Abstract

A key challenge on the path to developing agents that learn complex human-like behavior is the need to quickly and accurately quantify human-likeness. While human assessments of such behavior can be highly accurate, speed and scalability are limited. We address these limitations through a novel automated Navigation Turing Test (ANTT) that learns to predict human judgments of human-likeness. We demonstrate the effectiveness of our automated NTT on a navigation task in a complex 3D environment. We investigate six classification models to shed light on the types of architectures best suited to this task, and validate them against data collected through a human NTT. Our best models achieve high accuracy when distinguishing true human and agent behavior. At the same time, we show that predicting finer-grained human assessment of agents’ progress towards human-like behavior remains unsolved. Our work takes an important step towards agents that more effectively learn complex human-like behavior.

1. Introduction

Developing agents capable of learning complex human-like behaviors is a key goal of artificial intelligence research (Winston, 1984). Progress in this area would lead to new applications, from engaging video games (Soni & Hingston, 2008), to intuitive virtual assistants (Cowan et al., 2017) and robots (Scheutz et al., 2007). Yet, despite striking successes in training highly skilled agents (Vinyals et al., 2019), quantifying and making progress towards learning human-like behavior remains an open challenge.

We propose an automated Navigation Turing Test (automated NTT or ANTT), a novel approach to learn to quantify the degree to which behavior is perceived as human-like by humans. While human judges can be highly accurate when judging human-like behavior, their scalability and speed are limited. As a result, it is prohibitive to use human assessments in, e.g., iterations of training machine learning-based agents. We aim to remove this bottleneck, providing a scalable and accurate proxy of human assessments.

We demonstrate our NTT approach on the complex task of goal-directed navigation in 3D space. This task is highly relevant and practical because: (1) it has been studied from a wide range of perspectives in both biological (Banino et al., 2018; de Cothi, 2020) and artificial (Alonso et al., 2020; Chaplot et al., 2020) agents, so is relatively well understood; and (2) artificial approaches achieve a high level of task success (Alonso et al., 2020), so proficient agents can be obtained enabling us to focus on human-likeness. While the majority of research on learning 3D navigation has focused on task success, we show that a high success is not sufficient for human-like behavior.

Our key contributions compared to prior work (Sec. 2) are: (1) we introduce an automated NTT for assessing human-like behavior and investigate six representative architectures that capture principal design choices of this task (Sec. 3); (2) we devise methodology for validating our automated NTT against a human Navigation Turing Test (HNTT) and empirically demonstrate its effectiveness (Sec. 4); (3) our results show high agreement of ANTT with human judges when comparing human behavior to state-of-the-art reinforcement learning agents (Sec. 5). At the same time, we demonstrate that capturing finer gradations of human-likeness in artificial agents remains an open challenge. In addition to the potential to drive learning of more human-like behavior, our work opens the door to a deeper understanding of human decision making when assessing human-like behavior.

2. Related Work

The question of how to systematically assess whether an automated agent behaves in a human-like manner was compellingly addressed by Alan Turing (Turing, 1950). Initially called the “imitation game”, and now adapted and widely known as “Turing Test”, it asks a human judge to identify which of two respondents is human, and which is machine.
Turing’s initial thought experiment considered a setup where the judge could interact with each respondent through an appropriate natural language interface. Since its conception, the idea has been widely adapted to assess the degree of human-likeness in a range of scenarios, including to assess human-like play in video games (Yannakakis & Togelius, 2018). Closest to our focus, a Turing Test was used to assess bot behavior in the video game Unreal Tournament as part of the 2K BotPrize Contest (Hingston, 2009) and human-like play in the Mario AI Championship (Shaker et al., 2013).

While a Turing Test, if well designed, can provide strong ground truth assessments of whether humans judge behavior as human-like, several drawbacks exist. Most important from a machine learning perspective is the need to collect human judgments for every iteration of an agent, making the approach extremely costly. This makes it prohibitive for use in typical machine learning settings. Consequently, previous work has proposed automated proxies of human-likeness (Kirby et al., 2009; Karpov et al., 2013; de Cothi, 2020).

