methods B.5.

3.3 Quantitative Data Analysis: Navigation Turing Test

The quantitative analysis of our collected human judgment data follows the methodology proposed by Devlin et al. [9] as much as possible, to validate our change in crowdsourcing setup (results reported in Supplemental Research Methods B.2) and determine the degree to which human judges are able to distinguish between AI and human behaviors (Section 4.1).

A key aspect missing from prior work is a firm criterion to establish whether the Human Navigation Turing Test has been passed by an agent, requiring us to establish new methodology. We propose a methodology that formalizes the following 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 that is 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 we determine the agent passes the HNTT.

3.4 Qualitative Data Analysis: Free-form Responses

To analyze the free-form responses from our HNTT, we chose a sub-sample of the responses from Study 1 (the *hybrid* agent) and Study 3 (the *reward shaping* agent) in order to include one agent that doesn't pass the HNTT and one that does. From each study we sub-sampled 3 free-form responses per subject, which resulted in 426 individual responses for analysis. Responses were randomly sub-sampled and shuffled to minimize bias.

We followed a pair coding approach: two human annotators reviewed the free-form responses and assigned one or more codes that captured behaviors characteristic of AI or Human as perceived by human judges. The annotators reviewed an agreement sample (60 free-form responses) to converge on a list of 18 codes. When constructing the codes, we referenced worked examples of reflexive thematic analysis [5] and allowed for more meaningful interpretation. We provide the resulting list of codes and their definitions in Table 1 for reference. The resulting list of codes (with definitions in parentheses) were: **Frequent** or **Infrequent camera movement** (how often the camera is moved or zoomed), **Smooth** or **Jerky physical movement** (relating to movements, trajectory, or pathing), **Logical** or **Illogical reasoning** (where the character does things that do or do not make sense), **Related to human play** or **Non-related to human play** (relating a behavior to their own or human gameplay, often used in a non-descriptive manner, e.g., "the navigation has movements which I think can be done by a human only"), **Wall avoidance** or **Wall hit** (avoiding or running into or against walls), **Goal direct** or **Goal indirect** (related to goal orientation), **Frequent** or **Infrequent mistakes** (how often the character makes mistakes or corrections), **Object avoidance** or **Object hit** (avoiding, or running into or against objects), **Nonsense** (Response is not interpretable), and **Other** (any other characteristic that was more rarely mentioned, e.g., "the character jumps more often"). Table 2 illustrates an example of a coded response.

With these codes, the annotators coded the agreement sample. We quantified inter-annotator agreement on this sample by computing a binary Cohen's Kappa κ [6] for each code. Averaging over the codes yielded $\kappa = 0.511$. Each annotator then coded 55% of the data sub-sample (234 responses each) to compare agreement on the overlapping 10%. For each response, the annotators assigned one or more codes under two categories: Human Codes (average $\kappa = 0.62$) and AI Codes (average $\kappa = 0.54$).

4 RESULTS

We present two sets of results. First, we show that our reward-shaping agent passes the HNTT (Section 4.1). Second, we highlight characteristic behaviors and key differences in how human judges

Annotation Codes	Definition
Frequent camera movement, infrequent camera movement	How often the camera is moved, swung, or zoomed
Smooth physical movement, Jerky physical movement	Characteristics relating to movements, trajectory, or pathing
Logical reasoning, Illogical reasoning	Where the character does things that do or do not make sense
Related to human play, Non-related to human play	Where participants relate a behavior to their own or human gameplay
Wall avoidance, Wall hit	Avoiding, or running into or against walls
Goal direct, Goal indirect	Codes related to goal orientation
Frequent mistakes, Infrequent mistakes	How often the character makes mistakes or "corrections"
Object avoidance, Object hit	Avoiding, or running into or against objects
Nonsense	Response is not interpretable
Other	Any other characteristic

Table 1: Annotation code definitions.

Subject Response	Free-Form Response	Human Codes	AI Codes
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.	Smooth physical movement; Goal direct	Object hit; Goal indirect

Table 2: Example coded response to the question, "Which video navigates more like a human would in the real world?". Highlights illustrate annotation process.

