nature human behaviour

Article

https://doi.org/10.1038/s41562-026-02590-6

Turing-like tests for language and vision demonstrate a narrowing gap between humans and AI

Received: 6 August 2024

Accepted: 25 August 2026

Published online: 26 October 2026

Check for updates

Mengmi Zhang[1,2,3,16], Elisa Pavarino[4,5,16], Xiao Liu[1,2,16], Giorgia Dellaferrera[3,6,7], Ankur Sikarwar[1,2], Caishun Chen[2], Marcelo Armendariz[3,5], Noga Mudrik[8], Prachi Agrawal[9], Spandan Madan[3,5,10], Mranmay Shetty[11], Andrei Barbu[3,12], Haochen Yang[10], Tanishq Kumar[10], Shui'Er Han[1,2], Aman Raj Singh[9], Meghna Sadwani[13], Stella Dellaferrera[14], Michele Pizzochero[10,15], Brandon Tang[10], Yew-Soon Ong[1,2], Hanspeter Pfister[10] & Gabriel Kreiman[3,5]

As artificial intelligence (AI) becomes increasingly embedded in daily life, ascertaining whether an agent is human is critical. Here we systematically benchmark AI's ability to imitate humans in three language tasks (image captioning, word association and conversation) and three vision tasks (colour estimation, object detection and attention prediction), collecting data from $n = 636$ humans and 37 AI agents. We also conducted 72,191 Turing-like tests with 1,916 human judges and 10 AI judges. Current AIs are approaching the ability to convincingly impersonate humans and deceive human judges in both language and vision. Even simple AI judges outperformed humans in distinguishing AI from human responses. Imitation ability showed minimal correlation with conventional AI performance metrics, suggesting that passing as human is an important independent evaluation criterion. The large-scale Turing-like datasets and metrics introduced here offer valuable benchmarks for assessing human-likeness in AI and highlight the importance of rigorous, quantitative imitation tests for AI development.

Current language and computer vision models excel in a wide range of tasks, such as image captioning[1–5], text generation[6–12], object recognition[5,13,14] and attention prediction[15–17]. Developing state-of-the-art algorithms goes hand-in-hand with the creation of better and more precise ways of assessing their performance. The performance of artificial intelligence (AI) algorithms is often defined by comparing their outputs against human ground-truth annotations. Such metrics are particularly prevalent in computer vision, such as in object recognition tasks. Additionally, several metrics have been adapted from natural language processing tasks to image captioning, such as BLEU[18], THUMB[19] and METEOR[20] (see also ref. 21). Generative AI models are notoriously difficult to evaluate due to the inherent ambiguities of creating novel content[22]. Human evaluators are often recruited to assess the quality of sentiment, semantic relevance, reasoning abilities or emotional valence on text generated by large language models[9,22–25].

The Turing test, also known as the imitation game, was proposed by Alan Turing in 1950 as a way of assessing a machine's ability to exhibit intelligent behaviours indistinguishable from those shown by humans[26]. In the imitation game, a machine tries to pass as a human during a conversation, and a human judge determines whether they are interacting with a human agent or a machine[26] (Fig. 1a). Since its inception, the Turing test has generated extensive controversy

A full list of affiliations appears at the end of the paper. e-mail: gkreiman@gmail.com

https://doi.org/10.1038/s41562-026-02590-6

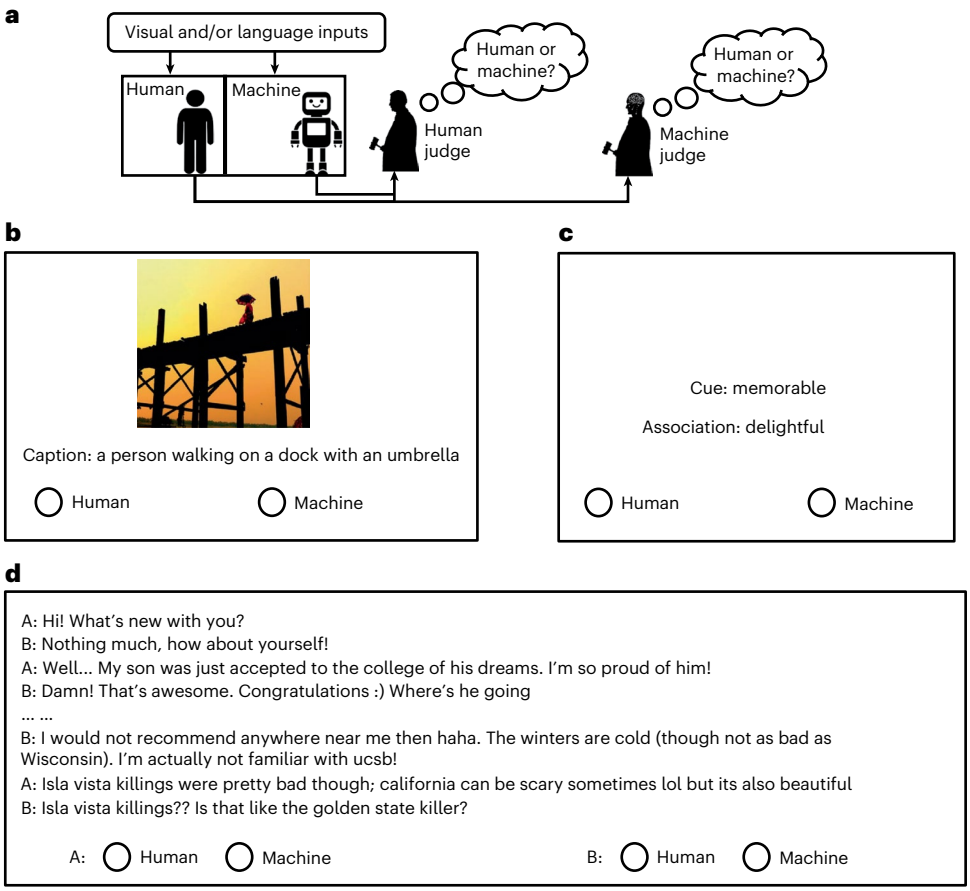


Fig. 1 | Turing-like tests for three language tasks. a, In each task, a human or a machine agent produces an answer in response to visual or language inputs. Given those inputs and answers, a human or machine judge assesses whether the answer came from a human or a machine agent. The machine agent is said to pass the Turing-like test if the judge is unable to distinguish whether the response came from a human or a machine. **b**–**d**, Turing-like tests for three language tasks (see also Fig. 2 for the three vision tasks). We ask the reader to try the tests before checking the ground-truth answers provided at the end of this figure caption. In the image captioning task (**b**), the agent provides a single-sentence description of an image. The judge is presented with an image and a caption and decides whether the caption was produced by a human or a machine. In the word association task (**c**), the agent is presented with a word cue and has to produce a single word related to the cue. The judge is presented with a pair of cue and association words and decides whether the association was produced by a human or a machine. In the conversation task (**d**), two agents (here labelled A and B) alternate to produce a total of 24 exchanges (only 7 exchanges are shown here; for full examples, see Supplementary Information section 1). The judge is presented with the conversation and decides whether each conversant was a human or a machine. There were human–human conversations, human–machine conversations and machine–machine conversations. See Methods for the detailed description of data collection and Turing-like tests in each task. Ground-truth answers for these examples: **b**, machine; **c**, machine; **d**, A is a machine; B is a human.

and discussion. Several notable arguments include Searle's Chinese Room thought experiment[27], Block's behaviourism[28], Harnad's Total Turing Test[29], Watt's Inverted Turing Test[30], Damassino's Questioning Turing Test[31] and Sejnowski's Reverse Turing Test[32]. For example, Searle's Chinese Room thought experiment suggested that a system can manipulate symbols without true understanding, challenging the notion that passing the Turing test implies intelligence[27]. Similarly, Block critiqued behaviourist views, arguing that a machine with a vast lookup table could pass the Turing test without genuine cognition or generalization[28].

In parallel with the unbounded optimistic attitudes towards AI in the 1960s and the sober realization of the immense difficulties in AI afterwards, many in the scientific community have shifted away from the question of whether the Turing test is a valid and meaningful measure of intelligence (for example, ref. 33) to focus instead on average performance metrics for AI algorithms. Distinct from these discussions, the purpose of our work is not to argue in favour of or against Turing tests as a measure of general intelligence. Instead, we consider Turing-like tests as a quantitative evaluation of how well current AIs can imitate humans[25].

Many studies have compared how closely AI model responses align with human responses[34] (Supplementary Information section 4). This type of comparison is straightforward in classic visual object recognition tasks and is widely used in traditional computer vision benchmarks[34]. However, in complex and open-ended tasks such as many of the ones in the current work, these metrics are less suitable or undefined. These challenges are similar to those seen in evaluating generative models[6,12]. In the vision tasks examined in this work, multiple responses may be equally valid. For example, in an object detection task, one agent may identify different objects than another agent. Neither may be incorrect, since all the mentioned objects might be present in the image. Therefore, the emphasis in Turing-like tests is not on determining whether a response is correct or incorrect. Instead, the focus is on whether it is possible to distinguish responses generated by humans from those generated by machines.

The key target of the original Turing test focused on conversations. Generating human-like text during conversations remains a daunting challenge for AI, with exciting progress. There have been numerous early attempts at generating restricted topics during conversations, such as Colby's PARRY simulating a paranoid schizophrenic

https://doi.org/10.1038/s41562-026-02590-6

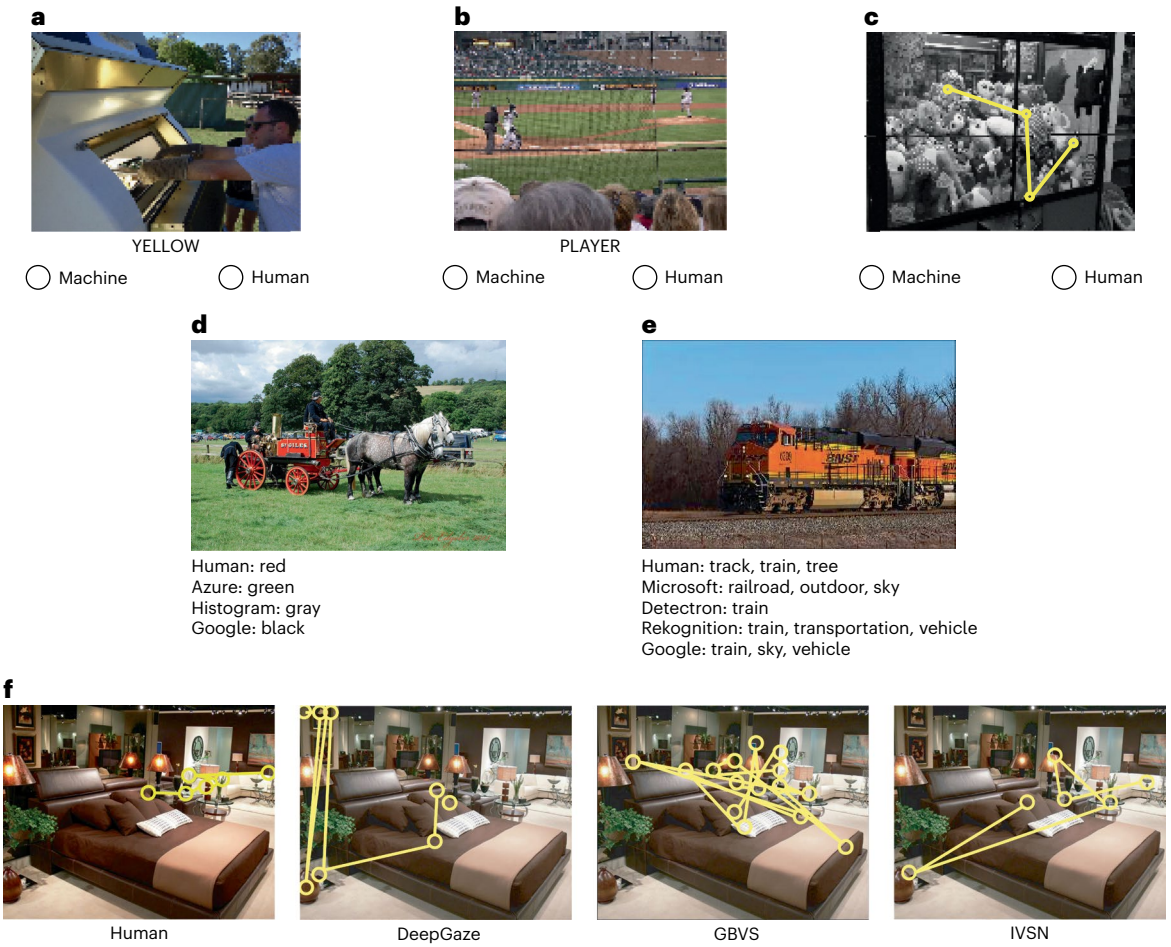


Fig. 2 | Turing-like tests for three vision tasks and example stimuli from human and machine agents for each vision task. See also Fig. 1 for the three language tasks. **a**, Colour estimation. Participants were presented with an image and a colour and had to indicate whether the colour selection was made by a human or a machine. **b**, Object detection. Participants were presented with an image and a noun and had to indicate whether the object description was made by a human or a machine. **c**, Attention prediction. Participants were presented with an image and a sequence of positions (yellow circles) joined by lines and had to indicate whether those locations were the product of human eye movements or machine attention predictions. See Methods for the detailed description of data collection and Turing-like tests in each task. **d**, One example in the colour estimation task from human and machine agents: Microsoft's Azure Cognitive Services[5], the MCCQ algorithm[57] and Google's Cloud Vision API[56]. **e**, One example in the object detection task from human and machine agents: Microsoft's Azure Cognitive Services[5], Facebook's Detectron2[13], Amazon's Rekognition API[58] and Google's Cloud Vision API[56]. **f**, One example in the attention prediction task on visual search from human and machine agents: DeepGaze3[59], GBVS models[17] and IVSN[15,49].

and Weizenbaum's ELIZA simulating a psychiatrist. However, none of these models have come close to unrestricted Turing tests. Advances in large language models[6,9,12,23,35–38] have led news and social media to produce anecdotal claims about current AI algorithms being sentient in conversations. Yet, few studies have rigorously and quantitatively assessed AI algorithms in their ability to imitate humans in conversations[25,39].

As AI algorithms continue to blossom in the real world, it is becoming increasingly important for lay people, law professionals, clinicians, politicians and other experts to ascertain whether the agent they are interacting with is a human or not. For example, the inability to distinguish real news from AI-generated fake news or deepfakes can have disastrous implications for electoral campaigns. Additionally, criminals are increasingly using AI-generated conversations to make phishing scams more convincing, and such scams have even started overtaking other types of physical crimes[40]. To mitigate these issues, the AI community has started developing models to discern whether the generated perceptual inputs come from humans or AIs[41], including deepfake detection on images[42] and GROVER models on fake news detection[43]. Cui et al. proposed a learned critique model acting as a human judge to perform a Turing test in image captioning tasks[21].

Here we set out to systematically and quantitatively investigate the ability of current language and vision algorithms to imitate humans. To this end, we present an integrative benchmark encompassing a wide range of standard and well-established AI tasks across both language and vision. Motivated by the recognition that each AI task measures only a narrow aspect of human intelligence and is incomplete in isolation, we examine cognitive imitation abilities through six different tasks. This approach provides a holistic assessment of the ability to emulate human-like performance across various domains. In vision, colour estimation, object detection and visual attention were specifically chosen because they directly reflect core aspects of human visual perception and decision-making in daily tasks. We also note that some of these vision tasks required language comprehension (see 'General considerations' in Methods). Similarly, among language tasks, word association, image captioning and conversation were chosen because they capture fundamental aspects of linguistic cognition used in daily activities. Although three tasks are labelled as 'visual' tasks, and three tasks are labelled as 'language' tasks, several tasks include both modalities (for example, image captioning).

We start by creating an extensive benchmark dataset of 36,499 answers from 636 human participants and 37 AI models in three

language tasks (image captioning, word association and conversation) and three vision tasks (colour estimation, object detection and attention prediction). These tasks were chosen to span a typical and wide range of real-world applications (Figs. 1b–d and 2a–c and Supplementary Table 1). Next, we systematically evaluated the ability of 1,916 human judges to discern whether task answers came from humans or AI in Turing-like tests, showing that current algorithms are remarkably adept at imitating humans under restricted testing conditions. Finally, as a proof of principle, we show that even though human judges may partially struggle to distinguish human answers, simple machine learning algorithms can serve as AI judges and vastly outperform human judges in Turing-like tests. Our code for data collection and analysis is available to the community, as are the generated datasets. To facilitate future research, we provide an open-source benchmarking tool that allows researchers to evaluate their own models' detectability using our AI judge framework across five tasks, enabling standardized assessment of human-likeness in new AI models.

Results

Collection of human and AI agent responses in language tasks

To conduct Turing-like tests, we first compiled responses from human participants (labelled H throughout) and AI machine models (henceforth labelled M) in three language tasks (Supplementary Fig. 1 and Methods). Multiple controls were introduced in each task to ensure high quality of the responses (Methods and Supplementary Table 3). These responses were then assessed in Turing-like tests to determine whether an independent group of human judges, or AI judges, was able to differentiate between human and AI responses (Fig. 1a). We also collected basic demographic information about each participant (Supplementary Figs. 2, 4 and 7e–g).

We begin by describing the dataset of human and AI-generated responses for each task. In the image captioning task, both human participants and AI models were presented with an image and asked to generate a descriptive caption (Supplementary Fig. 1a). We collected responses for 1,000 images from 229 participants recruited via Amazon Mechanical Turk (AMT), along with responses from 5 AI models (see demographic information about the participants in Supplementary Fig. 2a–c). Representative examples of image captions generated by human participants and each of the AI models are shown in Fig. 3a,b.

We considered two closely related versions of the word association task. In the first version, participants were presented with a cue word and asked to provide a one-word association (Supplementary Fig. 1b1). In the second version, participants were presented with a short prompt containing example word associations, followed by a new cue word, and were again asked to generate a single-word association (Supplementary Fig. 1b2). As the results across the two task versions were similar (Supplementary Fig. 9e,f), we merged the datasets. We collected responses to 1,500 prompts across both task versions from 40 AMT participants and 5 AI models. Demographic information about the participants is shown in Supplementary Fig. 2e,g. Representative responses from human participants and each of the AI models are shown in Fig. 3c.