Domain specific approaches to detecting (non) human-like behavior automatically often rely on heuristics or rules. In navigation scenarios this can include methods for detecting mistakes such as “getting stuck” (Kirby et al., 2009; Karpov et al., 2013). Kirby et al. (2009) focus on socially acceptable robot navigation with experiments performed in simulated navigation tasks (e.g., passing people in a hallway). Human preferences were explicitly encoded as constraints (e.g., a preference for moving to the right to let someone pass), and learned behaviors were evaluated in simulation, by assessing how often a given preference was violated. Karpov et al. (2013) tackle the problem of “believable” bot navigation in the context of the BotPrize competition (Hingston, 2009).

Believable behavior is assessed using rules for detecting mistakes such as getting stuck in cyclic behavior, collisions with other entities, failing to move, or falling off a map.

A key drawback of rule-based automation is that they require high manual effort by domain experts to construct. In addition, due to this high effort, they are necessarily coarse and unable to capture nuanced aspects of human-likeness.

Addressing both the limitations of scalability of human Turing Tests and rule-based approaches, learned models can serve as a proxy measure of human-likeness. If such a model can learn accurate predictions of human-likeness given ground truth judgments, these approaches could be deployed rapidly and at scale to drive the development and training of more human-like agents. Closely aligned with this goal of our work is the recent development of imitation learning approaches that rely on a discriminator to learn to distinguish between learned imitation and demonstrated expert behavior (Ho & Ermon, 2016; Peng et al., 2018). However, while closely aligned with the goals of our work in principle, the current state of the art in this field has not yet reached the level required for empirically strong performance on complex tasks and human-like behavior, with many evaluated on relatively less complex behaviors and imitating other automated agents. Our work has the potential to pave the way towards learning complex human-like behavior, and provides important insights into the key design choices required for capturing notions of human-likeness. We include representative examples closest to this approach as baselines (model SYM-FF and VIS-FF, Tbl. 1).

Closely related to our work is the approach taken by de Cothi et al. (2020) to compare artificial agent behavior in a maze navigation task to human and rat behavior. They used a support vector machine classifier trained on a small set of expert designed features to distinguish between model-free, model-based, and successor-representation-based trained agents. The trained classifier was then applied to behavior of humans and rats to estimate which agent behavior was closest to that of biological agents (de Cothi et al., 2020; de Cothi, 2020). We base our approach on this setup, but extend it to take as input lower-level features (e.g., trajectories, images) to avoid the need for expert designed feature extraction and better align with the line of work on imitation learning that forms the second inspiration to our work.

While our approach makes no specific assumption about the type of task for which we aim to detect human-like behavior, we focus on the task of navigation in 3D space. This task is well suited as it has been widely studied from a range of perspectives, including neuroscience (Iaria et al., 2003; Banino et al., 2018; de Cothi, 2020) and machine learning (Alonso et al., 2020; Chaplot et al., 2020). In addition, recent years have seen rapid progress on machine learned agents that achieve high task success, allowing us to focus our attention on whether the learned successful behavior is indeed human-like. A driver of progress on learning to navigate has been the availability of simulated environments and evaluation frameworks. Anderson et al. (2018) survey recent progress and evaluation frameworks for 3D navigation. In line with the majority of work in this area, the focus is on evaluating task success. Recent work on learning to navigate with a machine learning focus includes improvements of sample efficiency, e.g., through unsupervised pretraining (Li et al., 2020) and object semantics (Chaplot et al., 2020). Here we leverage the work of Alonso et al. (2020), as it most closely aligns with our requirement for achieving high task success in a 3D game environment. At the intersection of machine learning and neuroscience, machine learned models have been proposed as models for studying biological behavior (Banino et al., 2018; de Cothi, 2020). As before, the primary evaluation of these focused on task success (Banino et al., 2018), with the notable exception of de Cothi et al. (2020) which forms the starting point of our work.
3. Method

In this work we approach the task of predicting human judgments of whether an observed behavior appears human-like as a classification task, as detailed in Section 3.1. Our key focus is the question of which type of input space (Sec. 3.2) and model (Sec. 3.3) can best capture human judgments.\(^1\)

3.1. Problem Setup

We formalize the task of assessing human-likeness as binary classification. We assume input data in the form of time-indexed trajectories \(x_{0:T}\) of length \(T\), where \(x_t \in \mathbb{R}^d\) denotes the \(d\)-dimensional observation at time step \(t\).