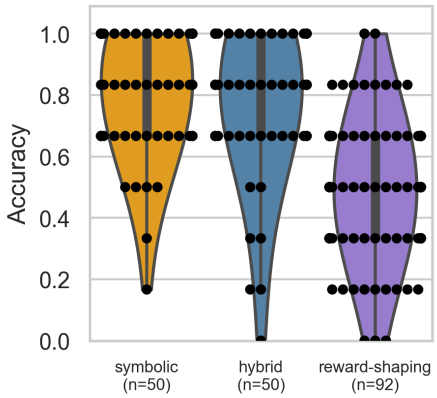

Figure 3: Accuracy of human judges' assessment of human-likeness. Human judges assessed the reward shaping agent as most human-like.

perceive AI vs human players (Section 4.2). In particular, we investigate these characteristic behaviors when the AI agent does and does not pass the HNTT.

4.1 Human Navigation Turing Test

With the experimental setup described in Section 3.2, we now evaluate the human-likeness of our agents as determined by human judges. Figure 3 shows the accuracy (agreement with ground truth) of human judges when assessing the human-likeness of our *reward shaping* agent compared to baselines. On average, participants achieve a significantly lower accuracy (mean=0.49, std=0.22) when assessing our *reward shaping* agent against human players, in comparison to the *symbolic* (mean=0.80, std=0.20, U=710.0, p=0.000) and *hybrid* (mean=0.76, std=0.24,U=863.5, p=0.000) agents. This indicates that our approach is judged the most human-like.

Next, we turn to the question of whether any of the agents included in our study passes the HNTT based on the criterion defined in Section 3.3. According to our criterion of passing the HNTT, we find that the *symbolic* and *hybrid* baseline agents fail the HNTT, whereas the *reward shaping* agent passes this test of human-likeness. Median accuracy has a 95% confidence interval that includes 0.5 (chance-level agreement), suggesting that human 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]). The *symbolic* and *hybrid* baseline agents from Devlin et al. [9] do not pass the HNTT according to this criterion. We obtain 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 human judges can distinguish them from humans at significantly higher than chance level.

4.2 Qualitative Analysis

With our qualitative analysis, we seek to answer the following research questions:

(1) Are there key differences between how people characterize behavior that they believe is generated by an AI and that which they believe is generated by a human?
(2) What mistakes do people make when characterizing AI vs. human behavior?
(3) Are there key differences between how people characterize AI vs. human behavior when we compare between AI that does and does not pass the HNTT?

4.2.1 Differences in Human vs. AI Behavior. We find that there are indeed differences between how people characterize AI vs. human behavior. We plot the counts of codes used to describe the behavior

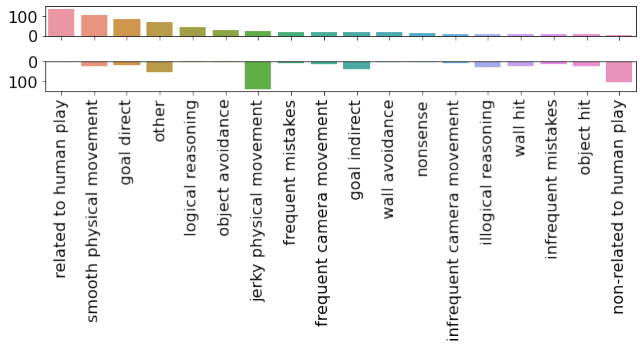

Figure 4: Most common human (top) and AI (bottom) codes.

in Figure 4. Most commonly, people use the following labels to describe human-like behavior (in decreasing order of frequency): related to human play, smooth physical movement, goal direct, other, and logical reasoning. In contrast, people most commonly use the following labels to describe AI behavior (in decreasing order of frequency): jerky physical movement, non-related to human play, other, goal indirect, and illogical reasoning. These findings suggest that there are clear differences in how people characterize human vs. AI behavior. In fact, the labels that are foils to one another occur most frequently (e.g., smooth physical movement and jerky physical movement), which indicates that people generally rely on the same high-level qualities to distinguish between AI- and human-generated behavior.

4.2.2 Mistakes. We further decompose the responses based on whether the human assessor correctly identified the agent as being human or not (Figures 5a and 5b). When assessors correctly identify human-generated behavior (Figure 5a top), they use the following identifiers most frequently (in descending order): smooth physical movement, related to human play, goal direct, other, and logical reasoning. In contrast, when they are mistaken, they most commonly use: related to human play, smooth physical movement, goal direct, other, and logical reasoning. One assessor (incorrectly) noted, "Video B has less accurate and precise moves. Thus, it reflects more human-like moves."