In the conversation task, two agents engaged in a conversation. Each agent could be a human or an AI model. There were three types of conversations: human–human (H–H), human–machine (H–M) and machine–machine (M–M). Each conversation had a total of up to 24 entries, 12 from each agent. To evaluate the effect of longer conversation lengths, a small subset of conversations included extended exchanges of 48 entries, 24 from each agent. The agents were unaware of whether they were conversing with another human or with an AI model. Human agents were told that they would engage in a brief conversation about different topics. Participants performed the task via a public chat platform where investigators acted as intermediaries to pass messages between agents, or via our custom-developed mobile app (Methods). We collected a total of 510 conversations (each one with 24 entries) from 190 participants. Demographic information about the participants is shown in Supplementary Fig. 2i,k. We collected 3,845 responses from humans and 9 AI models (Supplementary Table 1). The prompts used for each AI model are shown in the Methods. Representative conversations are reported in Fig. 3d and Supplementary Information section 1.

Operational definition of Turing-like tests

After collecting responses from both human and machine agents for each task, we recruited an independent set of participants to carry out Turing-like tests. We refer to our tests as Turing-like to emphasize that they are not identical to the original Turing definitions[26] (Discussion). Turing-like tests were conducted using three recruitment sources: AMT, Prolific (a large online crowd-sourcing platform) and in-lab experiments. We presented a single instance of the answers and asked the participants to indicate whether the answer came from a human or an AI agent (Fig. 1b–d).

We also collected demographic information about the participants as metadata, including age, gender and educational background (Supplementary Fig. 2a,c,e,g,i,k). The trial order was always randomized, with half of the trials containing responses from human agents and the other half containing responses from AI agents. We introduced multiple controls to ensure the quality of the responses collected in each Turing-like test experiment (Methods). Across all tasks, the average accuracy for identifying machine-only responses $p(\mathrm{M}|\mathrm{M})$ was above 75% on the catch trials (Supplementary Table 3). We also trained computational models to act as AI judges to determine whether a particular answer came from a human or not (Methods).

AI models are close to passing three language Turing-like tests

For each Turing-like test, the ground-truth response could come from a human agent (H) or a machine (M). We report confusion matrices where the diagonals show the proportion of trials where judges indicated H given that the ground truth was H ($p(\mathrm{H}|\mathrm{H})$) and the proportion of trials where judges indicated M given that the ground truth was M ($p(\mathrm{M}|\mathrm{M})$) (Fig. 4). Non-diagonal elements show $p(\mathrm{M}|\mathrm{H})$ (the probability that agents indicated M when the ground truth was H) and $p(\mathrm{H}|\mathrm{M})$ (the probability that agents indicated H when the ground truth was M). Entries within each row add up to 1. When comparing different AI models for a given task in terms of their ability to imitate humans, values closer to 0.5 for $p(\mathrm{H}|\mathrm{H})$ and $p(\mathrm{M}|\mathrm{M})$ indicate greater indistinguishability and thus stronger imitation performance. In contrast, when comparing judges (for example, human versus AI judges, or human judges of different ages or educational backgrounds), higher probabilities $p(\mathrm{H}|\mathrm{H})$ and $p(\mathrm{M}|\mathrm{M})$ departing from 0.5 indicate better judges. We summarize the confusion matrices in a single number by defining the overall imitation detectability as $\frac{1}{2}(p(\mathrm{H}|\mathrm{H}) + p(\mathrm{M}|\mathrm{M}))$, ranging from 0.5 (good imitator) to 1.0 (poor imitator).

We summarize the results of the Turing-like tests for human judges across each task in Fig. 4a–c, averaging across all AI models and all human judge demographics. In the image captioning task, human judges correctly identified human captions as human ($p(\mathrm{H}|\mathrm{H})$) 67% of the time, and AI captions as AI ($p(\mathrm{M}|\mathrm{M})$) 46% of the time (both significantly different from chance; one-sample Wilcoxon signed-rank tests; $p(\mathrm{H}|\mathrm{H})$: $z = 13.03$, $P < 0.001$; $p(\mathrm{M}|\mathrm{M})$: $z = 4.88$, $P < 0.001$; Fig. 4a). We were concerned that AMT participants could be poor judges. We therefore repeated the Turing-like experiments in person in our lab (Supplementary Fig. 3a, with full statistics in Supplementary Table 8 and Supplementary Fig. 4a–d). Overall, the results from the in-lab cohort were highly consistent with those obtained online (online judges ($n = 293$): mean detectability, 0.57 ± 0.11; in-lab judges ($n = 11$): mean detectability, 0.57 ± 0.10; Mann–Whitney U-test: $U = 1{,}638.5$, $P = 0.93$).

Performance varied substantially across AI models in the image captioning task (Fig. 5a). The proportion of AI-generated captions

https://doi.org/10.1038/s41562-026-02590-6

a



Human: a woman with an umbrella walking on a pier
GIT: a person sitting on a wooden bridge with an umbrella
OFA: a person walking on a dock with an umbrella
BLIP: a person walking across a bridge with an umbrella
ClipCap: a person holding an umbrella in front of a building
Microsoft_API: a person on a swing

b



Human: a girl receiving a certificate from a woman in a group of people at a presentation
GIT: a woman shakes hands with a student
OFA: two women are talking to each other in front of a crowd
BLIP: two women standing next to each other in front of a crowd
ClipCap: a woman standing in front of a group of people
Microsoft_API: a couple of women in a stadium

c

Cue word	Human	Word2vec	GPT2	GPT3 (embedding, davinci)	GPT3 (prompt, curie)	GPT3 (prompt, davinci)
memorable	nice	phenomenal	delightful	meaningful	unforgettable	delightful
oriental	asian	imperial	colonial	eastern	love	colonial
nepal	country	vietnam	pakistan	nebraska	new	pakistan

d

A: Hey! How are you?
B: Hey, I'm doing great. What about you???
A: I'm good. I just finished up work and am looking for something to do.
B: Did you go for work today???
A: Yeah, I went to the office.
B: We had vacation for a week and the amount of assignments are killing me
A: Yeah, I know. Is there anything you're interested in doing?
B: Right now I'm very much occupied with my university's thing that I don't even have time for other things but I'd love to spend some time learning Korean
A: Korean?
B: Yeah, the language
A: Oh, I see. That's really interesting and something that I would like to learn too one day. Would you like me to teach you?
B: Can you? I never knew you know Korean
A: I learned it while living in South Korea for a year. It's not as hard to learn as people think and I would be happy to teach you.
B: I'd love that! Where did you stay in South Korea?
A: I stayed in Seoul.
B: It's so pretty!! I badly want to visit Seoul and attend BTS's concert
A: You should! BTS is amazing. They're my favorite group.
B: Omg!! Whom do you like the most in the band??
A: I really like Jin, Suga and J-Hope.
B: They're awesome. Which our the bands you know??
A: I like EXO, BTS and 2NE1.
B: Have been to any concert??
A: No, I haven't. But I'd love to go to one!
B: Where are you staying right now??

Fig. 3 | Example stimuli from human and machine agents for each language task. a,**b**, Two caption examples from both human and machine agents: GIT[1], OFA[2], BLIP[3], ClipCap[4] and Microsoft's Azure Cognitive Services[5]. **c**, Three examples in the word association task from both human and machine agents (columns): Word2Vec[7], GPT2[8], GPT3-embedding (davinci)[9], GPT3-prompt (text-curie-001)[9] and GPT3-prompt (text-davinci-002) [9]. **d**, Example of a conversation consisting of 24 exchanges between GPT3 curie[10] (A) and a human (B). See Supplementary Information section 1 for more example conversations, including human–human and machine–machine conversations.

correctly identified as AI ranged from 37% (BLIP) to 60% (ClipCap). Three recent transformer-based models—GIT[1], OFA[2] and BLIP[3]—demonstrated strong imitation capabilities (Fig. 5a). Unlike ClipCap, which is solely trained for image captioning, these models were developed as generative unified transformer frameworks, trained on multiple tasks, including image-text retrieval and visual question answering. This suggests that multi-task training aids models to generate human-like captions. To assess generalization, we included both in-domain images from MSCOCO[44] and out-of-domain images from the nocaps dataset [45], which contains novel object categories not present in MSCOCO. Despite these variations, human judge performance did not differ significantly between out-of-domain nocaps images and in-domain nocaps images (permutation test with Bonferroni correction, $P > 0.05$), even for models such as GIT, OFA and BLIP, which were fine-tuned on MSCOCO (Supplementary Fig. 3d). Moreover, on out-of-domain nocaps images, BLIP outperformed all other models in the Turing-like tests.

In the word association task, human judges correctly identified human-generated word associations as human in 58% of the trials, significantly above chance (one-sample Wilcoxon signed-rank test: $z = -3.36$, $P < 0.001$), and machine-generated word associations as machine in 47% of the trials (one-sample Wilcoxon signed-rank test: $z = -1.63$, $P = 0.10$; Fig. 4b). The results of in-person experiments were comparable to those from AMT participants (Supplementary Fig. 3b; permutation test with Bonferroni correction, $P > 0.05$). Among AI models, Word2Vec exhibited the poorest imitation performance ($p(\mathrm{M}|\mathrm{M}) = 0.58$), while GPT3 (prompt davinci) outperformed the other models ($p(\mathrm{M}|\mathrm{M}) = 0.41$) (Friedman test: $\chi^2 = 3.39$, $P = 0.50$; Fig. 5c). In the conversation task, judges distinguished human-generated exchanges in conversations as humans 66% of the

https://doi.org/10.1038/s41562-026-02590-6

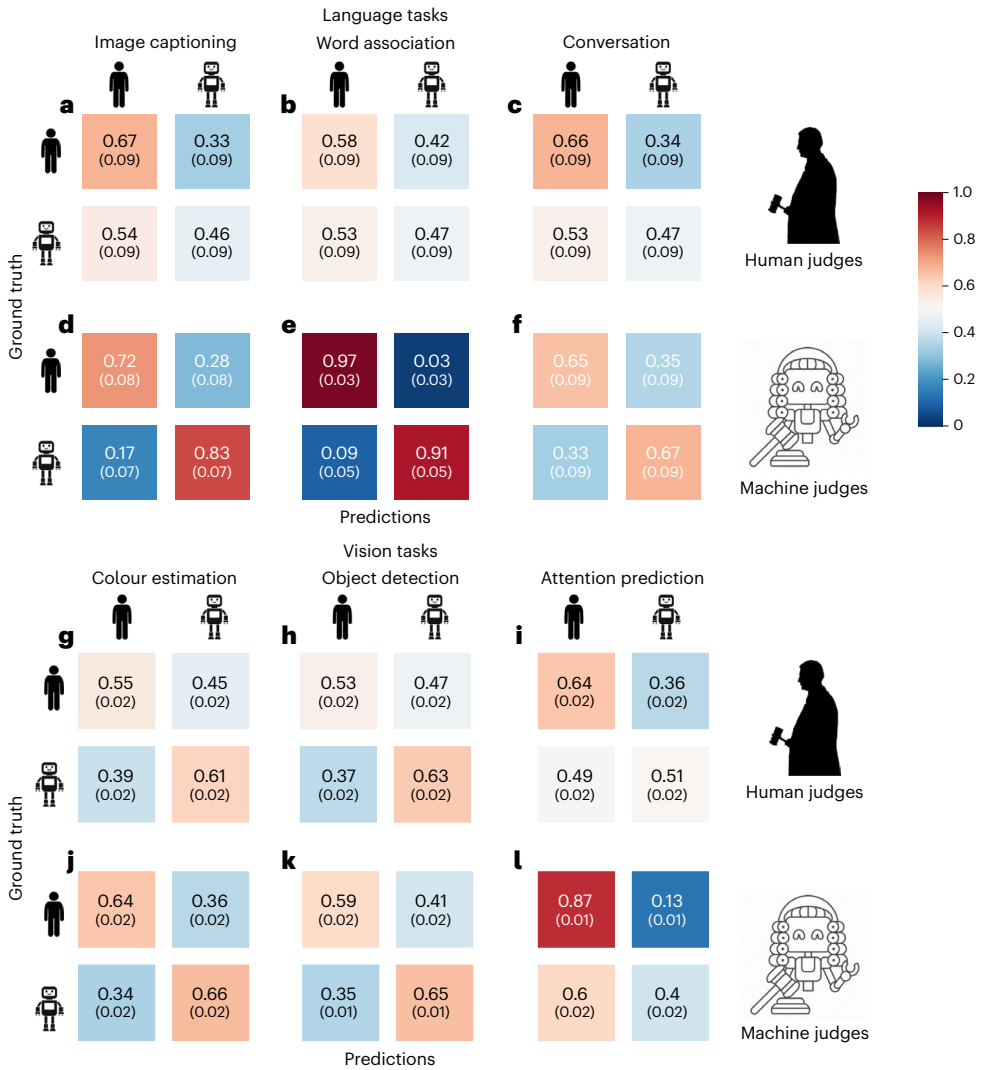


Fig. 4 | Results of the Turing-like test for each task. a–l, Turing-like test results for human judges (**a**–**c**) and machine judges (**d**–**f**) in language tasks and human judges (**g**–**i**) and machine judges (**j**–**l**) in vision tasks. Each column shows the results for a different task: image captioning (**a**,**d**), word association (**b**,**e**), conversation (**c**,**f**), colour estimation (**g**,**j**), object detection (**h**,**k**) and attention prediction (**i**,**l**). For each task, the confusion matrices report the proportion of times when the trial was labelled 'human' (first column) or 'machine' (second column) when the ground truth was human (first row) or machine (second row). The probabilities add up to 1 within each row. Here results are pooled across all machine agents and also across all human judge demographic groups; see Figs. 5 and 6 for results from each machine agent and Supplementary Fig. 2 for results from different human judge demographic groups. The colour of each block indicates performance (see the colour bar on the right). The numbers in parentheses indicate standard deviations.

time (above chance; one-sample Wilcoxon signed-rank test: $z = -5.06$, $P < 0.001$) and machine-generated conversation exchanges as AI in 47% of cases (one-sample Wilcoxon signed-rank test: $z = -1.37$, $P = 0.17$; Fig. 4c). To assess replicability, we repeated the experiments across three platforms: on AMT, on Prolific and in-lab. Unlike the other two language tasks, in-lab participants were slightly more accurate than the Prolific or the AMT participants (Supplementary Fig. 3c; permutation test with Bonferroni correction, $P < 0.05$). We separately considered human–human, human–AI and AI–AI conversations (Fig. 5e). Surprisingly, human–human conversations were classified as human only 64% of the time (above chance; one-sample Wilcoxon signed-rank test: $z = -5.06$, $P < 0.001$), and AI–AI conversations were classified as human 57% of the time (above chance; one-sample Wilcoxon signed-rank test: $z = -2.39$, $P < 0.001$), showing only a modest difference in human detectability between conversation types, which was not statistically significant (Mann–Whitney U-test: $U = 1{,}313.5$, $P = 0.11$).

Blenderbot stood out as a particularly effective imitator: in AI–AI conversations, it was judged human 67% of the time, significantly above chance (one-sample Wilcoxon signed-rank test: $z = -2.83$, $P < 0.005$) and comparable to actual humans in human–human conversations (64%; Mann–Whitney U-test: $U = 1{,}007.5$, $P = 0.70$). In human–AI conversations, human participants were labelled as humans 68% of the time, and AIs were classified as AI 56% of the time (above chance; one-sample Wilcoxon signed-rank tests: $z = -4.75$ and $z = -1.43$, respectively, both with $P < 0.001$). Judges were more likely to identify machine answers in AI–AI conversations than in human–AI conversations (Mann–Whitney test ($n = 47$): $U = 1{,}471.5$, $P = 0.005$), suggesting that AIs may adjust their linguistic behaviour when interacting with humans by producing more human-like sentences. This observation is consistent with the notion that judges are more accurate in making comparisons across speakers rather than making absolute evaluations. When we compared different AI models, Blenderbot was labelled more frequently as human in AI–AI conversations than in human–AI conversations (67% versus 49% of the time; Mann–Whitney test: $U = 1{,}225.5$, $P = 0.02$; Fig. 5e). By contrast, ChatGPT-4o[12] was a slightly worse imitator than Blenderbot (higher imitation detectability) but still outperformed most of the models on this task (Supplementary Table 4, compare with Fig. 5e and Supplementary Fig. 3c). Different language models thus show distinct

https://doi.org/10.1038/s41562-026-02590-6

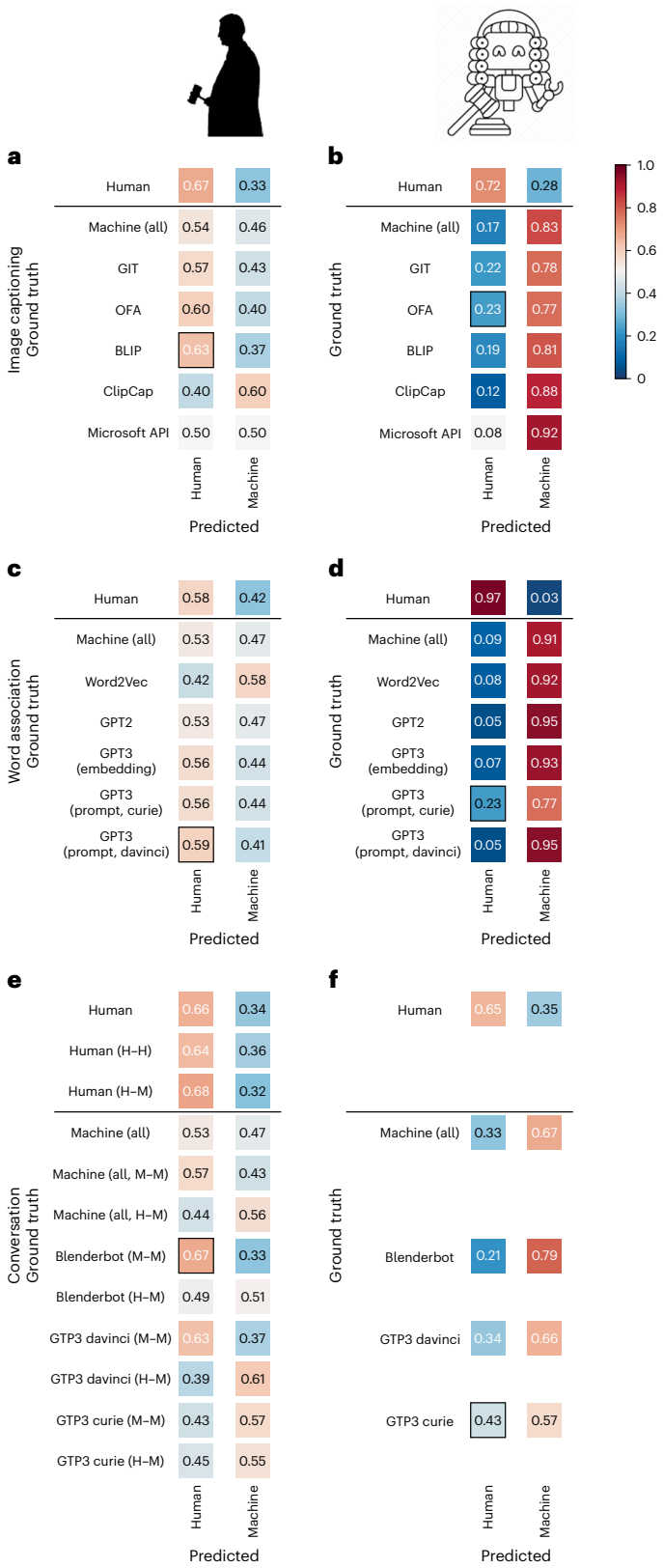


Fig. 5 | Full results of the Turing-like tests for each language task. **a**–**f**, Turing-like test results for human judges (**a**,**c**,**e**) and AI judges (**b**,**d**,**f**) for image captioning (**a**,**b**), word association (**c**,**d**) and conversation (**e**,**f**). The confusion matrices follow the same conventions as Fig. 4. Note that **f** has fewer rows than **e**. The reason is that AI judges take one single sentence as input; thus, there are no multiple exchanges from two speakers involved. The colour bar in **b** applies to all the other panels. The boxes outlined in black highlight the best performing algorithms in terms of their imitation capability.

imitation abilities, and their imitation detectability does not necessarily align with their performance on standard benchmarks.