Our goal is to obtain a classifier that achieves high alignment with ground truth human judgments (detailed in Sec. 4.4). Ground truth takes the form \(y^* : x^{(1)} \succ x^{(2)}\) to indicate that a single human judged the trajectory represented by \(x^{(1)}\) to be more human-like than \(x^{(2)}\). We obtain repeat measurements from several human judges for each pair. We focus on two evaluation measures: (1) accuracy compared to human ground truth (aggregated by majority vote), and (2) Spearman correlation (a non-parametric rank correlation measure robust to outliers (Croux & Dehon, 2010)) compared to a ranking of agent trajectories by the proportion of judges that assessed them as human-like. Together, these capture coarse and fine-grained agreement between our classification models and human judgments of human-likeness.

In this work, we assume that human ground truth judgments are only available at test time. This is a realistic assumption given the resource intensive nature of this type of data. To address this challenge, we propose training our models with proxy labels, which are derived from known agent identity \(\hat{y}\).

Thus, our classifiers are trained to maximize accuracy when predicting whether a trajectory was truly generated by a human player or an artificial agent, using a binary cross entropy loss:

\[
L = -[\hat{y}_{H=1} \log(p_{H=1}) + (1-\hat{y}_{H=1}) \log(1-p_{H=1})],
\]

where we denote by \(\hat{y}_{H=1}\) the pseudo label indicating that trajectory was truly generated by a human player, and by \(p_{H=1}\) the probability that the current model assigns to the trajectory having been generated by a human player.

A key question is whether the resulting models can generalize to align well with the degree to which human judges actually perceive a trajectory as more or less human. A key design choice that we expect to influence this ability to capture human judgments is the input space the classifier operates on. We discuss these choices in the next section.

3.2. Choice of Input Space

Classifier input can be represented in a variety of ways. For example, presenting input in the form of video frames derived directly from the game footage, is likely to align well with the information available to human judges when they decide on whether a given behavior appears human-like. We hypothesize that observation spaces that are well aligned with the information available to human judges will result in higher accuracy on the test task of capturing human judgments of human-like behavior. However, more complex observations, such as video, may suffer from low amounts of training data, leading to a potential trade-off between model fidelity and sample efficiency.

We consider the following choices (samples in Fig. 1):

**SYM.** Symbolic observations provide a low-dimensional representation of each step along the agent trajectory. Here, we use absolute agent location in 3D space as it provides a minimal representation of agent state.

**VIS.** Visual observations provide images of size 320x200 of the actual game footage, rendered without overlays such as game score. This observation type is expected to be well aligned with the information available to a human judge when assessing whether a behavior appears human-like.

In our architectures, we explore feed-forward and recurrent models for encoding observations, as detailed in Section 3.3. In addition, we consider two compressed representations that abstract the respective observations in a single image:

**TD.** Top-down trajectory projection (a single 2D image per trajectory), obtained by projecting the symbolic representation (agent position) along the "up" direction (z-coordinate).

**BC.** Bar-code encodes visual observations (a single 2D image per trajectory video). It is obtained by collapsing

---

\(^1\)See Appendix A.1 for training details and hyperparameters.
Table 1. Overview of the ANTT model architectures investigated in this work. See Sections 3.2 and 3.3 for details.

<table>
<thead>
<tr>
<th>Name</th>
<th>Input (Sec. 3.2)</th>
<th>Encoder</th>
<th>Pre-Training</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYM-FF</td>
<td>Symbolic</td>
<td>Feedforward Network</td>
<td>[None]</td>
<td>Equivalent to typical discriminator architectures in adversarial imitation learning (Ho &amp; Ermon, 2016)</td>
</tr>
<tr>
<td>SYM-GRU</td>
<td>Symbolic</td>
<td>GRU (Cho et al., 2014)</td>
<td>[None]</td>
<td></td>
</tr>
<tr>
<td>VIS-FF</td>
<td>Visual</td>
<td>VGG-16 (Simonyan &amp; Zisserman, 2014)</td>
<td>Imagenet (CNN)</td>
<td>Equivalent to discriminators for adversarial imitation learning from visual observations (Li et al., 2017)</td>
</tr>
<tr>
<td>VIS-GRU</td>
<td>Visual</td>
<td>CNN (VGG-16) + GRU</td>
<td>Imagenet (CNN)</td>
<td></td>
</tr>
<tr>
<td>TD-CNN</td>
<td>Top-down</td>
<td>CNN (VGG-16)</td>
<td>Imagenet (CNN)</td>
<td></td>
</tr>
<tr>
<td>BC-CNN</td>
<td>Bar-code</td>
<td>CNN (VGG-16)</td>
<td>Imagenet (CNN)</td>
<td></td>
</tr>
</tbody>
</table>

every video frame into a single line by averaging along the y-dimension of the frame, and then stacking those lines into a single image. If the input video had 600 frames of size 320x200, the resulting bar-code has size 320x600.