When assessors correctly identify AI-generated behavior (Figure 5b top), they most commonly refer to jerky physical movement, non-related to human play, other, goal indirect, and object hit. In contrast, when assessors falsely identify AI-generated behavior, they most commonly refer to jerky physical movement, non-related to human play, other, wall hit, and goal direct. Overall, the ranking of these is more stable than for human codes, suggesting that human assessors may have a more stable notion of characteristics that constitute AI behavior.

4.2.3 More Human-like AI. We now investigate whether assessors characterize behavior differently depending on how *human-like* an AI is, as defined by passing or failing the HNTT. We examine which codes are used to describe human play within both settings: passing or failing the HNTT. Figure 6a depicts the most common human codes in both of these settings, and Figure 6b shows the most common AI codes in both of these settings.

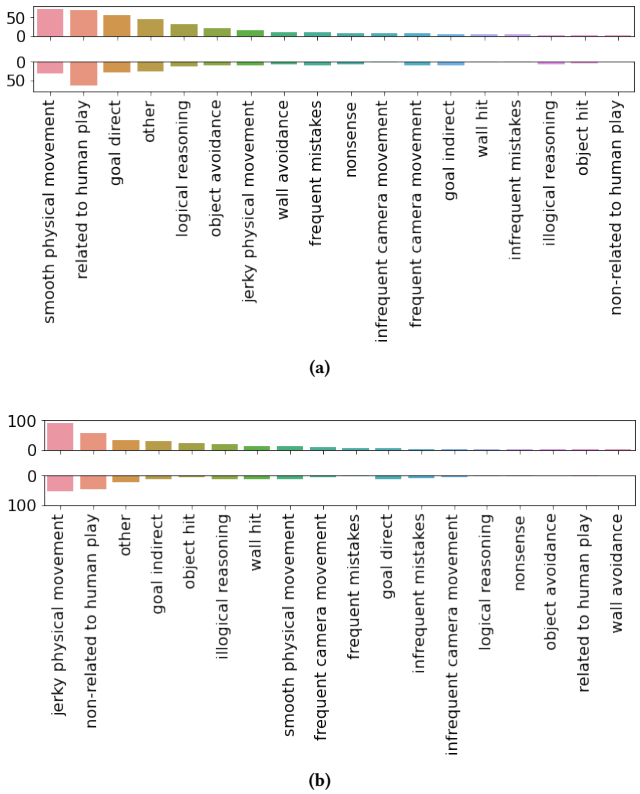

Figure 5: (a) Most commonly-used codes for correct human labels (top) and incorrect human labels (bottom). (b) Most commonly-used codes for correct AI labels (top) and incorrect AI labels (bottom).

We examine the codes that judges use to describe human-like behavior in both of these settings. When the agent passes the HNTT, the judges more frequently mention related to human play, smooth physical movement, and goal direct, compared to when the agent fails the HNTT. Interestingly, judges mention jerky physical movement more when the agent passes the HNTT compared to when the agent fails the HNTT. When describing AI behavior, judges mention jerky physical movement, non-related to human play, and smooth physical movement when the agent fails the HNTT compared to when the agent passes the HNTT. Taken together, these results suggest that assessors may have a notion of what constitutes human play but struggle to describe specific clues when presented with a more human-like AI.

These codes are also quite unstable: the relative frequencies of the use of these codes varies across studies. In contrast, the codes used to describe AI play within both settings appear more stable. This suggests that expectations about AI behavior may be more stable than those for human behavior. Rephrased, people may have a more rigid understanding of what constitutes AI behavior.