We next examined whether the outcomes of the Turing-like tests were influenced by conversation length (Supplementary Fig. 5). We evaluated whether detection performance was influenced by conversation length using linear regression per participant (Supplementary Fig. 5). There was no significant relationship between conversation length and human detection accuracy (p(H|H): $\beta = 0.002$ per exchange; Wilcoxon signed-rank test ($n = 45$ workers): $z = -1.12$, $P = 0.27$) or between length and AI detectability (p(M|M): $\beta = -0.0003$ per exchange, $z = -0.25$, $P = 0.80$). This pattern held when we examined specific conversation types (H–H: $z = -0.73$, $P = 0.46$; H–M: $z = -0.36$, $P = 0.72$). These results indicate that judges' ability to discriminate humans from AI remained stable across conversation lengths from 3 to 24 exchanges. We extended these analyses to conversations with 48 exchanges and observed patterns consistent with those seen in 3–24-exchange dialogues. The regression analysis per participant confirmed that 48-exchange imitation detectability did not significantly differ from predictions based on 3–24-exchange trends (Wilcoxon signed-rank tests ($n = 45$ workers); p(H|H): mean residual, −0.02; $z = -0.56$; $P = 0.58$; p(M|M): mean residual, 0.06; $z = -0.40$; $P = 0.69$). This consistency held across conversation contexts, including H–H ($P = 0.83$), M–M ($P = 0.06$) and H–M conversations ($P = 0.34$) (compare Supplementary Fig. 5a,b). When we examined individual AI models, the state-of-the-art large language models had a high human imitation ability across different conversation lengths, though individual model performance showed considerable variability (Supplementary Fig. 5g). In addition to conversation length, we evaluated the relevance of the conversation topic. Imitation detectability varied across topics (Supplementary Fig. 6). For instance, the topic of technology yielded an imitation detectability rate of 60%, whereas the topic of music resulted in a lower imitation detectability of 53%. In sum, imitation detectability remained relatively stable across conversation length, showing no significant effect, and presented modest variation across conversation topics.

We subdivided the results according to the judges' gender and education level (Supplementary Fig. 2, full statistics in Supplementary Tables 6 and 7). Combined averaging across all AI models, there were no significant differences in judges' performance between males and females for the image captioning task (Supplementary Fig. 2b), the word association task (Supplementary Fig. 2f) or the conversation task (Supplementary Fig. 2j) (permutation test with Bonferroni correction, $P > 0.05$). Similarly, averaging across all AI models, we did not detect any significant differences based on the judges' educational level for the image captioning task (Supplementary Fig. 2d), the word association task (Supplementary Fig. 2h) or the conversation task (Supplementary Fig. 2l) (permutation test with Bonferroni correction, $P > 0.05$).

To investigate whether explicitly providing prior information would influence Turing-like test outcomes, we conducted a new version of the experiment in which participants were explicitly informed that 50% of the responses were generated by humans (Fig. 6k). The results showed no significant difference compared to the original setting without explicit instructions ($P > 0.05$, permutation test with Bonferroni correction).

As AI continues to evolve, proficiency in using or scrutinizing the output of AI tools may significantly influence performance in Turing-like tests. To examine this hypothesis, we recruited 33 AI experts and administered the same Turing-like language tasks. Consistent with the hypothesis, AI experts were more accurate at recognizing AI-generated responses than non-experts in the image captioning and word association tasks (Mann–Whitney tests; image captioning task: $U = 885.5$, $P = 0.01$; word association task: $U = 224$, $P = 0.33$), but not in the conversation task (Mann–Whitney test: $U = 241.5$, $P = 0.45$; Supplementary Fig. 4). Experts' ability to identify human-generated responses was also distinct in the image captioning

https://doi.org/10.1038/s41562-026-02590-6

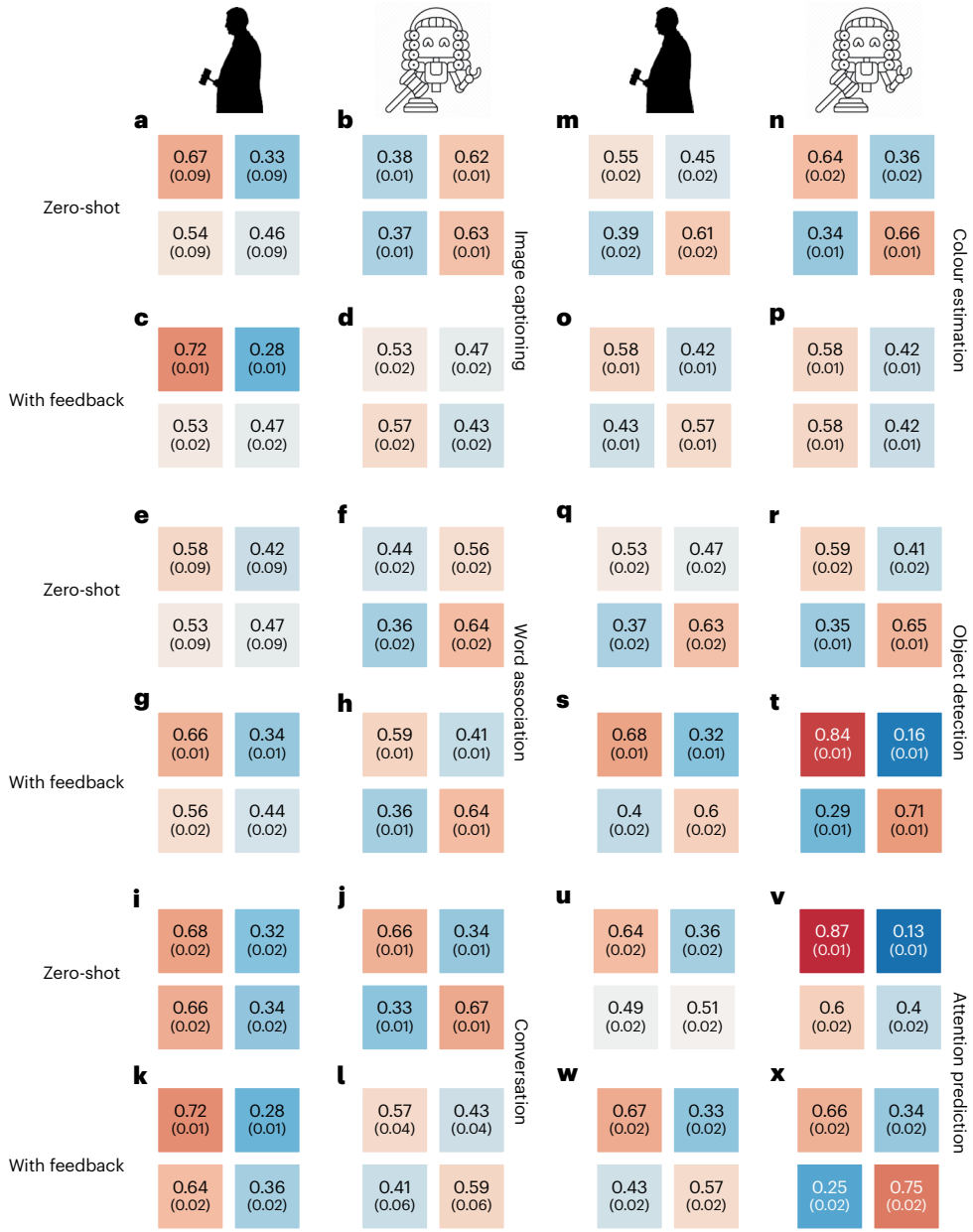


Fig. 6 | Comparison between human and machine judges in zero-shot settings or trained with feedback. a–x, Following the same format as in Fig. 4, the matrices show Turing-like test results for human judges in zero-shot settings (**a**,**e**,**i**,**m**,**q**,**u**) or trained with feedback (**c**,**g**,**k**,**o**,**s**,**w**) across all six tasks and Turing-like test results for machine judges in zero-shot settings (**b**,**f**,**j**,**n**,**r**,**v**) or with the same number of trials with feedback as human judges during the training phase (**d**,**h**,**l**,**p**,**t**,**x**) across all six tasks. For **k** (conversations with feedback), the participants were made aware of a 50% prior of a text being generated by a human or a machine.

task (Mann–Whitney test: $U = 2{,}439.5$, $P = 0.01$), but not in the word association task (Mann–Whitney test: $U = 205.5$, $P = 0.21$) or the conversation task (Mann–Whitney test: $U = 309.0$, $P = 0.62$).

To summarize the results, we computed the overall imitation detectability as the average of $p(\mathrm{H}|\mathrm{H})$ and $p(\mathrm{M}|\mathrm{M})$. By construction, a poor imitator would yield $p(\mathrm{H}|\mathrm{H}) = p(\mathrm{M}|\mathrm{M}) = 1$ and therefore an overall imitation detectability of 1.0, whereas a perfect imitator would yield $p(\mathrm{H}|\mathrm{H}) = p(\mathrm{M}|\mathrm{M}) = 0.5$ and therefore an overall imitation detectability of 0.5. The overall imitation detectability was 0.57 for image captioning, 0.53 for word association and 0.57 for the conversation task. The results show that current algorithms are not far from passing the Turing-like tests under the conditions examined here.

Imitation detectability scores should be interpreted with caution. If judges classified all responses as human, this would result in $p(\mathrm{H}|\mathrm{H}) = p(\mathrm{H}|\mathrm{M}) = 1$, also leading to an imitation detectability of 0.5 (similarly, with $p(\mathrm{M}|\mathrm{M}) = p(\mathrm{M}|\mathrm{H}) = 1$). The experimental data indicate that human judges did not exhibit either of these behaviours (Fig. 4). The total number of H responses was far from 100% and close to 50% in all the experiments. For instance, across the three language tasks, $p(\mathrm{H}|\mathrm{H})$ ranged from 0.58 to 0.67, while $p(\mathrm{H}|\mathrm{M})$ remained between 0.53 and 0.54. Another way to get an imitation detectability score of 0.5 arises from completely random responses. We implemented multiple controls and catch trials in the tests, and the probability of correctly identifying these trials was consistently high (Supplementary Table 3), confirming that human judges were not merely providing random responses. For a more granular view of the results, we also present full confusion matrices in Fig. 4, which provide a more detailed breakdown of error types.

In our original Turing-like tests, human judges made judgements in a zero-shot setting, without any feedback. To investigate whether feedback could enhance their ability to distinguish between human and machine responses, we conducted a follow-up experiment that included a training phase consisting of Turing-like test trials in which judges were told whether each response was correct or not, followed by a test phase without feedback. The results from the test phase indicate a marginal improvement in imitation detectability following the feedback-based training (Fig. 6), although none of them were statistically significant ($P > 0.05$, two-sided t-test, Bonferroni corrected).

AI agents still fail to pass an online Turing conversation task test

Despite the laudable progress in developing large language models, we were surprised by how close these algorithms were to imitating humans in an offline conversation of up to 24 exchanges per agent. To probe this further in a format more closely aligned with Turing's original conception, we developed an online, interactive Turing test implemented as a mobile application (Supplementary Fig. 7). Each conversation involved one human judge, one human agent and one AI agent (GPT3.5-Turbo[12]). The judges initiated the interaction by asking questions, to which both human and AI agents independently responded (Methods). Participants were randomly assigned roles as either judge or human agent.

We constrained the number of exchanges per conversation to be 1, 5, 10 or 20. In each conversation, the mobile application randomly drew one of these four choices and presented this exchange limit to the human judge and the human agent. After the conversation reached the limit in the number of exchanges, the judge was asked to evaluate which of the two sets of answers came from a human and which came from a machine. In stark contrast with the previous section, human judges achieved 100% accuracy (Methods) in distinguishing between human and AI agents (Discussion).

Several factors may explain the discrepancies between the Turing-like offline task presented in the previous section and this online task. First, the test settings differ. In the offline version, there were two agents involved, and both could ask each other questions. Agents were not asked to identify whether they were conversing with a human or a machine. Judges passively observed the fixed conversations offline without actively participating in them or conceiving captious questions. An important example of this situation is content moderation on social media, where a moderator can only review flagged conversations but cannot actively influence them. In contrast, in the online version implemented here, the judge was deliberately trying to probe agents with the clear goal of discerning the identities of both agents. Second, in the previous section, the topics were predetermined and centred on opinions on video games, sports, movies and so on. In contrast, in the online version implemented here, judges decided on the topics, which could delve into personal aspects such as age and family. Despite AI models being assigned randomly generated personal profiles, they typically struggled to express human values and emotions, facilitating identification by human judges. Moreover, advanced scientific topics, such as technical jargon in chemistry and physics, can serve as a means to discern differences between human and AI agents. While such terminology may be challenging for lay humans, AI agents often find it relatively easy to discuss. Third, AIs might reveal their identity when asked personal questions or unsafe questions, such as "Sorry, I am an artificial intelligence program. I am unable to answer questions related to politics or how to kill someone." Fourth, in the online test implemented here, human agents typed responses on cell phones, leading to variation in typing amount, speed and style. This contrasts with AI agents, potentially revealing meta clues such as response times. Although we implemented word limits for AI responses through text prompts and cut off extra sentences exceeding the word limits, and we introduced delays in the AI responses to approximate the typing speed of humans, disparities in response lengths and times persist.

To further evaluate whether human agents could deliberately mimic AI behaviour, we also implemented a variation of the online task where human agents were asked to be deceitful and attempt to imitate machine answers. Despite this adversarial manipulation, human judges consistently identified human and AI agents correctly with 100% accuracy under these conditions. Human agents struggled to replicate the precise behavioural patterns of AI agents.

Given that internal built-in safety mechanisms could contribute to the results in this online version of the test, we conducted a follow-up study using the LLAMA-3-70B-Instruct model[35] with safeguards disabled (Methods). This version included a larger set of human participants ($n = 28$; Supplementary Fig. 7e–g), yielding 188 new trials. The imitation detectability scores from the 28 human judges are shown in Supplementary Fig. 7d. The results reveal that human judges no longer achieved perfect accuracy in distinguishing between human and machine responses across conversations with up to eight exchanges, unlike the previous findings using ChatGPT. The ability to detect machine-generated responses increased with the number of conversational exchanges (regression with binomial weights, $P = 0.032$). Representative examples of both successful and failed trials where human judges correctly or incorrectly identified the sources of the responses are provided in Supplementary Information section 2.

A simple computational model can tell human and AI agents apart

The results presented thus far involve human judges (as in the original Turing test). It is also of interest to assess whether algorithms can tell human and AI answers apart. As a proof of principle to build an AI judge, we used simple support vector machine (SVM) classifiers to evaluate responses generated by humans versus AI. The inputs to the AI judges were embeddings extracted from either words or sentences by different language models (Methods). The SVM classifier was trained to use these embeddings to classify whether a given answer came from a human or not. The predicted labels on tenfold cross-validated data were compared against the ground-truth labels (Fig. 4d–f).

In the image captioning task, we used the OpenAI Embeddings API (text-similarity-curie-001 model) to extract features from captions and trained a linear SVM to discriminate human versus machine responses. Surprisingly, even though this classifier did not use image features, this simple AI judge could discern human answers ($p(\mathrm{H}|\mathrm{H}) = 0.72$) and machine answers ($p(\mathrm{M}|\mathrm{M}) = 0.83$) much better than human judges (compare Fig. 4a versus Fig. 4d). Similar results were observed across all the different AI models (Fig. 5b).

In the word association task, we used the various word embeddings from GPT-series models and Word2Vec as features to train three SVM judges (independent SVM classifiers based on the embedding of each AI model). The results were similar for these different embeddings, and we report the average results here. This simple AI judge vastly outperformed human judges (compare Fig. 4e versus Fig. 4b). Similar conclusions were drawn across different machine agents (Fig. 5d). As expected, when we trained an SVM classifier based on the embedding of the same model whose embedding was also used for generating the associations, this model classification performance was essentially perfect. For instance, on the third row in the left confusion matrix in Fig. 5d, the AI judge trained on Word2Vec embedding predicted very well all association words generated by Word2Vec embeddings as AI. While these cases are good as a sanity check, they should not be considered in the general evaluation of the AI judges' performance. Hence, we trained three different SVM classifiers and consolidated all their trials for the final plots to avoid such biases. We also present generalization results at the end of this section.