We hypothesize that the compressed TD and BC representations capture key patterns of human-likeness while providing higher information density than symbolic or visual observations, leading to more effective classification.

3.3. Models

Following from our problem definition above (Section 3.1), we now detail our proposed classification models. Each model tests specific hypotheses about how to best learn to align with human judgments. A summary of all models is provided in Table 1, with details as follows:

FF. Feedforward models extract features from individual states, without taking into account the sequential structure along a trajectory. We construct this type of model as a representative baseline in line with discriminators typical of adversarial imitation learning (Ho & Ermon, 2016).

GRU. Recurrent models explicitly model sequence information and are expected to more directly align with human assessments (human assessors view observations in the form of videos, c.f., Section 4). We encode sequences using Gated Recurrent Unit (GRU) networks (Cho et al., 2014).

CNN. Convolutional models are applied to image input (visual, top-down and bar-code observations). We use a VGG-16 (Simonyan & Zisserman, 2014) pre-trained on Imagenet (Deng et al., 2009) to extract visual features.

For all model training, input batches of individual observations (feedforward) and sequences (GRU) are sampled uniformly from trajectories. At test time, predictions are aggregated per trajectory using the majority vote. Learning aggregation at the trajectory level is an important direction of research, however improving simple and robust aggregation baselines with learned approaches is challenging (Radford et al., 2018; Lan et al., 2020).

Before moving forward, we validate that the proposed models are effective on the task they are trained to perform - distinguishing between behavior generated by human vs artificial agents. The validation curves in Figure 2 show that all proposed models learn to distinguish between these on our validation set. This confirms that we have a strong basis for evaluating which of the proposed approaches matches ground truth provided by human assessors. Our main results on this question will be presented in Section 5, after we detail our approach to validating the classifiers in Section 4.

4. Validating ANTT by Human NTT

This section details our approach to validating the methods proposed for automatically evaluating human-like behavior (Sec. 3) with a human NTT (HNTT). The following components are required for this validation, and are discussed in turn below. First, we detail the 3D navigation task on which human-like behavior is assessed (Sec. 4.1). Second, we need to collect behavior data from human players (Sec. 4.2) and artificial agents (Sec. 4.3). Third, we obtain ground truth human judgments of human-likeness (Sec. 4.4).2

4.1. Definition of the 3D Navigation Task

Our experiments are conducted in a modern AAA video game.3 A screenshot of the navigation task is shown in Figure 3, together with a top-down view of the map. The task is set up on a map that is representative in size and

2See Appendices for further details: A.2 (agent training), A.3 (human data collection) and A.4 (evaluation).

3We thank our colleagues at Ninja Theory for providing access to this game for research purposes only.
Figure 3. Screenshot of the 3D navigation task as perceived by human players (left). Mini map of the level (right), indicating as a blue circle the player’s spawning location (bottom) and in green the 16 possible goal locations. The main map is approximately 500x320 meters in size and includes complex 3D elements. The screenshot is not representative of the final game play and visuals.

complexity for a modern 3D video game, and comparable to recent navigation experiments, such as the large game map explored by Alonso et al. (2020). The players’ spawning position is on an island outside the main map, connected by a set of 3 available jump areas. Agents must learn to navigate to a target location on the main map, represented by the blue containers in the left screenshot in Figure 3. On each navigation episode, the target spawns uniformly at random in one of 16 possible locations. In preliminary tests we confirmed that human players can successfully navigate to all goal locations within 26 seconds on average.

4.2. Collecting Human Navigation Data

We now describe how we collected human navigation data.

**Requirements.** Data from a total of 7 human players was collected using the replay capture engine in an experimental game build implementing the navigation task (Section 4.1). Human players’ expertise with video games varied, ranging from casual to advanced players. All players had high familiarity with the game map and task, and recorded replays on machines that met the system requirements of the experimental game build, including GPU rendering support.