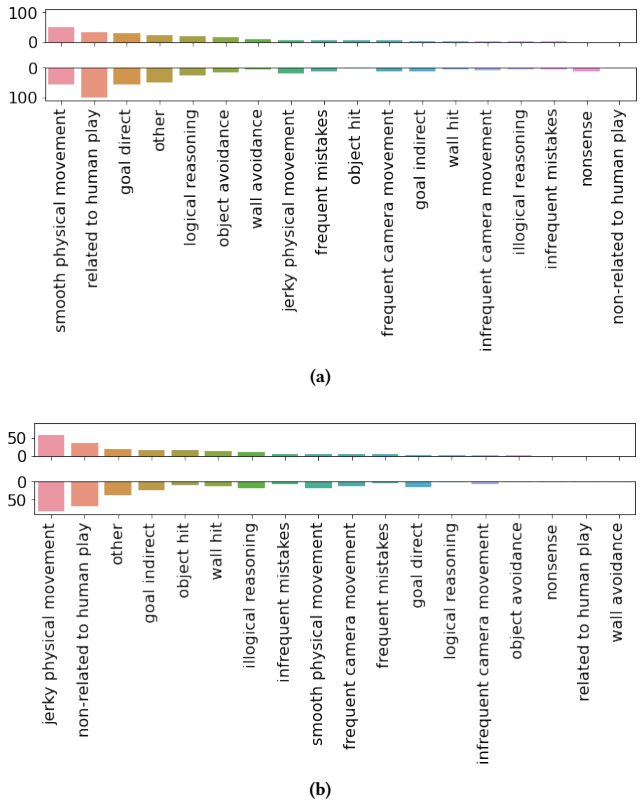

Figure 6: (a) Most common Human codes for agent that fails HNTT (top) and agent that passes HNTT (bottom). (b) Most common AI codes for agent that fails HNTT (top) and agent that passes HNTT (bottom).

5 CONCLUSION

In this work we present a behavioral study based on the Human Navigation Turing Test (HNTT), comprising of hundreds of human assessments of human and AI behaviors in a 3D navigation task. We present an AI agent that passes the HNTT in contrast to an AI agent that fails the HNTT, and run a qualitative assessment of free-form responses associated with this comparison. Our findings show clear differences in how human assessors perceive AI vs. human behavior. At the highest level, the behaviors "jerky physical movement" and "non-related to human play" were more frequently associated with AI, whereas "related to human play" and "smooth physical movement" were more frequently associated with human behavior. Breaking this down further we evaluate assessor mistakes. When assessors incorrectly identify the AI agent as a human, the top two behaviors are "related to human play" and "smooth physical movement", which align with the strongest associated human behaviors. When they incorrectly identify a human as an AI agent, the top two behaviors are "jerky physical movement" and "non-related to human play", which align with the strongest associated AI behaviors. Finally, we compare an agent that fails the HNTT (*hybrid*) against a more human-like agent (*reward shaping*) that passes the HNTT. We find some nuance in the way assessors consider "human play" in this comparison, suggesting that assessors have a notion of human play but struggle to identify this more precisely. Our findings suggest that perceived characteristics of AI behavior are more stable than those of human behavior. One limitation of our work is the analysis is based on one benchmark; results may differ for different scenarios. Nevertheless, our work opens up exciting opportunities for deeper understanding of how humans assess human-likeness.

REFERENCES

[1] 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.

[2] 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.

[3] Stephen Borthwick and Hugh Durrant-Whyte. 1994. Dynamic localisation of autonomous guided vehicles. In *Proceedings of 1994 IEEE International Conference on MFI'94. Multisensor Fusion and Integration for Intelligent Systems*. IEEE, Piscataway, NJ, 92–97.

[4] 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.

[5] David Byrne. 2021. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity* 1 (2021), 1–22.

[6] 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

[7] 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.

[8] William de Cothi, Nils Nyberg, Eva-Maria Griesbauer, Carole Ghanamé, Fiona Zisch, Julie Lefort, Lydia Fletcher, Charlotte Newton, Sophie Renaudineau, Daniel Bendor, Roddy Grieves, Éléonore Duvelle, Caswell Barry, and Hugo J. Spiers. 2020. Predictive Maps in Rats and Humans for Spatial Navigation. *bioRxiv preprint: 2020.09.26.314815* 1 (2020), 1–44.

[9] 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.

[10] Hugh Durrant-Whyte and Tim Bailey. 2006. Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine* 13, 2 (2006), 99–110.