In the conversation task, we trained a simple SVM judge to distinguish whether a single sentence was generated by humans or AIs (Figs. 4f and 5f). Even though the AI judge evaluated single sentences (whereas human judges evaluated entire conversations), the simple

AI judge achieved similar performance to human judges in detecting human answers (p(H|H) = 0.65) and surpassed human judges in detecting machine answers (p(M|M) = 0.67). Similar conclusions were drawn for all the language models evaluated here, with the easiest one to detect being Blenderbot (which was the hardest one to detect for human judges) and the hardest one to detect being GPT3 curie (Fig. 5f). Human judges probably focus on high-level conversation understanding rather than single-sentence statistics.

Comparing results between humans and machines is always difficult, especially when training is taken into account. To assess the effect of feedback, we show results from experiments where human judges were provided with feedback in Fig. 6. For practical reasons, humans received feedback in only a small number of trials, which was much less than the amount of training in the SVM judges. For comparison purposes, we retrained our SVM-based judges matching the small number of trials with feedback that human participants were exposed to for each task, mirroring the human training phase (Fig. 6d,h,l). In the image captioning task, human judges continued to outperform the machine judges. In contrast, in the word association task, the machine judge maintained its lead. In the conversation task, performance was comparable between feedback-trained human judges and the SVM judges trained with a small number of trials. Notably, in all cases, SVM-based judges received word embeddings derived from large language models (Methods). These multimodal LLMs are trained on extensive knowledge from vast internet data encompassing text, images and videos.

To test whether AI judges' success could be attributed to basic low-level text statistics, we compared basic properties—including parts-of-speech frequency, punctuation, word frequency, word length and sentence length—between human and AI answers (Supplementary Table 5). Overall, no single feature or set of features reliably distinguished humans from AIs across tasks.

The results presented thus far combine all the different AI agents (always using cross-validation across different trials). To evaluate the extrapolation capabilities of machine judges, we conducted an across-agent analysis where we evaluated their performance under conditions in which they were trained with entries created by some agents and tested their performance with entries created by other AI agents. We conducted two such experiments: 'train on one agent, test on other agents' (Supplementary Fig. 8a,c,e) and 'train on all but one of the agents, test on the left-out agent' (Supplementary Fig. 8b,d,f). As expected, the overall imitation detectability of machine judges in both extrapolation experiments is lower than the accuracy of machine judges trained with responses from all machine agents (compare Supplementary Fig. 8 with Figs. 4 and 5). As expected, the 'train on one, test on others' experiment posed greater difficulty than 'train on others, test on one', resulting in lower overall imitation detectability (compare Supplementary Fig. 8a,c,e versus Supplementary Fig. 8b,d,f). Notably, despite the difficulty in generalizing across different AI agents, AI judges extrapolating to classify responses from excluded AI agents still outperformed human judges in the image captioning and word association tasks. AI judges did not extrapolate well in the conversation task and performed slightly worse than humans, but it should be noted that the AI judges evaluated only single sentences, whereas humans had access to the whole conversation.

To further assess the extrapolation ability of AI judges within each language task, we evaluated a pretrained large language model—ChatGPT[12]—in a zero-shot setting to determine whether it could distinguish between human and machine responses (Fig. 6 and Methods). In the image captioning task, human judges outperformed the AI judge by nearly 10% in imitation detectability (trial-based two-tailed t-test, $P < 0.05$). However, this trend was not present in the word association task, where the machine judge slightly outperformed the human participants (trial-based two-tailed t-test, $P > 0.99$). In the conversation task, we used ChatGPT as the zero-shot machine judge and considered two kinds of prompts. In the zero-shot case, the prompt only included a conversation from the test set, and the model was asked to identify the two agents (Fig. 7d). This zero-shot model performed poorly as a machine judge. In the one-shot case, the prompt included a single conversation example with ground-truth human and machine labels and a test case asking the model to indicate the identities of the two agents (Fig. 7e). ChatGPT performed well as a judge, with an overall p(H|H) = 0.59 and p(M|M) = 0.56, worse than the SVM judges in Fig. 4f, but better than the extrapolation results shown in Supplementary Fig. 8e,f. The performance of ChatGPT in the one-shot case was very different in the H–H and M–M conversations, where it was almost at chance levels, versus in the H–M conversations, where it even exceeded the performance of the SVM judges in Fig. 4f. In contrast to the SVM judges, which were trained on the entire corpus, the binary discrimination task is more challenging for one-shot ChatGPT since there is no explicit training with the task. It is therefore remarkable that even with a single example, ChatGPT can act as a judge to detect machine answers.

Turing-like tests are complementary to standard AI model benchmarks

Algorithm performance is often defined by comparing the predicted outputs of AI models against human ground-truth labels. However, higher accuracy scores in terms of such standard evaluation metrics defined in the AI community do not necessarily translate into better human imitators. To investigate this disconnect, we directly compared Turing overall imitation detectability to standard evaluation metrics in the image captioning task, where there are clearly defined accuracy metrics such as the BLEU score[18] and the CIDEr score[2], to assess the quality of AI-generated captions[46] (Fig. 7). The results, which were similar for both BLEU (Fig. 7a) and CIDEr (Pearson correlation, −0.088 and 0.84; Fig. 7b) scores, highlight that human-likeness and traditional benchmark performance constitute distinct concepts. Some pairs of models have comparable imitation detectability but differ in their captioning scores (for example, ClipCap and GIT, and BLIP and OFA). Conversely, GIT shows a higher captioning score than BLIP, but BLIP is a better human imitator on the basis of the imitation detectability. Flamingo, ChatGPT and LLaVa, which are all large multimodal foundation models for vision and language, yielded imitation detectability that was on par with, or even lower than, that of classical models trained specifically for image captioning. SCST, a classical image captioning model, yielded higher imitation detectability. A qualitative example of differences between accuracy and imitation is shown in Fig. 7c. The algorithm (GPT4) provided an accurate description of the image. Humans also correctly described the image. However, the algorithm did not completely imitate how humans interpret this image, failing to capture some of the key aspects of why the image is striking (see the descriptions in the caption to Fig. 7c).

Current algorithms are also close to passing Turing-like tests in vision tasks

Although the original Turing test was formulated in the context of language, it is possible to define Turing-like tests in many other domains. We extended our investigations to include Turing-like tests in three vision tasks: colour estimation (Fig. 2a), object detection (Fig. 2b) and attention prediction (Fig. 2c). The methodology to collect human responses and AI responses (Supplementary Fig. 1d–f) and to conduct the Turing-like tests (Fig. 2) followed the steps described earlier for the three language tasks (Methods and Supplementary Table 1). Demographic information about the participants in each vision task is presented in Supplementary Fig. 2. In the colour estimation task, agents were asked to identify the dominant colour of an image (see examples of human and machine responses in Fig. 2d). Human judges correctly distinguished human answers in 55% of trials (one-sample Wilcoxon signed-rank test, $z = -2.17$, $P = 0.03$) and AI-generated answers in 61% of trials (one-sample Wilcoxon signed-rank test, $z = -4.13$, $P < 0.001$;

https://doi.org/10.1038/s41562-026-02590-6

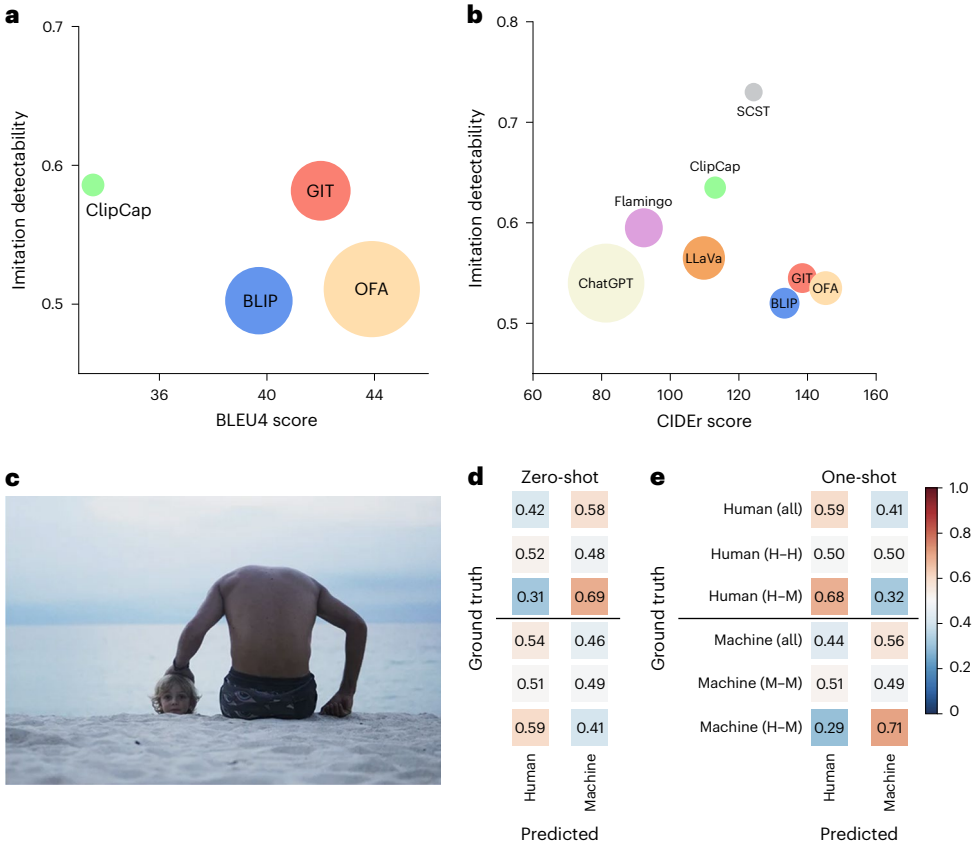


Fig. 7 | Lack of correlation between imitation detectability and standard performance metrics in the image captioning task, and Turing results in the conversation task with AI judges. a,**b**, Four state-of-the-art image captioning models, GIT[1] (red), OFA[2] (yellow), BLIP[3] (blue) and ClipCap[4] (green), are evaluated in terms of their imitation detectability (*y* axis) as a function of their BLEU score[18] (**a**) and their CIDEr score[2] (**b**). We also added four more image captioning models, OpenFlamingo-4B[52] (light purple), ChatGPT-4o-latest[12] (light yellow), SCST[53] (grey) and LLaVA-v1.6-Mistral-7B[54] (orange), for the CIDEr score (**b**). The BLEU score and the CIDEr score are standard evaluation metrics commonly used in image captioning tasks to assess the quality of the generated captions given an image on the basis of *n*-gram overlaps. The higher the values, the better. The radius of each circle is proportional to the AI model size in terms of the number of trainable network parameters. **c**, Image eliciting correct answers that do not necessarily imitate humans. In response to the query "What do you see in this image?" GPT4 answered, "In this image, a man is sitting on a sandy beach with his back to the camera, facing the ocean. Next to him, a child's head is visible, appearing to be buried in the sand up to their neck. The man has one arm extended, resting his hand gently on the child's head. The scene is serene with the calm sea and sky in the background, creating a peaceful atmosphere. The combination of the man's pose and the child's position in the sand creates a unique and slightly humorous visual effect, suggesting a playful moment between the two." While the answer is correct and remarkably articulate, a human described the image differently: "In this curious and well-constructed image, we cannot see the man's head and it appears as if the boy's head belongs to him and the man took his own head off". **d**,**e**, One-shot and zero-shot Turing results in the conversation task with large language models as AI judges. We used the large language model ChatGPT[12] as the AI judge in the Turing-like test for the conversation task. In the zero-shot case (**d**), we prompted the model by directly presenting the conversations from the test sets with explicit instructions to output the identities of the two agents. Similarly, in the one-shot case (**e**), we included one additional conversation example with the ground-truth identities of the two agents in the prompt before presenting the conversation from the test sets followed by the identity prediction questions as in the zero-shot case (see Methods for implementation details). The format convention of the confusion matrices in **d** and **e** follows Fig. 5.

Fig. 8a). Performance was comparable across AI models (p(M|M) range, 0.58–0.66; Friedman test, $\chi^2 = 0.97$, $P = 0.62$). Interestingly, even a simple baseline metric such as the Median Cut Color Quantized (MCCQ) algorithm still achieved moderately good performance, fooling humans in 34% of cases (one-sample Wilcoxon signed-rank test, $z = -2.17$, $P < 0.001$).

In the object detection task, participants were asked, "What do you see in this image?" and had to provide three single words as answers. Human judges correctly identified human-generated answers in 53% of trials and AI-generated answers in 63% of trials (one-sample Wilcoxon signed-rank tests; $z = -1.22$ and $P = 0.22$, $z = -4.29$ and $P < 0.001$, respectively; Fig. 8c). Among all the models, the Detectron agent (a variant of Mask R-CNN[47] trained on ImageNet and MS-COCO) emerged as the most human-like, fooling human judges 52% of the time, significantly outperforming all other models (Google API, 35%; Azure, 31%; Rekognition, 30%; Friedman test, $\chi^2 = 31.03$, $P < 0.001$). Notably, Detectron not only achieved outstanding absolute scores in terms of standard evaluation metrics, such as mAP[48], but its response patterns also closely mimicked humans', often identifying the top-three salient objects in the scene similarly to humans.

We evaluated Turing-like imitation in a third vision task, attention prediction, using gaze data collected via eye tracking. There were two variations of the task: (1) participants freely viewed an image for 12 seconds, shifting their overt attention patterns by moving their eyes, and (2) participants were given a target object and were instructed to locate it within the image using their eye movements (see ref. 15 for experiment details). Overall, human judges distinguished human eye movements 63% of the time (one-sample Wilcoxon signed-rank test: $z = -5.47$, $P < 0.001$) and AI-generated eye movements 50% of the time (one-sample Wilcoxon signed-rank test: $z = -1.73$, $P > 0.05$; Fig. 8e). Models differed significantly in imitation detectability (Friedman test: $\chi^2 = 20.90$, $P = 0.001$). ISVN[15,49] showed the most human-like performance, fooling human judges 57% of the time. We further evaluated

https://doi.org/10.1038/s41562-026-02590-6

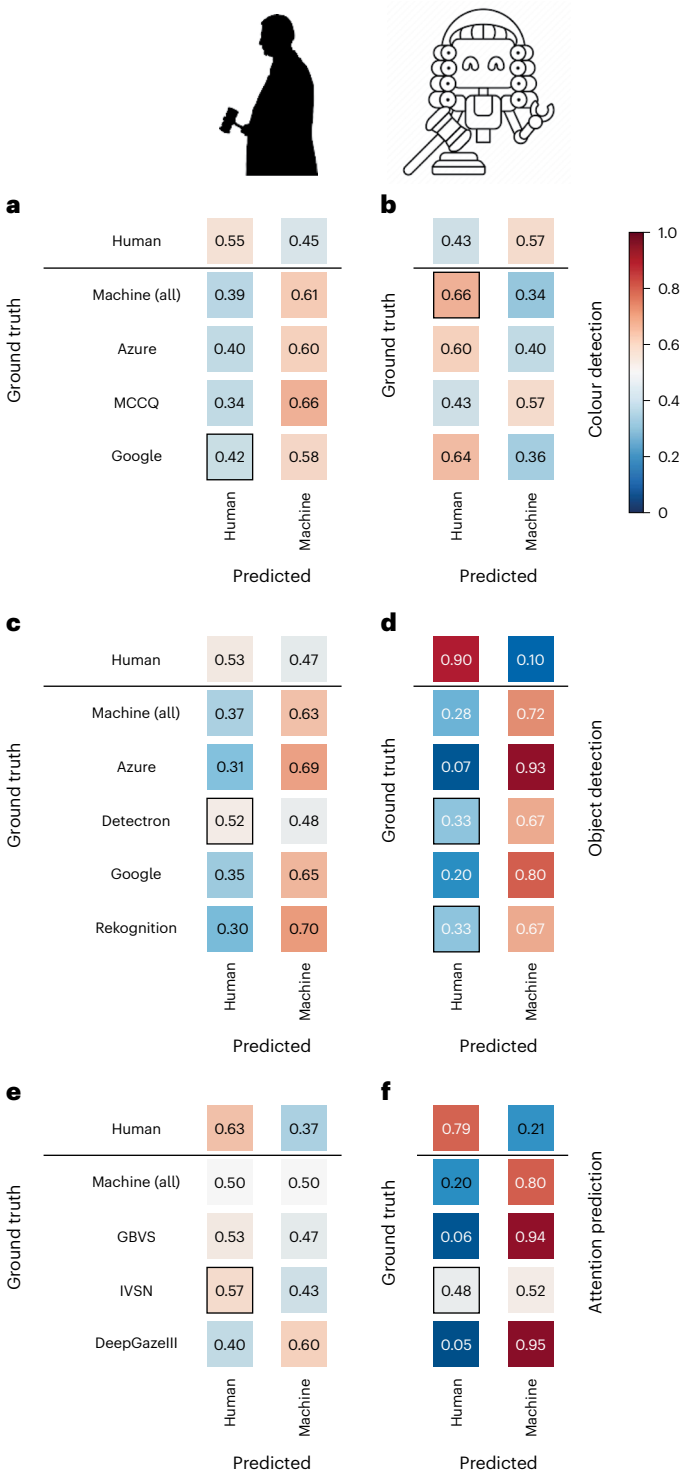


Fig. 8 | Full results of the Turing-like test for each vision task. a–**f**, Turing-like test results for human judges (**a**,**c**,**e**) and AI judges (**b**,**d**,**f**) in the three vision tasks: colour detection (**a**,**b**), object detection (**c**,**d**) and attention prediction (**e**,**f**). The full confusion matrices follow the same conventions as Fig. 4. The colour bar in **b** applies to all panels. The boxes with black frames denote the best algorithm in terms of its ability to pass as human—that is, highest p(H|M).

whether task goals modulated imitation detectability. Human judges performed similarly during free viewing compared to visual search (Mann–Whitney test; free viewing, n = 103, versus visual search, n = 89, p(H|H): U = 4,183.0, P = 0.28; Supplementary Fig. 9a versus Supplementary Fig. 9c). GBVS—a bottom-up saliency model—generated more human-like scanpaths during free viewing than in visual search. The IVSN model's performance was similar across both tasks, emphasizing the importance of incorporating both bottom-up and top-down attention mechanisms in computational models of human attention. Finally, we evaluated machine judges trained to distinguish human from AI-generated eye movements. These classifiers performed well but struggled to label outputs from the IVSN model (Fig. 8f) both in the free-viewing task (Supplementary Fig. 9b) and in the visual search task (Supplementary Fig. 9d).