**Parameters and Constraints.** The navigation target was indicated on-screen in a mini-map that was shown throughout the navigation task (Fig. 3). There was a limited action space: agents and human players could only navigate (e.g., jump, attack and special abilities were disabled). The human players used keyboard and mouse navigation only.

**Video Generation.** The episode replays were saved as MP4. A total of 140 human recordings were collected, which we split into 100 videos (from 4 players) for classifier training and validation, and 40 (3 remaining players) for testing.

**Post-Processing.** We applied post-processing edits to the videos for final inclusion into the HNTT ground truth assessment (Section 4.4). These edits are defined by a set of rules which were uniformly applied to all the videos. Rule 1: Mask identifying information (computer name tag). Rule 2: Cut all videos to start the moment the character lands on the main map.\(^4\)

**Rule 3:** Cut the last 1-3 seconds of the human player videos to just before they stopped moving to correct for an effect of the human data collection process, where human players had to manually end the recording and thus appeared briefly “stopped” at the end of each episode.

4.3. Collecting Agent Navigation Data

This section details training of the artificial agents used to generate agent navigation data, including agent architectures, training process, and agent navigation data collection.

Figure 4 shows our two reinforcement learning agent architectures: symbolic and hybrid. Both are representative examples of current state-of-the-art approaches to learning 3D navigation in video games, closely following very recent work by Alonso et al. (2020). We chose this starting point as a representative approach that learns effective navigation in a game similar to ours, and select two observation spaces and architectures for which we expect high task performance based on the ablation studies of Alonso et al. (2020).

**Symbolic.** The *symbolic* agent (Fig. 4a) observes a low-dimensional representation of the goal (relative angle and distance) and game state (agent position, average frame depth), processed using a feed forward network. We hypothesize that this compact observation space enables sample-eficient learning but relatively less human-like behavior.

**Hybrid.** The *hybrid* agent (Fig. 4b) receives the symbolic observation, as above, and a 32x32 depth buffer of the game footage. Image features are extracted using a CNN before concatenating with the output of the feed forward layers that process symbolic features. Given access to (coarse) visual and depth information, we hypothesize the agent to be better aligned with the type of information used by human players and to consequently learn more human-like behavior.

Both agents use a discrete action space: moving forwards, left and right to a set of given degrees (30, 45, 90), as well as turning in place. Both agents receive the same reward signal during training. This consists of a dense reward for minimizing the distance between the goal location and the agent’s position, a +1 reward for reaching the target, and a -1 penalty for dying, as agents are able to jump off the edges of the map. In addition, a small per-step penalty of −0.01 encourages efficient task completion. Episodes end when agents reach the goal radius or after 3,000 game ticks (≈ 50 seconds, 210 agent steps).

The agents were trained using PPO (Schulman et al., 2017), implemented in Tensorflow (Abadi et al., 2015). Behavior\(^4\)This is to focus our current validation on ground-based navigation. Assessing human-like flying is left for future work.
policy rollouts during training were collected through a scalable distributed framework on 60 parallel game instances. In each training episode, the agent spawned in one of the 26 predefined locations, both on the main map and in the spawn area, with a 34% preference for the latter. This uniform domain randomization (Tobin et al., 2017) was found sufficient for sample efficient learning in preliminary experiments.

Before proceeding to the next stage, we verify that the selected agent models successfully learn the navigation task. Figure 5 verifies that the agent models used here are representative of the state of the art of learning to navigate in complex 3D games. As shown, agents learn to reliably reach the goal within the first 5M training steps. On average, the hybrid agent achieves task completion 10% faster than the symbolic agent. This demonstrates that our trained agents achieve the required level of task success, allowing us to focus on whether the learned behavior is also human-like.

To capture agent navigation data we rolled out the model checkpoints after 20M training steps. The rolled out policies are non-deterministic. The starting position for all rollouts was the spawn area (as for human players). The episodes were saved as videos and then sampled to match the navigation goals in the selected human navigation samples. The agent navigation was post-processed in line with the human navigation data (Sec. 4.2).

4.4. Collecting Ground Truth Human Judgments

We designed two behavioral human NTT (HNTT) studies to establish ground truth based on collecting human judgments of agent vs. human navigation in the game. Each study was administered in the form of an online survey.