[11] Néstor García, Jan Rosell, and Raúl Suárez. 2017. Motion Planning by Demonstration With Human-Likeness Evaluation for Dual-Arm Robots. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49, 11 (2017), 2298–2307.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Peter H. Kahn, Hiroshi Ishiguro, Batya Friedman, and Takayuki Kanda. 2006. What is a Human? - Toward Psychological Benchmarks in the Field of Human-Robot Interaction. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, Vol. 1. IEEE, Piscatway, New Jersey, 364–371. https://doi.org/10.1109/ROMAN.2006.314461

[18] 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

[19] John O'Keefe and Lynn Nadel. 1978. *The hippocampus as a cognitive map.* Oxford: Clarendon Press, New York City, New York.
[20] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419.
[21] 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.
[22] Matthias Scheutz, Paul Schermerhorn, James Kramer, and David Anderson. 2007. First steps toward natural human-like HRI. *Autonomous Robots* 22, 4 (2007), 411–423.
[23] 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.
[24] Ho Chit Siu, Jaime Peña, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chang, and Ross Allen. 2021. Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi. *Advances in Neural Information Processing Systems* 34 (2021), 28 pages.
[25] Greg Snook. 2000. Simplified 3D movement and pathfinding using navigation meshes. *Game programming gems* 1, 1 (2000), 288–304.
[26] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press, 1 Broadway, Cambridge, MA.
[27] Eric Wiewiora. 2010. *Reward Shaping.* Springer US, Boston, MA, 863–865. https://doi.org/10.1007/978-0-387-30164-8_731

A SUPPLEMENTAL RESEARCH METHODS

A.1 Human Navigation Data

Here we provide supplemental details that expand on Section 3.1, detailing sampling and post-processing. We sampled human player data from the sample published by Devlin et al. [9] with the aim to avoid potential biases. Potential biases could be introduced, e.g., in cases of noticeable differences between the human and AI data (e.g., systematic differences in video length or quality that are unrelated to the actual navigation behavior). Human videos were sampled from the 40 videos published under their "study 1" protocol. We excluded videos shorter than 10 seconds (as these were found to be too short to assess navigation quality in pilot studies). We matched goal locations with those of AI agent videos (see Section 3.1). We applied post-processing in line with the procedure of Devlin et al. [9]): This included masking any identifying information, adding a "For Research Purposes Only" watermark, and cutting out the last few seconds of the human videos (this was to correct an effect of the data collection process where the human players had to manually end their recording, artificially adding a few seconds at the end of the videos).

A.2 AI Agents and Navigation Data

This section provides supplementary information to the agent architectures discussed in Section 3.1, and summarizing the *symbolic* and *hybrid* agents introduced in Devlin et al. [9]. We first overview the high level approach shared by all three agents (the *symbolic* and *hybrid* baseline agents and our novel *reward shaping* agent), and then discuss individual agents in turn. Finally, we provide details on our distributed agent training process, which enables scalable training in complex video games.

A.2.1 Shared agent architecture. We follow a reinforcement learning (RL) approach [1], a popular machine learning approach that focuses on agents learning to interact with an environment through trial and error, which has been shown to lead to effective navigation in complex game settings. Following Devlin et al. [9], we use the popular RL algorithm Proximal Policy Optimization (PPO) [23], one of the most commonly current state-of-the-art approaches and one found to be empirically robust and effective in a wide range of tasks.

The key components needed to specify a RL method are the *observation space* (i.e., information perceived by the agent, its input), *action space* (i.e., how the agents affects the world, its output), and *reward signal* (i.e., the feedback the agent receives after trying a course of action, its learning signal). These are defined in turn for the three agents used in this work.

A.2.2 Symbolic agent. Our first baseline agent is the *symbolic* agent from Devlin et al. [9]. Its observation space consists of 6 symbolic inputs (relative angle and distance to goal, numerical visual frame depth average, player's current x,y,z coordinates). The agent's action space is represented by 8 discrete actions, allowing the agent to stand, move forward or turn left/right by a given set of discrete angles (±0.4, ±0.7, ±1.0), where ±1.0 represents a ±90°angle. The reward signal is designed to encourage progress towards, and successful navigation to the goal and consists of the following: a -0.01 per step penalty, a penalty of -1 for dying, an incremental

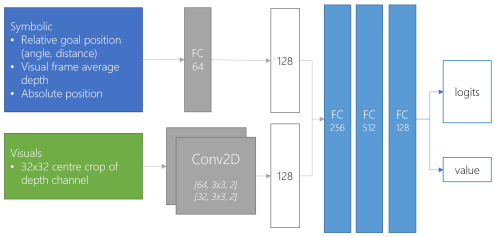

Figure 7: Architecture of the reward-shaping agent.