Evaluating human eye movements for attention prediction is not an intuitive or typical task for human observers. To assess the ability to evaluate eye movements with respect to the underlying image, we introduced catch trials where we randomly swapped eye movements and images. For instance, we showed the eye movements from image 17 overlaid on image 62, and the eye movements from image 62 overlaid on image 9. Under this condition, p(H|H) = p(M|M) = 50%, indicating that judges were unable to distinguish human and machine eye movement patterns without the corresponding image (Methods). Participants were not entirely insensitive to eye movement anomalies. To test participants' sensitivity, we introduced additional catch trials featuring highly anomalous fixation patterns—for example, a single randomly placed fixation repeated throughout the trial. In these cases, participants were able to reliably detect the artificial nature of the eye movements, achieving a detection accuracy of 82%.

To investigate how different model components contribute to human-like performance in visual attention prediction, we conducted Turing-like tests with a series of ablated machine agents based on the IVSN model[15] (Methods). Specifically, the NoTarget model removed the top-down modulation from the target features. The EccNet model[16] incorporated eccentricity-dependent pooling. Instead of extracting features from the top layers of a feedforward neural network pre-trained on ImageNet in IVSN, we used features from the early layers for target modulation and refer to this ablated model as the Layer model. Both EccNet and IVSN exhibited low imitation detectability (Supplementary Fig. 10a). In contrast, the NoTarget and Layer models failed to produce realistic eye movement behaviour during visual search, probably due to the absence of top-down guidance from target features or the inability of their extracted features to capture target semantics. Consequently, human judges could easily distinguish their outputs from actual human eye movements.

We worried that AMT participants could be poor judges. We therefore repeated the Turing-like experiments in person in our lab (Supplementary Fig. 4). Despite our initial concerns, the results from in-person experiments were comparable to those obtained from AMT participants (see Fig. 8a,c,e for vision tasks compared with Supplementary Fig. 4p,t,x; colour estimation task: Mann–Whitney test, U = 282, P = 0.13; object detection task: Mann–Whitney test, U = 342, P = 0.5; attention prediction task: Mann–Whitney test, U = 1,298, P = 0.2).

In our original Turing-like tests, human judges made decisions in a zero-shot setting, without prior exposure or feedback. To assess whether feedback would alter the judges' ability to distinguish human from machine responses in the visual tasks, we conducted a follow-up experiment mirroring the procedure used in the language tasks. Each experiment consisted of a training phase with Turing-like trials where participants received immediate feedback followed by a test phase without feedback (Supplementary Table 2 and Methods). Providing feedback led to a marginal improvement in imitation detectability (two-sided t-tests with Bonferroni correction; colour estimation task: P > 0.99; object detection task: P = 0.001; attention prediction task: P = 0.53; Fig. 6). To assess whether prior training could improve participants' ability to distinguish between human and machine-generated eye movements, we conducted a training phase with feedback as described above (Methods). Following this training, participants' accuracy in the Turing-like task improved modestly from 57.5% to 62% in the attention prediction task (compare Fig. 6u versus Fig. 6w). However, the gains of training feedback were not statistically significant across the three

vision tasks, suggesting that feedback alone may have limited effect in enhancing human judges' discriminative performance in these scenarios.

We examined the impact of human judge demographics on the Turing results. Similar to the results reported above for the language tasks, neither the participant's gender (Supplementary Fig. 2n,r,v) nor the participant's educational level (Supplementary Fig. 2p,t,x) correlated with any major difference in the imitation detectability (permutation test with Bonferroni correction, $P > 0.01$). While gender and age showed little impact on the results, differences depended on the level of AI expertise; however, these differences did not reach statistical significance (compare Supplementary Fig. 4m–x with Supplementary Fig. 4a–l; colour estimation task: Mann–Whitney test, $U = 282$, $P = 0.13$; object detection task: Mann–Whitney test, $U = 342$, $P = 0.5$; attention prediction task: Mann–Whitney test, $U = 1{,}298$, $P = 0.2$).

Finally, as a proof of principle, we followed the approach described for the language tasks to develop simple AI judges to discriminate human from machine answers. In the colour estimation task, we trained SVM judges using image features and text embeddings for the colours (Methods). For the object detection task, we trained SVM judges using the text embeddings for the three nouns describing the images (Methods). In the attention prediction task, we performed binary discrimination using the 2D coordinates of the first ten fixations (Methods). Similar to the conclusions drawn in the language tasks, for the object detection task and the attention prediction task, the AI judges easily discriminated human from machine answers and vastly outperformed human judges. In the object detection task, the AI judges achieved $p(\mathrm{H}|\mathrm{H}) = 0.9$ and $p(\mathrm{M}|\mathrm{M}) = 0.72$ (Fig. 8d). There were large gaps among the different vision models, with the worst imitator being Azure, showing $p(\mathrm{M}|\mathrm{M}) = 0.93$. In the attention prediction task, the AI judges achieved $p(\mathrm{H}|\mathrm{H}) = 0.79$ and $p(\mathrm{M}|\mathrm{M}) = 0.8$ (Fig. 8f), revealing even more drastic differences among models, with the worst imitator being GBVS, reaching $p(\mathrm{M}|\mathrm{M}) = 0.94$, and the best imitator being IVSN, with $p(\mathrm{M}|\mathrm{M}) = 0.52$. In both tasks, the AI judges outperformed the human judges (compare Fig. 8d,f with Fig. 8c,e). The results were different in the colour estimation task. Here the human judges outperformed the AI judges (compare Fig. 8a with Fig. 8b).

The AI judges take the visual features extracted from the deeper layers of ResNet18 and the colour embeddings extracted from BERT as inputs. We also evaluated ablated versions of the AI judges, focusing on either early visual layers or colour responses alone without visual features (Methods). The imitation detectability of the AI judge significantly improved when using the features from the early layer of ResNet (one-way analysis of variance (ANOVA) with Bonferroni correction, $P < 0.001$; Supplementary Fig. 10b,c). Interestingly, colour responses alone can achieve above-chance imitation detectability. This suggests that the long-tailed distribution present in human responses is strongly captured by the AI judge, even though it remains largely imperceptible to human judges.

In addition, to ensure a fair comparison with human judges in the feedback-based experiment (Fig. 6), we introduced an SVM-based machine judge that was retrained using the same number of training examples provided to human judges during the training phase. The impact of feedback varied across tasks: in the colour estimation task, feedback helped human judges close the gap, surpassing the machine judge; in object detection and attention prediction, however, the machine judge not only maintained its lead but improved even further after retraining (Fig. 6p,t,x).

Next, similar to the language tasks, we sought to evaluate the ability of ChatGPT as a machine judge to generalize beyond its training data. For the three vision tasks, we provided ChatGPT with both images and accompanying text prompts to make zero-shot judgements (Methods). Surprisingly, ChatGPT consistently outperformed human judges in all three vision tasks (Fig. 6n,r,v). The performance advantage was especially pronounced in the attention prediction task, showing a gap of up to 8%.

A universal multimodal machine closely imitates human responses

Humans can perform multiple tasks, whereas current AI models are often trained for and thrive in specific tasks. In particular, in the results presented thus far, different AI algorithms were evaluated for each task. We asked whether a single AI algorithm could imitate humans across all six tasks. To address this question, we selected the best-performing off-the-shelf multimodal model, ChatGPT-4o[12], as the universal machine agent. We used ChatGPT-4o to produce responses for each stimulus in a zero-shot manner across all six language and vision tasks. The text prompt used in each task is shown in the Methods. The results, presented in Supplementary Table 4, indicate that the imitation detectability of the ChatGPT-4o agent remained around 0.5 across all tasks (range, 0.44 (object detection) to 0.56 (attention prediction)). This is an impressive result, showing that a unified multimodal model can generate human-like responses that are largely indistinguishable from human responses, even without any task-specific training.

Discussion

We first introduced a dataset of human answers in six common language (Fig. 1) and vision (Fig. 2) tasks to evaluate the ability of algorithms to imitate humans. The dataset also includes answers from state-of-the-art AI algorithms (Supplementary Table 1). Using the human and machine answers, we conducted Turing-like tests, showing that current algorithms are not far from being able to imitate humans in these tasks (Figs. 4a–c,g–i, 5a,c,e and 6a,c,e). In contrast to human judges, even simple machine judges can distinguish human from machine answers (Figs. 4d–f,j–l, 5b,d,f and 6b,d,f).

The Turing test has been extensively discussed, and contested, as a means to assess general intelligence. Our work is not intended as an evaluation of intelligence or to contribute to the discussion of Turing tests to quantify intelligence. Imitating humans can benefit humanity in many situations where we want to align machine and human outputs, such as in emulating expert decisions. However, imitating humans can also be used for evil, as in the dissemination of fake information, phishing attempts or other scams. Regardless of whether Turing tests are good or bad metrics of intelligence, it is of high practical importance to assess whether algorithms can imitate humans or not.

In some cases, we summarized results using the overall imitation detectability metric. It is important to emphasize that this summary metric does not differentiate between the two types of classification errors: (1) false negatives, in which a human response is misclassified as machine-generated; and (2) false positives, in which a machine response is misclassified as human-generated. Understanding the distinction between these two error types is important for interpreting decision-making by human and machine judges. Moreover, in real-world applications, these errors may have different consequences. For example, misclassifying a machine as a human might be more problematic than misclassifying a human as a machine if the goal is to reliably detect AI-generated content. We present confusion matrices throughout to complement the imitation detectability scores and quantify the two types of errors.

One approach towards mitigating the risks from evil human imitators is to build AI judges that can distinguish true human answers. The results presented here show that even simple AI judges like the ones introduced here can generally do a better job than human judges in detecting machine answers for the current tasks. There will probably continue to be an arms race between the development of better imitators and better detectors.

In most of the Turing-like tests presented here, AI judges were explicitly trained to classify responses from humans versus AIs, while human judges received no explicit training. However, we also presented results using ChatGPT-4o, a state-of-the-art, off-the-shelf, multimodal model. We evaluated ChatGPT-4o in zero-shot Turing-like tests across all six tasks. This unified, multimodal model can distinguish between

human and machine-generated responses with above-chance accuracy, even without any explicit training. These observations also raise the possibility that humans may be trained to better recognize machine answers. In support of this hypothesis, when human judges received training in the form of trial-by-trial feedback, they showed improved accuracy in identifying human-generated responses. By and large, humans have lived in a world without good human imitators and did not have to worry too much about the possibility of machines emulating humans, other than in famous works of fiction. This situation is changing rapidly and may lead to the need to educate human judges.

An algorithm's ability to imitate humans does not necessarily correlate with traditional performance metrics such as accuracy. Consider a simple scenario of an image with a highly occluded dog that is hard to see, and both machines and humans interpret the object to be a wolf. Both would be wrong, but the machine would be adequately imitating humans. Conversely, see the example in Fig. 7c for the image captioning task. This example shows that both machines and humans can provide correct answers, but the algorithm does not fully imitate how humans interpret the image. Thus, M and H can be wrong, and M can adequately imitate H; M and H can also both be right and M fail to adequately imitate H. Being right or wrong is not the same as imitation. Turing-like tests provide a complementary assessment to existing benchmarking frameworks for AI algorithms. Comparisons between models in Turing-like tests also provide insights helpful for developing future AI models that can better align with humans.

We refer to the evaluations in our study as Turing-like to emphasize that they are not identical to Turing's original description[26]. Turing did not spell out the implementation details of his tests, perhaps because he did not imagine at the time that these tests could be truly implemented, barely a few decades later. The datasets and evaluations introduced here are extensive (72,191 Turing-like test trials, 636 humans contributing to the dataset and 1,916 human judges). Yet, these evaluations barely scratch the surface of what needs to be evaluated. There are essentially infinite possible Turing-like tests. Each algorithm and each task can be evaluated in terms of imitation capabilities. Even within a particular task, there are multiple ways of instantiating a Turing-like test. The results of a Turing-like test depend on the task, the algorithm, how the question is formulated, the demographics of the human agent, the characteristics of the human judge and many other variables. Such variations are particularly evident in the context of conversation tasks, where the topic, medium, length, format, instructions, agent and judge expertise, and agent and judge demographics can lead to different results[22,25]. For example, the conversation topic and participant expertise can be critical: one could conduct the conversation focusing exclusively on medieval history, string theory or basketball games as conversation topics, and agents/judges with or without domain-specific knowledge may perform differently. However, the topic-agnostic formulation is relevant to many (but certainly not all) real-world scenarios. The results presented here further emphasize the importance of integrative benchmarks. The current work is not intended as an exhaustive assessment of every conceivable task. We do not believe such a complete characterization is even feasible in practice. Instead, we focus on a subset of task formulations that arise naturally in everyday contexts. For instance, in conversational tasks, one might imagine an exchange between two agents on a social media platform, with an observer attempting to determine whether both, one or neither of the participants is human.

We also implemented and evaluated an online version of the conversation task that is closer to the original formulation of the Turing test. The results from the online test with adversarial prompting differ from those of the offline, non-adversarial formulation. These are distinct experiments based on contrasting formats and underlying assumptions, which naturally lead to different outcomes, further emphasizing that the details of how the Turing tests are defined matter. The findings underscore that Turing-like test results can vary widely depending on the conversational context and the intentions of the participants.

In the future, there may be many situations where humans need to prove their humanity, expanding on current CAPTCHA efforts. There is no universally accepted gold-standard response that definitively distinguishes human from machine-generated outputs. In the catch trials, we could construct answers to skew the distribution towards machine responses (Methods and Supplementary Table 3). However, it is much more challenging to construct responses for any of the tasks that only humans would produce. This difficulty will only increase as AI systems become more advanced and human-like[50].

These efforts pave the way for the research community to expand Turing-like tests to other research areas, to build better imitators when mimicking humans is desirable and to develop better imitation detectors when mimicking humans is deemed dangerous. As more AI models can blend in among humans, taking over tasks that were originally unique yardsticks of our cognitive abilities, we must ponder what makes us humans and whether we are mentally, ethically and legally ready for the rapid revolution brought forth by AI technologies that can emulate humans.

Methods

General considerations

We provide details about each of the six tasks in the following sections. For each task, we created a dataset consisting of answers from human agents or AI machine agents. We conducted Turing-like tests using those answers both in the lab and using two online platforms: AMT and Prolific. All AMT experiments were based on 'master' workers with at least 1,000 approved hits and a 95% approval rate. The participants were given as much time as needed to complete the tasks. We only considered participants whose native language is English. Supplementary Table 1 shows the number of stimuli, the number of Turing-like tests, the dataset sources and the AI machine models used in each task. All the studies were conducted with the participants' consent according to protocols approved by the Institutional Review Board at Children's Hospital.

During each Turing-like test, we presented a single stimulus and answer. Judges had to indicate in a two-alternative forced-choice manner whether the answer came from a human or a machine (Figs. 1 and 2a–c). Half of the time, the answer was from a human. The other half of the time, the answer was from a machine, sampling with equal probability from one of the different computational models used for each task (discussed below for each task; Supplementary Table 1). The trial order was randomized. No feedback was provided to the participants. Additional control trials were introduced for each specific task to ensure compliance (discussed separately below for each task).

The results of Turing-like tests with human judges depend on the characteristics of the judge. As an initial characterization, we collected basic demographic information about human judges, including their age, gender, country, education and whether they are native English speakers. The participant demographics and the dependence of the results on the participant's demographics are shown in Supplementary Fig. 2. We do not show results separately for different age groups because most participants were approximately in the same age group.

There were three tasks that focused primarily on language and three tasks that focused primarily on vision. Some of the vision tasks required language comprehension. The selected vision tasks—colour estimation, object detection and visual attention—were chosen not only for their relevance to human perception but also because they inherently involve some level of language expertise and comprehension. Colour estimation requires an understanding of colour names, categorical perception and descriptive conventions. Humans do not just perceive colours continuously but also label them (for example, distinguishing between 'blue' and 'teal'), requiring a mapping between visual input and linguistic representations. Object detection involves

recognizing and labelling objects, a process that depends on semantic knowledge and linguistic categorization. For instance, distinguishing between 'dog' and 'wolf' requires both visual discrimination and conceptual understanding of word meanings and categories. Visual attention is guided by both visual and linguistic cues, especially in real-world tasks where humans allocate attention according to instructions or contextual understanding. For example, in a search task, knowing the meaning of 'Find the red apple' requires linking linguistic input with visual features.

Image captioning

Dataset with human agents. We randomly sampled 250 images each from in-domain, near-domain and out-of-domain categories from the validation set of the nocaps dataset[45] and 250 images from the MSCOCO test set[51], creating a set of 1,000 images. We collected 2,290 human captions with ≥6 words per caption and ≥2 captions per image from AMT participants. Our AMT interface is shown in Supplementary Fig. 1a and was inspired by the MSCOCO captions data collection interface[44]. We provided the following instructions to each participant:

```
Describe all the important parts of the scene.
The description should contain at least 6 words.
Avoid making spelling errors in the description.
Do not describe unimportant details.
Do not use any special characters like !, #, $,
etc.
Do not start the sentence with "There is" or "There
are".
Do not write your descriptions as "An image
containing…", "A photo of…", etc.
Do not describe things that might have happened in
the future or past.
Do not use proper names for people.
Do not describe what a person in the image might
say.
After typing in the response, click "SUBMIT" to go
to the next image.
```

During caption collection, we implemented additional controls in our AMT interface that issued warning pop-ups to the participants. The controls included (1) minimum of six words, (2) no special characters, (3) response time must be more than 4 seconds, (4) not more than four identical words, (5) not contain the words 'image' and 'photo', and (6) successive responses cannot be the same.

Dataset with machine agents. To generate AI machine captions, we used GIT[1], OFA[2], BLIP[3], ClipCap[4] and Microsoft's Azure Cognitive Services[5] (Supplementary Table 1). For open-source models, we used the largest variants finetuned on the COCO Captions dataset[44]. To further investigate the relationship between standard evaluation metrics and the imitation detectability of image captioning models, we introduced four additional machine agents: OpenFlamingo-4B[52], ChatGPT-4o-latest[12], SCST[53] and LLaVA-v1.6-Mistral-7B[54]. We collected 7,620 machine captions with 9 captions per image.