IRB Approval. The survey was approved by our institution’s Institutional Review Board (IRB), authorizing us to collect data from internal employees. An IRB approved digital consent form was obtained before participation.

Participants. A total of 60 human assessors participated in the HNTT. We recruited participants from internal mailing lists, none of the participants were familiar with our study. Participants were required to be 18 years or older. No additional demographic information was collected. Participants were asked to indicate their familiarity with third-person action games and the game title used in this study.

Study Design. Each HNTT study consisted of 10 Turing Test trials, in each of which the stimuli consisted of two side-by-side videos, A and B, of an avatar navigating a game environment (Figure 6). The task of the participant was to watch each video and answer the question “Which video is more likely to be human?”, explain their rationale, and indicate uncertainty of their decision on a 5-point Likert scale varying from “extremely certain” to “extremely uncertain”.

In each study all participants viewed the same ten HNTT trials with the same stimuli, presented in randomized order per participant (within-subject design). Videos A and B were selected from a set of short videos in which the symbolic or hybrid agent (Section 4.3), or a human player (Section 4.2) completed the navigation task (Section 4.1).

Stimuli and conditions. The presented video stimuli and trial conditions differed between Study 1 and Study 2 according to Table 2, administered as follows.

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Hybrid</th>
<th>Symbolic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>-</td>
<td>Study 1</td>
<td>Study 2</td>
</tr>
<tr>
<td>Hybrid agent</td>
<td>Study 1</td>
<td></td>
<td>Study 1,2</td>
</tr>
<tr>
<td>Symbolic agent</td>
<td>Study 2</td>
<td>Study 1,2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Table of HNTT comparison conditions in Study 1 and Study 2. Each study consisted of 6 human vs. agent comparison trials, and 4 trials comparing the two agents.

Study 1. Human participants were asked to judge a human player vs. a hybrid agent in 6 trials with side-by-side video pairs, and a symbolic agent vs. a hybrid agent in 4 trials.
In this section we detail our main empirical results of two Navigation Turing Test studies, obtained according to the validation approach detailed above (Sec. 4).

5.1. HNTT: Human Assessments of Human-Likeness
We analyzed human behavioral data from two human NTT (HNTT) studies, in each of which participants observed 10 pairs of video stimuli, judged which is more human-like, and marked their uncertainty. These results were used to establish ground truth for automated NTT studies (Sec. 4.4). We first assessed the extent to which human participants accurately judged videos of real human players as more human-like than videos of RL agents (Fig. 7). Participants accurately detected human players above chance in both Study 1 (mean=0.84, std=0.16, n=30 throughout for both studies) and 2 (mean=0.77, std=0.16). On average, participants were slightly more accurate in judging humans as more human-like in Study 1 compared to Study 2 (Mann–Whitney $U = 341$, $p = 0.046$). We then analyzed the reported uncertainty of participants, reported on a scale of 1 (extremely certain) to 5 (extremely uncertain), Fig. 8.

In Study 1 participants indicated a mean certainty of 2.1 (std=0.47), corresponding to the level “somewhat certain”. In Study 2, participants indicated an average certainty of 2.6 (std=0.84), between “somewhat certain” and “neutral”. The distributions in the two groups differed significantly (Mann–Whitney $U = 257.5$, $p = 0.002$), indicating that participants were less certain of their decisions in Study 2.

Next we compared how participants judged symbolic and hybrid agents (Sec. 4.3). In both studies, and in line with our prediction, human participants judge the hybrid agent to be more human-like when directly compared with symbolic agents (Fig. 8). At the same time, we observed substantial variability among human judges. A comparison across studies reveals an interesting pattern. In hybrid-symbolic comparisons of Study 1, on average $0.78$ of all participants judged the hybrid agent to be more human-like (std=$0.25$). This proportion is substantially higher than that of Study 2 (mean=$0.62$, std=$0.30$), with a statistically significant difference (Mann-Whitney $U = 328.0$, $p=0.013$).

This finding suggests that there are differences in the degree to which the hybrid agent is judged as more human-like between the two studies. Possible explanations could be confounds, however we eliminated potential confounds by matching characteristics of hybrid and symbolic agent videos across the two studies, making this explanation unlikely. Another possibility are effects such as anchoring or exposure bias. If this is the case, this not a confound but a finding that needs to be taken into consideration in future...
Navigation Turing Test (NTT)

Figure 9. ANTT Classifier accuracy compared to ground truth identity or human judgments on held-out test data.