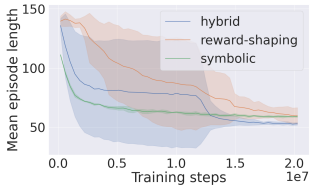

Figure 8: Hybrid, symbolic, and reward-shaping models converge to approximately optimal policies – we report episode length average and standard deviation (for reward-shaping, N=3; for hybrid and symbolic, N=4). All plots are smoothed with a rolling windows of 200.

reward for approaching the goal and a +1 reward for reaching the goal.

A.2.3 Hybrid agent. Our second baseline agent is the *hybrid* agent from Devlin et al. [9]. It differs from the *symbolic* agent only in its observation space. In addition to the symbolic observations, the agent receives a 32x32 cropped depth buffer visual input. The visuals present a third-person view of the agent in the environment. To process this additional visual channel, the hybrid agent is 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.2.4 Distributed agent training. 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 being collected from a scaleset of 20 low priority virtual machines (Azure NV6), each running 3 video game instances. The samples are then sent to one training head node, an Azure E32s virtual machine.

B SUPPLEMENTAL RESULTS

B.1 AI Agent Training

We provide additional results that confirm the effectiveness of our agent training. As discussed in Section 3.1, we considered three agents: the *symbolic* and *hybrid* baseline agents from Devlin et al. [9], and our proposed *reward shaping* agent, designed to minimize noticeable differences between agent and human behavior.

Figure 8 shows training curves for all three agents. We observe that the mean episode length decreases with training for all three agents, which shows that all three agents achieve high proficiency

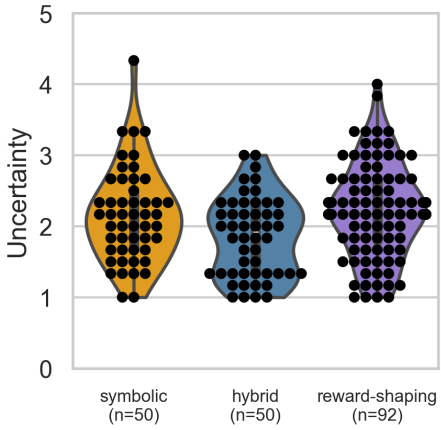

Figure 9: Uncertainty of human judges' assessment of human likeness.

on the task. Learning is slower for the reward-shaping agent, which can be accounted for by the need to learn to optimize a more complex reward signal. The lowest mean episode length at the end of training is obtained by the hybrid agent, however, as demonstrated by Devlin et al. [9], high task performance is not necessarily aligned with human-likeness.

Our results demonstrate that all three agents learn to effectively solve the navigation task, paving the way to our qualitative and quantitative study of the human-likeness of the learned behaviors.

B.2 Replication of HNTT Studies

We replicated Studies 1 and 2 from Devlin et al. [9] on the Amazon Mechanical Turk platform to verify that our switch to crowdsourcing did not significantly impact the validity of our findings. Table 3 summarizes the results. Our analysis shows no significant differences in judges accuracy in distinguishing human from agent behavior, validating our approach.

Interestingly, we see a small but statistically significant decrease in participants self-assessed level of uncertainty across both studies (Study 1 U=580, p=0.045; Study 2 U=478, p=0.003), indicating that crowdsourcing workers are more certain of their judgments. This is likely to result from the wider population from which participants are drawn in this setup. Factors impacting self-expressed uncertainty are an interesting topic for further study. For the purpose of this study, we increased sample sizes for our follow up studies to counter any potential increases in variance that could potentially result from a more diverse population or a higher level of uncertainty in judgments.