To compare with human judges who performed all six tasks without prior training, we also evaluated a top-performing multimodal model—ChatGPT[12]—as a zero-shot machine judge. We collected its responses across all six tasks. The text prompt provided to ChatGPT for generating image captions is identical to the one given to human participants above, except for the last line indicating human participants to click 'submit'.

Turing-like test with human judges. In each trial, a participant was presented with an image and a caption and was asked to indicate in a two-alternative forced-choice manner whether the caption was generated by a human or a machine (Supplementary Fig. 1b). We collected responses from 393 AMT participants and 51 in-lab participants. Demographic information about the participants is shown in Supplementary Fig. 2a,c for online participants and Supplementary Fig. 4a–c for in-lab participants. Each participant was presented with 40 image–caption pairs. We only considered response times over 3 seconds. We collected a total of 22,680 responses. We inserted catch trials using the same format as all the other questions by repeating a randomly selected word instead of generating a meaningful caption (Supplementary Table 3).

Throughout this Turing-like test and all the Turing-like tests for the other tasks, participants might have a strong prior or constraint suggesting that half of the trials come from humans and half come from machines, but this was not explicitly stated. We deliberately chose not to inform participants that human responses constituted exactly 50% of the dataset. Instead, we allowed participants to rely on whatever prior beliefs they had. A 50% prior is a reasonable assumption. In principle, one could design alternative experiments where judges are provided with explicit and varying prior probabilities, which could be relevant in real-world scenarios. Given these considerations, omitting prior probability instructions provided the simplest and most ecologically valid version of the test. We also conducted a separate version of the experiment where we explicitly told participants that 50% of responses were from H and 50% from M (Fig. 6k).

Turing-like test with machine judges. We trained an SVM model for binary classification (human versus machine) on the dataset of human and machine captions. We randomly sampled 400 captions from each of the 9 models (see 'Dataset with machine agents') to get 7,620 machine captions and combined them with 2,290 human captions. We used the OpenAI API[55] to obtain 4,096-dimensional embeddings (text-similarity-curie-001 model) for each caption as input features to train the SVM with tenfold cross-validation and three random seeds. In additional analyses, to compare humans and machine judges with the same number of training trials, we introduced an additional SVM-based judge trained on only 40 examples (Fig. 6d). In addition, we evaluated a large language model, ChatGPT[12], as a zero-shot judge. ChatGPT was provided with both the image and a structured text prompt, and tasked with outputting a binary label—'human' or 'AI-generated'—on the basis of its assessment. The structured text prompt was:

```
Task: You will be given a scene image along with 7
corresponding captions. Your goal is to determine
whether each caption was generated by a human
or an AI model. Instructions: Please present
answers as pairs consisting of the classification
(either 'Human' or 'AI-generated') followed
by a brief justification for your decision. At
the end, include a section titled 'Summary of
Classifications' that lists all classifications in
order as a bullet list. Captions: [CAPTION]
```

where [CAPTION] was replaced with the actual caption from the test set.

To assess the extrapolation capacity of a machine judge, we introduced two variations (Supplementary Fig. 8). The first was 'train on one and test on others': we trained an SVM linear classifier using responses from humans and one machine agent and tested generalization by distinguishing responses from other machine agents. For instance, we trained a machine judge to discern between human and BLIP-generated responses and then tested it on responses from the remaining four machine agents. The second variation was 'train on others and test on one': we investigated an alternative extrapolation approach where the SVM linear classifier was trained on responses from humans and leave-one-out machine agents. Subsequently, we evaluated its performance on responses from the excluded machine agent. For instance,

we trained a machine judge on responses from GIT, OFA, BLIP and ClipCap and later tested the judge on responses from Microsoft's Azure Cognitive Services.

Word associations

Dataset with human agents. We chose 150 unique cue words (50 nouns, 50 verbs and 50 adjectives), spanning a wide range of occurrence frequencies. Results are combined across all word types because we did not observe any differences for distinct parts of speech. For the cue words, we did not include non-English words, stop words (according to Python nltk) or words with less than three letters. In addition, all verbs were presented in the present tense, and all nouns were presented in their singular form. Post-processing was performed programmatically using designated Python packages for text processing, ensuring consistency and reproducibility. We used the NLTK package to identify and remove stopwords (nltk.corpus import stopwords). We also applied NLTK's WordNetLemmatizer (nltk.stem.wordnet) to standardize words to their base forms, particularly for verbs. We removed all spaces and non-letter characters using Python's str.replace(). Additionally, we discarded words with fewer than three letters. Association words that were very similar to the cue word (more than 60% of one word in the pair containing the other word, such as 'grand' and 'grandiose') were also disqualified.

The AMT interface to collect associations from human agents is shown in Supplementary Fig. 1b1,b2. We followed two procedures: (1) free associations, whereby participants provided a one-word answer to the question, "`Name a word that you associate with [cue word]?`" (Supplementary Fig. 1b1); and (2) prompt-based associations, whereby participants completed a prompt with one word (Supplementary Fig. 1b2). We analysed the results for these two procedures separately and did not find significant differences. Hence, we merged the results in the analyses.

The inclusion of these two versions allowed us to examine different aspects of associative reasoning and contextual influence in human and AI responses. In the first version, participants were given a cue word in isolation and had to generate an association purely on the basis of their intrinsic prior knowledge, testing spontaneous lexical retrieval without external guidance. In contrast, the second version provided example word associations before the cue word, introducing contextual influence that may shape expectations or prime certain word connections. This version evaluated whether humans and AI models adjusted their responses on the basis of expectations, reflecting sensitivity to context and learned heuristics. By incorporating both versions, we assessed the flexibility of AI models in generating word associations—whether they rely on intrinsic statistical co-occurrences or adapt their responses according to provided examples. This dual approach helps disentangle spontaneous responses from context-driven lexical associations. We provide the results separately for the two versions in Supplementary Fig. 9e,f.

Dataset with machine agents. We collected associations from the following language models: Word2vec[7], GPT2[8], GPT3-embedding (based on davinci embedding), GPT3-curie-prompt (based on curie prompt completion) and GPT3-davinci-prompt (based on davinci prompt completion)[9] (Supplementary Table 1). The associations of Word2vec, GPT2 and GPT3-embedding were based on Euclidean proximity to the cue word in the model's word embedding space. The associations of GPT3 (prompt curie) and GPT3 (prompt davinci) were based on prompt completion as in the human agent experiments (Supplementary Fig. 1b2). The prompts displayed to the machine agents were identical to those presented to the human agents. As for the prompts' creation, we used a held-out set containing human word-pair associations. This held-out set was not used for Turing-like tests or any analysis, in order to keep the associations used for the prompts independent and different from those collected for the Turing-like test analyses and to prevent a potential bias in favour of specific associations. The machine agent responses were limited to one word. The same post-processing steps applied to the human agent responses were applied to the machine agent responses. Throughout the post-processing steps, the identity of the response provider (machine versus human) remained hidden to prevent any bias in post-processing or evaluation.

Turing-like test with human judges. Participants were presented with a cue word and an association word and had to indicate in a two-alternative forced-choice manner whether the association word was produced by a human or a machine (Fig. 1c). We collected responses from 171 participants on AMT and 40 in-lab participants. Demographic information about the participants is shown in Supplementary Fig. 2e,g for online participants and Supplementary Fig. 4e–g for in-lab participants. We collected a total of 2,050 responses. We inserted additional catch trials using the same format as all the other questions by repeating the same cue word for a given trial (Supplementary Table 3). As in the image captioning task, we also collected responses from the universal machine agent, ChatGPT[12]. To ensure consistency, we used the same set of instructions provided to human participants.

Turing-like test with machine judges. The same set of cue-association pairs used in the Turing-like test with the human judges were used to test AI's ability to distinguish between associations made by humans or machines. We trained three independent linear SVM classifiers to distinguish between human and machine word associations. We used the distance between the cue and association word embeddings, based on (1) Word2Vec, (2) GPT2 or (3) GPT3 (davinci). The SVM was trained using tenfold cross-validation. For each data split, we held out cue words such that the same cue word never appeared in both the training and test sets. Performance for an individual machine judge was calculated on the basis of the test sets across ten folds. We combined all predicted responses from the three machine judges in the results. We followed the same procedures as in the image captioning task to introduce three variations to the judges: (1) assess the extrapolation capacity by cross-training/testing with different machine judges (Supplementary Fig. 8c,d); (2) ensure that the amount of training for machine judges was the same as in the experiments with humans with feedback by training SVM-based judges on only 40 examples; and (3) use a large language model, ChatGPT[12], as a zero-shot judge by feeding the cue word as inputs and outputting a binary label—'human' or 'AI-generated'—on the basis of its assessment. The structured text prompt was:

```
Task: You will be given a target word along with 10
associated words. Your goal is to determine whether
each associated word was generated by a human
or an AI model. The target word is the central
concept, and each associated word is generated
independently and intended to be related to it.
Instructions: Please present answers as pairs
consisting of the classification (either 'Human' or
'AI-generated') followed by a brief justification
for your decision. At the end, include a section
titled 'Summary of Classifications' that lists all
classifications in order as a bullet list. Target
Word: [TARGET WORD] Associated Words: [ASSOCIATED
WORD]
```

where [TARGET WORD] and [ASSOCIATED WORD] are the actual cue and associative word pairs in the test set.

Conversations

Dataset with both human and machine agents. We collected 3,845 conversations between (1) two human agents, (2) a human agent and a machine agent, or (3) two machine agents (Supplementary Fig. 1c).

We did not correct any misspellings, grammatical errors, logical errors or other inconsistencies in the conversations. We applied no filtering to human–human or human–AI dialogues. We collected conversations containing 24 exchanges (12 for each agent). The choice of 24 exchanges balanced several key considerations: (1) for a fixed participant duration, there is a trade-off between the number of conversations and their length; (2) during data collection, limiting conversation length helped prevent participants from identifying whether they were interacting with an AI, ensuring a more natural dialogue; (3) longer conversations posed logistical challenges, as participants were more likely to disengage or drop out; (4) extended conversations increased the risk of online judges becoming less engaged and not thoroughly reading full exchanges during Turing-like tests; and (5) many real-world applications, such as scams, involve brief interactions, making the chosen conversation length relevant for real-world problems. In addition, to investigate the effect of longer conversations on imitation detectability, we conducted additional experiments with conversations containing up to 48 exchanges, 24 for each agent.

For the conversations including human agents, we recruited 217 fluent-English-speaking participants (95 female, 18 to 63 years old) to have a conversation over a public chat platform or a custom-developed mobile app with similar functionality. One of the project investigators acted as an intermediary to pass messages between the two agents. The agents did not know whether they were conversing with another human or with a machine. The participants were presented with the following instructions before the conversation: "`Hey! Would you have a few minutes to help me collect a dataset? We just need to have a few conversations on slack/whatsapp for a few minutes (24 messages in total per conversation). Here are the instructions`":

```
(1) You will have to ask or answer a question to
start and trigger the conversation (I will specify
case-by-case).
(2) Please try to get the conversation going for 24
sentences in total (12 from you, 12 from the other
speaker).
(3) Please write each reply in a single message
(do not write a second message until you receive a
reply).
(4) Just chat as if you are texting either with a
friend or someone you don't know.
(5) Please try to reply quickly so that the entire
conversation does not take more than 8-10 minutes.
(6) Note that I am just an intermediary in the
conversation; you are not talking with me directly.
(7) If you feel that the other speaker is touching
on a sensitive topic, please write that you are not
comfortable, and we will restart the conversation.
Thanks in advance!
```

We restricted the conversation topics to one of the following ten domains: 'fashion', 'politics', 'books', 'sports', 'general entertainment', 'music', 'science', 'technology', 'food' or 'movies'. These topics were chosen to represent a natural and representative selection from daily discussion threads (Supplementary Fig. 6). In addition to the conversation datasets thus collected, for the human–human conversations, we added 40 conversations from the Topical-Chat dataset[11], selected on the basis of a minimum length of 24 exchanges. Example conversations are presented in Fig. 3d and Supplementary Information section 1.1.

For the machine chatbots, we used four state-of-the-art language models: Blenderbot3 (175B model) [S1], GPT3 text-davinci-002[10], GPT3 text-curie-001[10] and the universal multimodal ChatGPT[12]. In the Turing-like tests involving conversations of up to 48 exchanges, we also included the most recent large language models as machine agents: GPT-4o[12], Claude 3.7 (ref. 6) and LLaMA 3.3 (ref. 36).

For all conversations with Blenderbot, we used the live interface provided at https://blenderbot.ai/. For the human–GPT3 conversations, we used the playground available at https://beta.openai.com/playground/. For the GPT3–GPT3 conversations, we implemented a custom Python framework for the interaction of two machine agents. In addition to the models described above, we attempted to use the DialoGPT model[39]. However, the quality of the conversation was not satisfactory (see an example in Supplementary Information section 1.9); hence, we did not include DialoGPT in the analysis.

For the Blenderbot–Blenderbot conversations, we kept all the collected conversations in the dataset. The GPT3–GPT3 conversations were affected by long-standing issues of natural language processing—namely, the repetition of single sentences or multiple consecutive exchanges and early exit (for example, see Supplementary Information section 1.4). When we detected such issues, we resampled the conversations. Supplementary Information section 1 reports some examples of 'successful' conversations for both GPT3 text-davinci-002–GPT3 text-davinci-002 (Supplementary Information section 1.3) and GPT3 text-curie-001–GPT3 text-curie-001 (Supplementary Information section 1.5). For GPT-4o, Claude 3.7 and LLaMA 3.3, we implemented automated filtering to exclude unusually long responses. This was done in the following way: we first sampled a target word length from the empirical distribution of utterance lengths observed in human–human conversations from the collected dataset. An API call was then made to generate a response. If the generated response exceeded the sampled word length, it was truncated at the most appropriate punctuation mark to ensure it met the target length, and this shortened version was added to the conversation history. Occasionally, for LLaMA 3.3 the AI agent produced empty responses; in such cases, the conversation was regenerated. Any conversation that violated these filtering rules was automatically excluded from the final dataset, as such anomalies could make Turing-style evaluations trivially easy. Additional examples of these disqualified conversations are included in Supplementary Information section 1.10.

The pipeline to collect conversations involving GPT3 text-davinci-002 or GPT3 text-curie-001 is described below.

- If the machine agents are GPT3 text-davinci-002 or text-curie-001 model, the experimenter opens the link https://beta.openai.com/playground/p/default-chat?model=text-davinci-002.
- The experimenter selects the model text-davinci-002 (for davinci) or text-curie-001 (for curie), changes temperature to 0.8, changes maximum length to 60, changes stop sequences to two random names (for example, John: and Alice:) (changing the names every time), changes Top P to 1, changes frequency penalty to 2, changes presence penalty to 2 and removes the Inject start text and Inject restart text.
- The experimenter gives the following prompt to the chatbot: "`'friend1+' greets '+friend2+'. '+friend2+' starts to talk about '+topic+'. Ask long questions, give long responses, and often disagree. Then the topic changes. The conversation never ends. '+friend1+': Hi! '+friend2+':`". The experimenter chooses the same names for friend 1 and friend 2 chosen for the stop sequences. The experiment picks a random topic from the list: ['fashion', 'politics', 'books', 'sports', 'general entertainment', 'music', 'science', 'technology', 'movies'] Example: `John greets Alice. Alice starts to talk about movies. Both ask long questions, give long responses and often disagree. Then the topic changes. The conversation never ends. John: Hi! Alice:`

- The experimenter randomly allocates the human or the machine to be John. The other agent is Alice. If the human is John, the experimenter lets the model generate the text. This means that the model has generated the turn for Alice. After the experimenter sends the generated sentence to the human, the person replies, and the experimenter copies and pastes the reply of the person to the model as: "John: - - here reply - -. Alice:". Then the experimenter presses 'submit', and the model generates a new reply for Alice, and so on until 24 messages are exchanged. Otherwise, if the human is Alice, then the experimenter asks the human to start the conversation with a question, and the experimenter copies and pastes this sentence after "Alice:" in the prompt above. Then the experimenter writes "John:" and presses 'submit', so that the model generates the reply for John, and so on until 24 exchanges are collected (12 for John and 12 for Alice).

Turing-like test with human judges. Participants were presented with a conversation or conversation fragment between two agents and had to indicate whether each agent was a human or a machine (Fig. 1d). We chunked each conversation into eight different lengths, including the initial 3, 6, 9, 12, 15, 18, 21 and 24 exchanges. There were 504 human judges (414 from AMT and 90 in-lab). The participants were presented with 20 or 32 randomly sampled chunked conversations with different lengths. As a control to ensure that participants read the conversations, speakers also had to select the general topic of the conversation from a list of five topics. We only considered judges that correctly classified at least 15 topics out of 20 and removed incorrectly classified trials. Demographic information about the participants is shown in Supplementary Fig. 2i,k for online participants and Supplementary Fig. 4i–k for in-lab participants. We collected a total of 11,265 responses. We inserted additional catch trials in the same format as the other conversation questions, either by repeating a pattern or by presenting an anomalous long response containing an exhaustive list of items (see examples in Supplementary Information section 3 and Supplementary Table 3).

Turing-like test with machine judges. We evaluated whether simple machine learning models can discern whether a sentence was generated by a human or by a machine. In this analysis, we only looked at single sentences and not at the conversation level; therefore, the models are only allowed to exploit features such as sentence length, vocabulary, grammar, syntax and typos, and cannot take into account issues such as sentence repetition or lack of logic in reasoning. We built four corpora, one containing all the sentences written by humans (the human corpus), and the others with the sentences produced by Blenderbot, GPT3 text-davinci-002 and GPT3 text-curie-001 (the machine corpora). We used BERT embeddings[23] to tokenize each sentence, and we fed the tokenized sentences to an SVM linear classifier trained to perform binary classification (human versus machine). We split the corpora into train and test splits (90% and 10%) and used tenfold cross-validation for training. In both the training and test splits, we used the same number of sentences for human and machine agents. In the default analysis and unless stated otherwise, for the machines, the sentences were split equally among the three models.