Figure 10. Spearman rank correlation of ANTT rankings compared to human judgments on held-out test data.

work. If confirmed, such effects would show judgments of human-likeness to be context-dependent. Future agents could use meta-reasoning to adapt their behavior to be more human-like given a context. Our results show that identifying behavioral features based on which human judges assess human-likeness is a key challenge that needs meticulous investigation to be better understood.

Overall, our analysis of human judgments show that, in line with expectations and on average, videos of human navigation are judged more human-like than those of artificial agents. In addition, as hypothesized in Sec. 4.3, the hybrid agent is judged more human-like when directly compared with the symbolic agent (Figure 8). These results validate our HNTT methodology, enabling us to validate our classifiers (artificial judges) against the human judgment ground truth reported in this section. Future behavioral studies can explore individual differences in human judgments of both human and artificial navigation.

5.2. Evaluating the ANTT

We now turn to our results of evaluating our proposed ANTT. We aim to identify classification models (Sec. 3.3) that accurately capture human judgments of human-likeness. All models were trained to distinguish human from agent trajectories. Training and hyperparameter tuning was performed using 5-fold cross validation on trajectories generated by agent checkpoints and human players that were fully separate from those that generated test data. We collected human judgments (Sec. 5.1) for test trajectories only.

Figure 9 shows the accuracy of all considered classifier models (see Table 1 for model details). First, we verify that all models generalize to the held out training data. We assess this by comparing classifier predictions (aggregated on the trajectory level by majority vote) against the true identity of the agent (i.e., human or agent) termed identity accuracy. We find that all models generalize to the test data, with results for the symbolic models particular robust to overfitting. Results range from an accuracy of 0.583 for TD-CNN to 0.850 for both symbolic models. These results also validate our trajectory level aggregation by majority voting approach, demonstrating that is achieves high accuracy when comparing to trajectory level human judgments.

Our main results compare classifier models to human judgments, split into human-agent accuracy and hybrid-symbolic accuracy (Figure 9). As the majority vote aggregation of human responses was equal to the true identity for all questions that compared human videos with agent videos, we again observe the same high accuracy on human-agent judgments. However, accuracy is low for hybrid-symbolic pairs across all classifiers.

Our classification results demonstrate that classifier models
do generalize well to unseen test trajectories. Agreement with human judgments is high when assessing the human-likeness of agents compared to human players. Highest agreement is observed for symbolic models. This relative performance of classifiers is unexpected, given that the symbolic classifiers observe a representation that is very different from that used by human judges for decision making. Based on our results, it is likely that the visual models overfit given the richer representations and higher model capacity. In addition, we find that current classifiers perform poorly when assessing the degree of human-likeness of pairs of agents. Our results motivate important follow up work to close this gap between matching human judgments on human-agent vs agent-agent pairs.

6. Conclusion
We introduce the automated Navigation Turing Test (ANTT), a new approach to automatically evaluating the human-likeness of navigation behavior. Our results, comparing to a human NTT, show that a high level of agreement between automated and human judgments can be achieved, opening the door to using classification approaches as scalable proxies for driving the development of more human-like agents. Importantly, we do not (and do not recommend) directly using these models to reward learning human-like policies as doing so can lead to agents that exploit static classifiers (Hernández-Orallo, 2020; Shi et al., 2021).

At the same time, we show that detecting finer-grained human judgments that capture agents’ improvements towards human-likeness remains an open challenge for future work. Conceptually related work by (Groß et al., 2017) showed enabling the discriminator to interrogate the generator improved performance over passively learning to discriminate via supervised learning alone. The ability of the discriminator to interrogate was also a core part of the original Turing Test and may be key to improving finer grained discrimination between agents.

Our work takes an important step on the dual goals of understanding human assessment of human-like behavior and towards agents capable of learning complex human-like behaviors.

Acknowledgements
We thank Batu Aytemiz, Dave Bignell, Mikhail Jacob, Mingfei Sun, Robert Loftin, Sebastian Nowozin, and Stephanie Reikine for their feedback and valuable discussions. We are grateful to Ninja Theory and Microsoft for their support. We thank the anonymous participants of our human NTT studies for their time and effort.

References