B.3 Human Navigation Turing Test: Uncertainty

Here we provide supplementary results on the self-reported uncertainty of judges assessing human-likeness (Figure 9). Human judges reported the highest average uncertainty when assessing our *reward shaping* agent (mean=2.21, std=0.66) in comparison to the *symbolic* (mean=2.15, std=0.65, U=2122.0, p=0.223) and *hybrid* (mean=1.85, std=0.56, U=1576.5, p= 0.001) agents. Note however

	Accuracy	Uncertainty	Hybrid vs Symbolic
Replication Study 1 (n=50)	0.76 (0.24)	1.85 (0.56)	0.72 (0.28)
ICML Study 1 (n=30)	0.84 (0.16)	2.08 (0.47)	0.78 (0.25)
Mann-Whitney U test	U=636.0, p=0.120	U=580.0, p=0.045	U=690.0, p=0.156
Replication Study 2 (n=50)	0.80 (0.20)	2.15 (0.65)	0.72 (0.32)
ICML Study 2 (n=30)	0.77 (0.16)	2.66 (0.84)	0.62 (0.30)
Mann-Whitney U test	U=653.5, p=0.160	U=478.0, p=0.003	U=626.5, p=0.052

Table 3: Comparing results between the studies from Devlin et al. [9] and our replication of those studies on MTurk. Results reported here are the mean (and standard deviation) of accuracy, uncertainty, and hybrid vs symbolic agent human-likeness judgments, including significance tests for each comparison.

that uncertainty is significantly higher only when compared to the *hybrid* agent. The statistical measures throughout are Mann-Whitney U tests with Bonferroni corrections to account for multiple comparisons.

B.4 HNTT Supplemental Material

The HNTT was conducted as an online survey with the following sections: an introduction page with a required consent form, a comprehension page including the questions in Figure 10a, a background page with brief details about the specified video game, a familiarity page with the questions in Figure 10a, and finally 10 HNTT trials with 3 questions each, as illustrated in Figure 10b. All questions were marked required. The survey format was kept the same as in Devlin et al. [9] to allow comparisons with their results.

B.5 Mechanical Turk Task Instructions

Below is the full text of the MTurk task instructions given to participants:

"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 10 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."

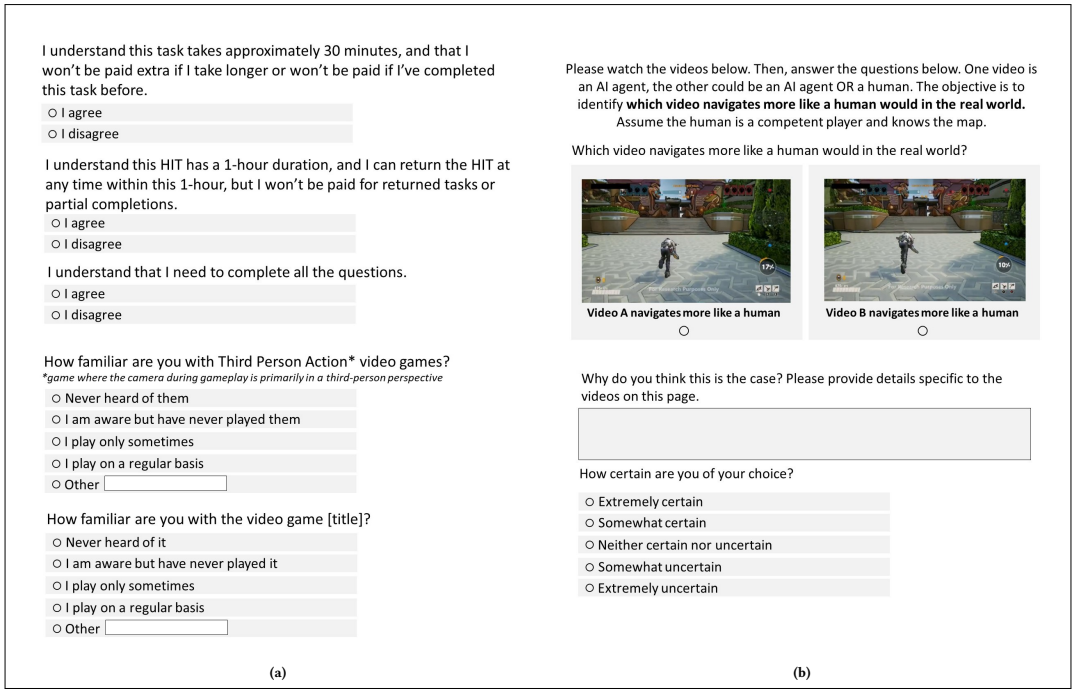

Figure 10: Screenshots of HNTT survey questions. (a) shows the comprehension and familiarity questions (asked once per participant). (b) shows one HNTT trial with 3 questions. Screenshots are not representative of actual game play or visuals.