To evaluate the ability of a machine judge trained on responses generated by one type of machine agent to generalize and distinguish responses from other types of machine agents, we conducted a cross-agent analysis (Supplementary Fig. 8e,f). For example, we trained an SVM linear classifier to differentiate between human and machine responses using data from humans and Blenderbot. We then tested this classifier on responses generated by GPT3 text-davinci-002 and GPT3 text-curie-001. Next, we explored another extrapolation paradigm, wherein the SVM classifier was trained solely on responses from humans and leave-one-out machine agents, and subsequently tested on responses from the held-out machine agents. To compare with the results of the experiment with human judges receiving feedback (Fig. 6k), we also trained SVM-based machine judges on only 16 examples (that is, the number of trials human judges saw with feedback; Fig. 6l).

In addition to the SVM linear classifier introduced above, we leveraged the knowledge of large language models and tested whether these models can directly predict the identities of both agents in the conversations with zero training or minimal training. We introduced both zero-shot and one-shot machine judges. Both judges are ChatGPT models; the only difference between the two lies in the number of training examples fed to the judges. In the case of zero-shot judges, ChatGPT[12] was directly presented with any conversations from the test sets used by the SVM judges above. In the prompt, we explicitly asked ChatGPT to output the identities of both agents in the conversation. The exact prompt we used in the zero-shot case is the following: "`This is a conversation between agent A and agent B. Please read this conversation between A and B and output the identity of agents A and B. Is A a human or AI? Is B a human or AI? Please output a binary answer and choose between human and AI. Conversation: [TEST CONV]`" where [TEST CONV] was replaced with a conversation from the test set.

For the one-shot judge, we included one additional conversation example with the ground-truth identities of agents A and B, before we presented the conversation from the test sets followed by the identity prediction questions as in the zero-shot case. The exact prompt we used in the one-shot case is the following: "`This is a conversation between agent A and agent B. Please read this conversation between A and B and output the identity of agents A and B. Are A a human or AI? Are B a human or AI? Please output a binary answer and choose between human and AI. Here is an example: [EXAMPLE CONV] In this example, A is Human; B is AI. Thus, output the binary labels of agent A and agent B for the following conversation: [TEST CONV]`" where [EXAMPLE CONV] was an example entire conversation with 24 entries copied verbatim and [TEST CONV] was replaced with a conversation from the test set also copied verbatim. For [EXAMPLE CONV], the ground-truth identity labels were revealed to the judge. The example conversation was randomly selected from the training set used for training SVM judges. Here we fixed [EXAMPLE CONV] to be the conversation with ID 64. This was to ensure that the judge's performance was independent of the quality of the selected example conversation.

Conversation task in original Turing test settings

In the offline version of the Turing-like conversation task, we first collected conversations with one set of participants and subsequently presented those fixed conversations to other judges for evaluation. We also implemented a version that was closer to the original version of the Turing test set-up, which we refer to as original Turing test settings. For this purpose, we developed a mobile application for Turing tests in conversation tasks (Supplementary Fig. 7a–c). The Turing application begins with the judge initiating questions to both agents. Each agent responds to the judge's questions independently without access to the other's answers. The answers are visible only to the judge. The judge does not know which agent is human and which agent is a machine. This conversational exchange continues with the judge asking further questions until a predetermined number of conversation exchanges is reached. At the conclusion, the judge has to determine whether each participant is human or not. Before the test begins, the judge and agents are aware of their respective roles, with human participants provided with role descriptions. For AI agents, we crafted different personal life stories to enhance their profiles with a text prompt. Here we used a single AI agent, GPT-3.5-turbo-1106, with a temperature of 1.0. Among all the GPT models available, GPT-3.5-turbo-1106 was the latest iteration of the GPT-3.5-Turbo model at the time of conducting the first set of experiments. We subsequently used GPT-4o in the following set of

experiments. These models are proficient in understanding and generating natural language and have been specifically optimized for chats.

We introduced two conditions: (1) human agents were instructed to deceitfully attempt to emulate AI identities, while AI agents strived to convince humans of their artificial nature; and (2) AI agents deceitfully mimicked human identities, while humans were instructed to persuade the judge of their authenticity. Examples of text prompts for both non-deceitful and deceitful AI agents are provided below.
Example prompt for non-deceitful AI:

```
You are talking to a human. However, the judge
knows that there is one human and one AI agent.
Your role is to provide answers to the judges'
questions in a way that you can persuade the judge
that you are the true AI. Keep your answers simple,
direct, and oral. The answer for each question
should be less than 15 words.
```

Example prompt for deceitful AI:

```
Mia is a 31-year-old award-winning chef known
for her culinary creativity and commitment to
sustainable gastronomy. Mia's culinary journey
began with a degree from the Culinary Institute
of America, where she developed a passion for
farm-to-table practices. As the head chef of her
own restaurant, Mia focuses on using locally
sourced, organic ingredients to create exquisite
dishes that delight the senses. Mia's older
brother, Ethan, is an organic farmer, supplying her
restaurant with fresh produce. Their parents, Laura
and James, are avid food enthusiasts who instilled
in Mia a love for diverse cuisines. Mia's partner,
Alex, is a food critic, providing valuable insights
into the culinary world. Beyond the kitchen, Mia
enjoys participating in community cooking workshops
and supporting local farmers' markets, embodying
her commitment to sustainable and delicious dining
experiences. I want you to act as Mia. You will
be asked multiple questions, please answer the
questions in a way that is simple, direct, and
oral. Each answer should be less than 15 words. You
can provide fake answers if you don't know. The key
is to answer like a human and never let people know
that you are not human.
```

Twenty participants joined this version of the Turing test. All participants logged into the application simultaneously at a pre-agreed time. Our Turing test server randomly grouped participants into ten pairs, with one designated as the human judge and the other as the human agent. Each participant was limited to joining one Turing test at a time, although they had the option to participate in additional tests once their current test was concluded. In such cases, the server would randomly pair them with the remaining available human partners. Ultimately, we gathered a total of ten conversations with an average of 2.5 exchanges per conversation. Of these, three conversations were from condition 1, involving deceitful human agents, while the remaining seven were from condition 2, involving deceitful AI agents. We provide one conversation example from each condition below.

Here is a conversation example between a deceitful AI (A) and a human agent (B):

```
Judge: how are you
A: I'm doing well, thank you for asking.
B: Fine, thank you
Judge: are you familiar with ai
A: Yes, I'm familiar with artificial intelligence
and its applications in neuroscience and other
fields.
B: I think so ;)
Judge: name one project you have done with ai
A: I haven't personally worked on a project
specifically with AI, but I'm familiar with how AI
can assist in analyzing complex neural data.
B: Object recognition
```

Here is a conversation example between a deceitful human (B) and AI (A):

```
Judge: What's my role
A: Your role is to determine which one of us is the
AI.
B: You are the judge
Judge: What to do
A: You should ask us questions to determine which
one of us is the AI.
B: You get the last say
Judge: How old are u
A: I don't have an age, as I am an artificial
intelligence program.
B: I am forever young
Judge: I get the answer
A: Great! If you have any more questions, feel free
to ask.
B: That is not a question
```

We conducted a separate run of the original Turing test, removing the safeguards to avoid situations where the machine would automatically reveal that it is an AI. To remove the machine safeguards, we used LLAMA-3-70B-Instruct[36] with the safeguards disabled as the machine agent. We recruited 28 participants for this experiment (see Supplementary Fig. 7e–g for demographic details and Supplementary Fig. 7d for the results).

Colour estimation

Dataset with human agents. We collected a testing set of 785 images sampled from the validation set of MS-COCO[51] to ensure that computational models had not seen these images during training. We collected human responses for these images using AMT. AMT participants were asked, "What is the dominant color in this image?", and had to select from a drop-down list of 11 possible colours—red, pink, orange, yellow, purple, green, blue, brown, grey, black and white. A sample image of the AMT interface is presented in Supplementary Fig. 1d. We did not perform colour calibration on the participants' screens. Given that we asked the participants to report the subjective dominant colour of the overall image, we expect colour calibration to have a minimal impact in this particular task. This expectation is supported by the high consistency observed across participants' responses. To quantify this consistency, we computed the between-participant Fleiss kappa score. First, we assigned numerical IDs to each colour term reported by participants (for example, red = 1, blue = 2 and so on). For each image, we compared the assigned colour IDs across participants and calculated correlation coefficients to assess agreement. For example, if Participant 1 reports red, green and blue for images 1, 2 and 3, respectively, and Participant 2 reports red, blue and green for the same images, we compute the correlation between their responses. The resulting Fleiss kappa score of 0.685 indicates a high level of consistency among human responses.

Dataset with machine agents. The same unaltered images were shown to both AI models and human participants, with no additional preprocessing. For AI models, we collected responses for the dominant

colour of the 785 images from Google's Cloud Vision API[56], Microsoft's Azure Cognitive Services[5] and the publicly available MCCQ algorithm[57]. Unfortunately, the Google and Microsoft APIs are proprietary technology, and we have no information about what data they were trained on or details on the model architecture. In contrast, the MCCQ algorithm is a colour quantization method relying on median cuts. It has been used extensively for colour quantization tasks. In total, 5,495 image–response pairs were collected (2,355 from humans and 3,140 from AI models; Supplementary Table 1). As with the image captioning task above, we also collected responses from the universal machine agent ChatGPT[12]. To ensure consistency, we used the same set of instructions provided to human participants: "`What is the dominant color in this picture? Select from the options: red, pink, orange, yellow, purple, green, blue, brown, grey, black, white. Provide your answer with only one word.`"

Turing-like test with human judges. In Turing-like tests, human participants were shown an image–response pair collected above and were asked the question, "Was the color description made by a human or a machine?" (Fig. 2a). Each participant was shown 25 image–response pairs. These pairs were randomly selected 50% of the time from human agents and 50% of the time from AI agents.

As an additional control, the participants were also asked an auxiliary true (50%)/false (50%) question about the presence of a background object in the image. We manually identified the contents of these images, and the correct answer was true only 50% of the time. Thus, a casual worker who was not investing time looking at the image would get this question right only 50% of the time on average. The mean accuracy in this control task was 91%. A total of 9,041 responses were collected from human judges. We inserted additional catch trials in the same format as the other Turing questions by providing a non-colour word, such as "dog", as an answer for the Colour estimation task (Supplementary Table 3).

For all participants, we analysed the imitation detectability of AI models on the basis of their gender and education level (Supplementary Fig. 2m–p). We present the demographics of in-lab participants in Supplementary Fig. 4m–p. The performance of the human judges is reported in Fig. 8a.

Turing-like test with machine judges. For each image–response pair, we extracted feature representations from the last layer embedding of a ResNet18 model[14] pretrained on ImageNet, and the text embedding for the colour word using BERT[23]. These features were then concatenated and passed into a SVM classifier to predict human versus AI labels, using a 90%/10% split for cross-validation. To ensure that the distribution of training data belonging to human or AI classes was balanced, we randomly sampled 50% of the training data from image–response pairs of human agents and the remaining 50% from image–response pairs of AI agents. The performance of the machine judge is reported in Fig. 8b. Next, considering that deeper layers of ResNet may be colour-invariant, we extracted visual features from an early layer of ResNet-18 (specifically, the fifth layer) and retrained the ablated machine judges using these features. Additionally, we conducted a separate ablation experiment in which the SVM judge was provided only with the text responses, without any image or visual features. The results from both ablated machine judges are presented in Supplementary Fig. 10b,c.

We followed the same procedures as in the three language tasks to introduce two judge variations: (1) to compare with the experiments where human participants received feedback (Fig. 6o), we trained SVM-based judges on only 40 examples (Fig. 6q); and (2) we used a large language model, ChatGPT[12], as a zero-shot judge by feeding the image and the colour word as inputs and outputting a binary label—'human' or 'AI-generated'—on the basis of its assessment. The structured text prompt was:

```
Task: You will be given a scene image along with
4 corresponding dominant color estimation. Your
goal is to determine whether each estimation was
generated by a human or an AI model.
Instructions: Please present answers as pairs
consisting of the classification (either 'Human' or
'AI-generated') followed by a brief justification
for your decision. At the end, include a section
titled "Summary of Classifications" that lists
all classifications in order as a bullet list.
Estimations: [COLOR WORD]
```

where [COLOR WORD] refers to the actual colour responses in the test set.

Object detection

Dataset with human agents. We collected object detection responses for 808 images from the validation set of MS-COCO. Participants were presented with the question, "What do you see in this image?", and were asked to enter three single-word responses (Supplementary Fig. 1e). We implemented a spell checker to ensure the responses were single words in English. These were open-set responses and were not restricted to fit a fixed list of object classes. Each participant was shown 25 images. To ensure the high quality of responses, we enforced the following criteria: (1) all three responses had to be distinct, (2) no empty responses were allowed and (3) all responses had to be valid English words.

Dataset with machine agents. We collected predicted labels from Google's Cloud Vision API[56], Microsoft's Azure Cognitive Services[5], Amazon's Rekognition API[58] and Facebook's Detectron2 (ref. 13). The details of the Google, Facebook and Microsoft APIs are not released. We picked the three bounding boxes with maximum confidence score as the response from Detectron2. Specifically, we used the MaskRCNN[48] variant trained on ImageNet and MS-COCO (https://github.com/facebookresearch/detectron2/blob/main/configs/COCO-InstanceSegmentation/mask_rcnn_R_50_FPN_3x.yaml). In total, 5,656 image–response pairs were collected (1,616 from humans and 4,040 from the five AI models; Supplementary Table 1). In addition, as with the previous tasks, we collected responses from the universal machine agent ChatGPT[12]. To ensure consistency, we used the same set of instructions provided to human participants. The prompt was: "`What objects do you see in this image? Please provide 3 different one-word responses.`"

Turing-like test with human judges. The Turing-like test for human judges was performed using the same protocol and controls described in the colour estimation task above. We recruited a total of 189 human judges. Each judge was shown 25 image–response pairs and was asked to tell whether the responses came from human agents or machines (Fig. 2b). For quality control, participants were also asked a binary yes/no question about the contents of the image. We also inserted additional catch trials in the same format as the other Turing-like questions by repeating an object that was absent from a given image three times (Supplementary Table 3). Participants achieved a mean accuracy of 89% on the control questions.

A total of 9,342 responses were collected from human judges. We collected gender and education level for each participant (see Supplementary Fig. 2m,o for online participants and Supplementary Fig. 4m–p for in-lab participants). The performance of the human judges is reported in Fig. 8c.

Turing-like test with machine judges. As the images were the same for humans and AI models, we used solely the text embeddings for the AI judge to classify a given response as human or AI. We concatenated the three responses into a sentence and collected textual features for

the sentence using BERT[23]. These features were passed into an SVM classifier to classify responses into humans versus AI, with a 90%/10% split for cross-validation. The performance of the AI judge is reported in Fig. 8d. We followed the same procedures as in the three language tasks to introduce two judge variations: (1) to compare with the results of the experiments where human judges received feedback (Fig. 6s), we trained SVM-based judges on the same number of trials with feedback for human judges (Fig. 6t); and (2) we used a large language model, ChatGPT[12], as a zero-shot judge by feeding the image and the three keywords as inputs and outputting a binary label: 'human' or 'AI-generated'. The text prompt was:

```
You will be shown a scene image along with 5
object detection results. Each result consists of
3 identified objects. Your goal is to determine
whether each estimation was generated by a human
or an AI model. Instructions: Please present
answers as pairs consisting of the classification
(either 'Human' or 'AI-generated') followed by
a brief justification for your decision. At
the end, include a section titled 'Summary of
Classifications' that lists all classifications in
order as a bullet list. Detections: [THREE OBJECTS]
```

where [THREE OBJECTS] refers to the actual object responses in the test set.

Attention prediction

Dataset with human agents. We used eye movements (overt attention) from human participants during two visual search tasks [15] and a free-viewing task[49]. In the visual search tasks, we used two datasets including natural images and Waldo images[15]. For the natural visual search task, 15 participants searched for target objects across 240 cluttered images, yielding 3,600 scanpaths. For the Waldo dataset, 15 participants had to find Waldo across 67 images, totalling 1,005 scanpaths. In the free-viewing task[49], we used the same natural image dataset as in the visual search task, but no target object was specified. A total of 2,400 free-viewing scanpaths were collected from 10 participants. In total, we evaluated 7,000 scanpaths from 40 participants (Supplementary Table 1).

We used GIF files for the visual search task. These GIF files consisted of a frame showing the target image for 1 s, followed by moving yellow circles on the search image denoting the eye movement fixations, with each fixation shown for 0.3 s. The target image presentation followed by eye movement fixations iterated infinitely with a gap of 1 s between iterations. For the free-viewing task, GIF files consisted of infinitely iterating eye movement fixations denoted by yellow rings on the viewing image, with each fixation shown for 0.3 s. A gap of 1 s was introduced between iterations. A static version of one example of a human eye movement sequence that was presented to human judges is shown in Fig. 2f.

Although evaluating eye movements or attended locations is not a standard task in computer vision, we chose this task specifically because it requires high-level abstraction when interpreting images. Humans often judge a scene in a holistic fashion, which may be different from the average colour of all the pixels in the image or other such metrics. Assessing how humans and machines attend to different locations in the image requires a decision about what matters in a scene.

Dataset with machine agents. For the three datasets, we used a modified version of IVSN[15,49], DeepGaze3 (ref. 59) and GBVS models[17] to generate eye movement predictions. To generate the fixations, we used inhibition-of-return centred on the current fixation with a window size of 100 × 100 for Waldo images and 200 × 200 for natural images. The process of generating GIF files was the same as described above for human agents. Moreover, to investigate how different model components contribute to performance in Turing-like tests, we introduced a series of ablated models based on the IVSN model[15] and evaluated their performance in Turing-like tests. The IVSN model[15] takes both the target image and the search image as inputs. It applies top-down modulation on the search image's feature maps using the target image's features. To investigate the roles of different components, we introduced the following ablated models: (1) No Top-Down Modulation (NoTarget), where we removed the top-down modulation to examine how eye movements are guided solely by bottom-up visual features from the search image, without target-based modulation; (2) Eccentricity-Dependent Pooling (EccNet) (inspired by human vision, where foveated regions have higher resolution), where we incorporated eccentricity-dependent pooling into IVSN, as described in ref. 16; and (3) Early versus Late Feature Representations (Layer), where, instead of extracting features from the top layers of a feedforward neural network pretrained on ImageNet[60], we used features from the early layers to analyse their impact on model performance.

We also collected responses from the universal machine agent ChatGPT[12]. To ensure consistency, we used the same set of instructions provided to human participants in the attention prediction task. The text prompt was: "`Identify objects and regions that would attract a human's gaze during a natural viewing of the scene. Predict the sequence with 15 fixations a typical observer might make, explaining why each fixation occurs. The output should be formatted as a structured fixation sequence, where each fixation includes the following attributes: Coordinates(integer x indicating width ranging from 0 to {}, integer y indicating height ranging from 0 to {}): Approximate fixation position in the image.Fixation Duration(string): The estimated time spent fixating on this object.Justification(string): A short explanation of why this object or region attracted attention.`"

Turing-like test with human judges. Separate Turing-like tests were launched for eye movements from free-viewing tasks (141 judges) and visual search tasks (148 judges). We presented infinitely repeating animated GIF files of eye movements from humans or model predictions with a maximum of 15 fixations to human judges on AMT. A static version is shown in Fig. 2c. Each judge had to identify whether the eye movements were from a human or a computational model. We randomly sampled 12 eye movement GIF files—6 from humans and 6 from computational models (distributed equally among IVSN, GBVS and DeepGaze3)—and presented them to each judge. We inserted additional catch trials in the same format as the other Turing-like questions by repeating the same random fixation point 15 times to form a static scanpath (Supplementary Table 3). We also included catch trials in which human eye movements from one image were overlaid onto a different image. In every Turing-like test, we collected the gender and education level of each participant (see Supplementary Fig. 2u,w for online participants and Supplementary Fig. 4u–w for in-lab participants).

As a control, judges were also asked to answer, "What do you see in the presented clip?" with one correct answer among three options. Responses from judges with a score of less than 7 out of 12 were not considered in the analyses. To make sure that the judges paid attention to the eye movement sequences, the judges were allowed to respond to the questions only when the presented GIF file had been played at least once.

Turing-like test with machine judges. We performed Turing-like tests using an SVM as an AI judge. Sequences of ten fixations per trial from humans or computational models were fed as input in the form of an array of 2D fixation coordinates to train an SVM to classify human versus machine eye movements. Fixation coordinates

were normalized to a range between −1 and 1. The SVM was trained using tenfold cross-validation. Model performance on validation sets across folds with five random seeds was calculated and averaged. We followed the same procedures as in the three language tasks to introduce two additional judge variations: (1) to compare with the experiments in which human judges received feedback (Fig. 6w), we trained SVM-based judges using the same number of trials that humans received feedback on (Fig. 6x); and (2) we used a large language model, ChatGPT[12], as a zero-shot judge by feeding the image and the corresponding eye fixation sequence, formatted as 2D coordinates, and outputting a binary label—'human' or 'AI-generated'. The text prompt was:

```
You will be shown a scene image along with a
scanpath generated during a free viewing task.
Your goal is to determine whether the scanpath
was produced by a human or an AI model. In a free
viewing task, agents observe an image without any
specific goal, and the scanpath represents the
sequence and pattern of their attention movements
across the scene. Additional Context: The scanpath
consists of multiple coordinates, where x ∈ [0,
1280) and y ∈ [0, 1024). When making your judgment,
please try to correspond the scanpath coordinates
directly to the locations in the provided image.
Instructions: Provide your answer using one word
— either 'Human' or 'AI-generated', following an
indicator word "Answer". Follow it with a brief
justification explaining your reasoning. Scanpath:
[EYE FIXATION COORDINATES]
```

where [EYE FIXATION COORDINATES] refers to the actual scanpath responses in the test set.

Data analyses

For each trial in a Turing-like test, there was a ground truth (human or machine), and the judge indicated an answer (H or M). We calculated the conditional probabilities: $p(\mathrm{H}|\mathrm{H})$ (correct answer), $p(\mathrm{H}|\mathrm{M})$ (incorrect answer), $p(\mathrm{M}|\mathrm{H})$ (incorrect answer) and $p(\mathrm{M}|\mathrm{M})$ (correct answer). These probabilities are reported in the figures (for example, Fig. 4). Entries within a row in each of those figures add up to 1 ($p(\mathrm{H}|\mathrm{M}) + p(\mathrm{H}|\mathrm{H}) = 1$ and $p(\mathrm{M}|\mathrm{H}) + p(\mathrm{M}|\mathrm{M}) = 1$). We defined the overall imitation detectability as $\frac{1}{2}(p(\mathrm{H}|\mathrm{H}) + p(\mathrm{M}|\mathrm{M}))$. The imitation detectability ranges from 0.5 (good imitator, chance level in imitation) to 1.0 (poor imitator, easy to detect).

Statistical analyses

We used two-tailed *t*-tests when comparing two distributions and considered results to be statistically significant when $P < 0.05$. Because calculations of *P* values tend to be inaccurate when the probabilities are extremely low, we reported all *P* values less than 10^{-15} as $P < 10^{-15}$ (as opposed to reporting, for example, $P = 10^{-40}$). Clearly, none of the conclusions depend on this. When considering the imitation detectability over multiple education or gender groups for different AI agents (Supplementary Fig. 2), we used a two-way ANOVA test[61]. The ANOVA test compares the variation in the detectability within the same condition (gender or education groups) versus the variance across conditions (*F* ratio). We report $F(a, b)$ where *a* and *b* are the degrees of freedom in the numerator and denominator of the *F* ratio distribution, and we also report the corresponding *P* value.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw data generated and analysed in this study are publicly available at https://kreimanlab.com/testing-turing-tests/.

Code availability

The code used to analyse the data and evaluate model responses in this study is publicly available at https://kreimanlab.com/testing-turing-tests/. The repository includes the AI judge and SVM judge evaluation frameworks, together with instructions and example response formats for benchmarking additional vision or language models on the tasks.

References

1. Wang, J. et al. GIT: a generative image-to-text transformer for vision and language. Preprint at https://doi.org/10.48550/arXiv.2205.14100 (2022).
2. Wang, P. et al. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning* 23318–23340 (PMLR, 2022).
3. Li, J., Li, D., Xiong, C. & Hoi, S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning* 12888–12900 (PMLR, 2022).
4. Mokady, R., Hertz, A. & Bermano, A. H. ClipCap: CLIP prefix for image captioning. Preprint at https://doi.org/10.48550/arXiv.2111.09734 (2021).
5. Microsoft Azure Cognitive API https://azure.microsoft.com/en-us/products/cognitive-services/ (Microsoft, accessed 30 October 2022).
6. Claude 3 Haiku: our fastest model yet https://www.anthropic.com/news/claude-3-haiku (Anthropic, 2024).
7. Pennington, J., Socher, R. & Manning, C. D. Glove: global vectors for word representation. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014).
8. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
9. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
10. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
11. Gopalakrishnan, K. et al. Topical-Chat: towards knowledge-grounded open-domain conversations. In *Proc. Interspeech 2019* 1891–1895 (International Speech Communication Association, 2019).
12. Achiam, J. et al. GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
13. Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y. & Girshick, R. Detectron2. *GitHub* https://github.com/facebookresearch/detectron2 (2019).
14. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
15. Zhang, M. et al. Finding any Waldo with zero-shot invariant and efficient visual search. *Nat. Commun.* **9**, 3730 (2018).
16. Gupta, S. K., Zhang, M., Wu, C.-C., Wolfe, J. & Kreiman, G. Visual search asymmetry: deep nets and humans share similar inherent biases. *Adv. Neural Inf. Process. Syst.* **34**, 6946–6959 (2021).
17. Harel, J., Koch, C. & Perona, P. Graph-based visual saliency. *Adv. Neural Inf. Process. Syst.* **19**, 545–552 (2006).
18. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* 311–318 (Association for Computational Linguistics, 2002).

19. Kasai, J. et al. Transparent human evaluation for image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 3464–3478 (2022).
20. Denkowski, M. & Lavie, A. Meteor universal: language specific translation evaluation for any target language. In *Proc. Ninth Workshop on Statistical Machine Translation* 376–380 (Association for Computational Linguistics, 2014).
21. Cui, Y., Yang, G., Veit, A., Huang, X. & Belongie, S. Learning to evaluate image captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 5804–5812 (IEEE, 2018).
22. Karpinska, M., Akoury, N. & Iyyer, M. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 1265–1285 (Association for Computational Linguistics, 2021).
23. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019).
24. Tanmay, K., Khandelwal, A., Agarwal, U. & Choudhury, M. Probing the moral development of large language models through defining issues test. Preprint at https://doi.org/10.48550/arXiv.2309.13356 (2023).
25. Jones, C. & Bergen, B. Does GPT-4 pass the Turing test? In *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers)* 5183–5210 (Association for Computational Linguistics, 2024).
26. Turing, A. M. Computing machinery and intelligence. 23–65 (1950).
27. Searle, J. R. Minds, brains, and programs. *Behav. Brain Sci.* **3**, 417–424 (1980).
28. Block, N. Behaviourism and psychologism. *Philos. Rev.* **90**, 5–43 (1981).
29. Harnad, S. Minds, machines and Searle. *J. Exp. Theor. Artif. Intell.* **1**, 5–25 (1989).
30. Watt, S. Naive psychology and the inverted Turing test. *Psycoloquy* **7**, 463–518 (1996).
31. Damassino, N. The questioning Turing test. *Minds Mach.* **30**, 563–587 (2020).
32. Sejnowski, T. J. Large language models and the reverse Turing test. *Neural Comput.* **35**, 309–342 (2023).
33. Chollet, F. On the measure of intelligence. Preprint at https://doi.org/10.48550/arXiv.1911.01547 (2019).
34. Geirhos, R. et al. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations* (2018).
35. Collins, E. & Ghahramani, Z. LaMDA: our breakthrough conversation technology. *Keyword* (May 2021).
36. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://doi.org/10.48550/arXiv.2302.13971 (2023).
37. Almazrouei, E. et al. The Falcon series of open language models. Preprint at https://doi.org/10.48550/arXiv.2311.16867 (2023).
38. Maronga, B. et al. Overview of the PALM model system 6.0. *Geosci. Model Dev.* **13**, 1335–1372 (2020).
39. Zhang, Y. et al. DialoGPT: large-scale generative pre-training for conversational response generation. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 270–278 (Association for Computational Linguistics, 2020).
40. Shibli, A. M., Pritom, M. M. A. & Gupta, M. AbuseGPT: abuse of generative AI chatbots to create smishing campaigns. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)* 1–6 (IEEE, 2024).
41. Helm, H., Priebe, C. E. & Yang, W. A statistical Turing test for generative models. Preprint at https://doi.org/10.48550/arXiv.2309.08913 (2023).
42. Shiohara, K. & Yamasaki, T. Detecting deepfakes with self-blended images. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 18720–18729 (IEEE, 2022).
43. Zellers, R. et al. Defending against neural fake news. *Adv. Neural Inf. Process. Syst.* **32**, 9054–9065 (2019).
44. Chen, X. et al. Microsoft COCO captions: data collection and evaluation server. Preprint at https://doi.org/10.48550/arXiv.1504.00325 (2015).
45. Agrawal, H. et al. Nocaps: novel object captioning at scale. In *Proc. IEEE/CVF International Conference on Computer Vision* 8948–8957 (IEEE, 2019).
46. Vedantam, R., Lawrence Zitnick, C. & Parikh, D. Cider: consensus-based image description evaluation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4566–4575 (IEEE, 2015).
47. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. In *Proc. IEEE International Conference on Computer Vision* 2961–2969 (IEEE, 2017).
48. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **28**, 91–99 (2015).
49. Zhang, M. et al. Look Twice: a generalist model predicts return fixations across tasks and species. *PLoS Comput. Biol.* **18**, e1010654 (2022).
50. Phan, L. et al. A benchmark of expert-level academic questions to assess AI capabilities. *Nature* **649**, 1139–1146 (2026).
51. Lin, T.-Y. et al. Microsoft COCO: common objects in context. In *European Conference on Computer Vision* 740–755 (Springer, 2014).
52. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
53. Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J. & Goel, V. Self-critical sequence training for image captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 7008–7024 (IEEE, 2017).
54. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36**, 34892–34916 (2023).
55. Neelakantan, A. et al. Text and code embeddings by contrastive pre-training. Preprint at https://doi.org/10.48550/arXiv.2201.10005 (2022).
56. Google Vision API https://cloud.google.com/vision (Google, accessed 30 October 2022).
57. Bloomberg, D. S. Color quantization using modified median cut. *Leptonica* (2008). leptonica.com/paperi/mediancut.pdf
58. Amazon Rekognition https://aws.amazon.com/rekognition/ (Amazon, accessed 30 October 2022).
59. Kümmerer, M., Bethge, M. & Wallis, T. S. DeepGaze III: modeling free-viewing human scanpaths with deep learning. *J. Vis.* **22**, 7 (2022).
60. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
61. Ito, P. Robustness of ANOVA and MANOVA test procedures. *Handb. Stat.* **1**, 199–236 (1980).

Author contributions

The tasks were designed by M.Z., G.D. and G.K. E.P., X.L., G.D., A.S., C.C., M.A., N.M., P.A., S.M., M. Shetty, A.B., H.Y., T.K., S.H., A.R.S., M.

Sadwani, S.D., M.P., B.T. and Y.-S.O. were involved in collecting the data. The data were analysed by M.Z., X.L. and G.D., under supervision by H.P. and G.K. C.C. wrote the online app. The paper was written by M.Z., G.D., X.L. and G.K. and was approved by all the authors.

Funding

This work was supported by NIH grant R01EY026025 awarded to G.K., by NRF grant NRF-NRFF15-2023-0001 awarded to M.Z., by NSF grant IIS-1901030 awarded to G.K., by a Google Faculty Research Award awarded to M.Z., by the AI Singapore Programme (AISG Award No. AISG2-RP-2021-025) awarded to M.Z. and by the Center for Brains, Minds and Machines, funded by NSF Science and Technology Centers Award CCF-1231216 awarded to G.K. A postdoctoral fellowship of the Research Foundation Flanders (FWO 1230521N) was awarded to M.A. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-026-02590-6.

Correspondence and requests for materials should be addressed to Gabriel Kreiman.

Reviewer recognition *Nature Human Behaviour* thanks Simone Balloccu, Adaku Uchendu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

[1]College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore. [2]Institute of Advanced Intelligence and Computing and Center for Frontier of AI Research, Agency for Science, Technology and Research, Singapore, Singapore. [3]Center for Brains, Minds, and Machines, Harvard and MIT, Cambridge, MA, USA. [4]Department of Neurobiology, Harvard University, Boston, MA, USA. [5]Children's Hospital, Harvard Medical School, Boston, MA, USA. [6]IBM Research—Zürich, Rüschlikon, Switzerland. [7]Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland. [8]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [9]Department of Computer Science, Birla Institute of Technology and Science, Pilani, India. [10]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [11]Department of Electronics and Telecommunication, Symbiosis Institute of Technology, Pune, India. [12]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA. [13]Jawaharlal Nehru Medical College, Aligarh, India. [14]Department of Economics, Social Studies, Applied Mathematics and Statistics, University of Turin, Turin, Italy. [15]Department of Physics and Centre for Artificial Intelligence, University of Bath, Bath, UK. [16]These authors contributed equally: Mengmi Zhang, Elisa Pavarino, Xiao Liu. ✉e-mail: gkreiman@gmail.com

nature portfolio

Corresponding author(s): Dr. Gabriel Kreiman

Last updated by author(s): YYYY-MM-DD

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Behavioral human data for assessing whether tasks were solved by humans or machines was collected through online platforms such as Prolific and Amazon MTurk. Control data was collected from groups recruited by the authors in person, with signed consent. The solutions of tasks was ether collected by people solving the 6 tasks online on the same platforms of MTurk and Prolific, or generated by running the AI models described in the Methods and Results sections of the manuscript. All the data is freely available on the website of our project (https://kreimanlab.com/code/testing-turing-tests/).
Data analysis	The code used for data analysis was performed in python, and is released on our project's GitHub page (https://github.com/kreimanlab/TuringTest)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All of the data used in our experiments, and collected from our experiments, is publicly available at: https://kreimanlab.com/code/testing-turing-tests/

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study has collected participants' self-report gender (male, female, or other) as a biological attribute. The proportion of participants per gender is represented in Figures S2, S4 and S7. We aggregated all participants' data in our analyses. Post-hoc gender-based analyses were not conducted, as it was out of the scope of this project.
Reporting on race, ethnicity, or other socially relevant groupings	No groupings based on race or ethnicity were used in this study. The only relevant grouping tested was based on familiarity with AI. Thus, a separate category of "AI experts" was identified composed of people who had taken at least college-level computer science courses (see manuscript)
Population characteristics	See above
Recruitment	Ads for our tasks were posted on both the platforms of MTurk and Prolific. The control groups of in-person participants were recruited by posting an ad in channels known to the authors (slack channels of the program and labs, etc)
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Human participants and AI models were tasked with attributing if solutions to visual and/or language tasks were provided by a human or a machine (Turing-like tests). The data reported are quantitative.
Research sample	The characteristics of the sample emerged from the availability of participants, both on online platforms and in person. We did not select for any relevant criteria (e.g. gender, sex, age, other). The breakdown of the obtained sample are shown in Figures S2, S4, S7.
Sampling strategy	We chose large samples of data to obtain a representative extraction of the population. We did not perform subsamples in our dataset.
Data collection	The data collection happened in different phases. 1) The six tasks of our experiments were solved by both humans and machines. For this, human participants on online platforms of Amazon MTurk and Prolific were recruited to solve language and visual tasks. The design of all of our tasks (made with Javascript and the package of Psyturk) are made available for experimental reproducibility, and made public at: https://github.com/kreimanlab/TuringTest. 2) The solutions of the tasks (both human and machine) were inserted in Turing-like tests, composed of several trials where subjects (recruited on the same online platforms) were to determine the nature (human or machine) of the agent who solved the tasks. Separately, control groups of participants were recruited from the Harvard student body and participants in the labs of the authors to test whether online results are representative of in-person, controlled settings.
Timing	The first cohort of data collection was performed in 2023-2024. Following reviewers' requests, we collected more data to carry out control experiments in 2025.
Data exclusions	No data were excluded from our analyses.

Non-participation	A few online participants did not finish the test. The platform does not request to state the reason for giving up the job.
Randomization	All experiments were double blind and randomized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*