OPEN ACCESS

Artificial Intelligence REPORT

The Innovation Informatics

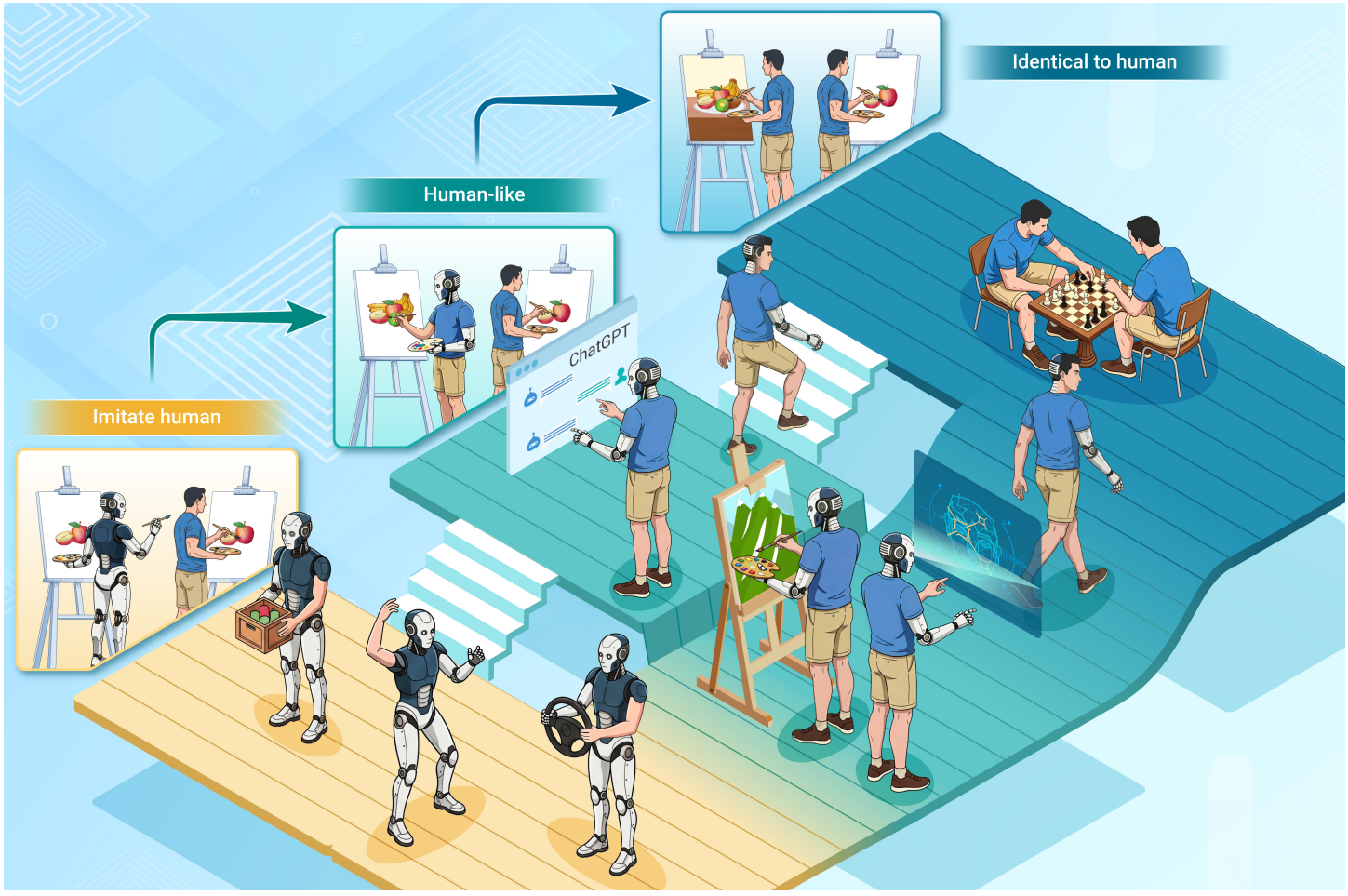
Measuring human likeness of artificial intelligence to natural intelligence

Yingji Xia,[1] Bing Zhu,[1] Maosi Geng,[1] Shengnan Zhu,[2,3,*] Sudan Sun,[4] Hui Chen,[2,*] Xiqun (Michael) Chen,[1,*] Xiaoxiang Na,[5] Jieping Ye,[6,7] and Christopher S. Tang[8]

*Correspondence: shengnanzhu1996@163.com (S.Z.); chenhui@zju.edu.cn (H.C.); chenxiqun@zju.edu.cn (X.C.)

Received: May 11, 2026; Accepted: May 11, 2026; Published Online: May 25, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100047

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Human likeness of artificial intelligence (AI) was quantified behaviourally and neurally.
- Comprehensive measurements were used to replace binary self-reports of the Turing test.
- Behavioural human likeness linearly increased when AI's behavioural pattern approached human.
- Neural human likeness did not increase until the AI performed identically to human.

REPORT Artificial Intelligence

OPEN ACCESS

www.the-innovation.org/the-innovation-informatics

Measuring human likeness of artificial intelligence to natural intelligence

Yingji Xia,[1] Bing Zhu,[1] Maosi Geng,[1] Shengnan Zhu,[2,3,*] Sudan Sun,[4] Hui Chen,[2,*] Xiqun (Michael) Chen,[1,*] Xiaoxiang Na,[5] Jieping Ye,[6,7] and Christopher S. Tang[8]

[1]Institute of Intelligent Transportation Systems, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China
[2]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou 310058, China
[3]Research Center of Brain and Cognitive Neuroscience, Liaoning Normal University, Dalian 116029, China
[4]Shulan International Medical College, Zhejiang Shuren University, Hangzhou 310015, China
[5]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK
[6]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor MI 48109, USA
[7]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor MI 48109, USA
[8]Andreson School of Management, University of California, Los Angeles CA 90095, USA
*Correspondence: shengnanzhu1996@163.com (S.Z.); chenhui@zju.edu.cn (H.C.); chenxiqun@zju.edu.cn (X.C.)

Received: May 11, 2026; Accepted: May 11, 2026; Published Online: May 25, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100047

Citation: Xia Y., Zhu B., Geng M., et al. (2026). Measuring human likeness of artificial intelligence to natural intelligence. *The Innovation Informatics* **2**:100047.

Natural intelligence (NI) is considered the biological foundation and ultimate aim of artificial intelligence (AI). Leveraging the steep increase in computational power and high-performance algorithms, many AI-empowered applications have claimed to attain human-level natural intelligence and behave like humans in certain tasks. However, these claims are primarily based on the classical Turing test and heavily rely on the binary self-reported ratings. Here, we introduced comprehensive behavioural and neural measurements to quantify the continuous distinctions between AI and NI. An adaptive-learning nonverbal Turing test was used to imitate the certainty and variability of human behaviours simultaneously, and typical LLM-based implementations were conducted as validation. Our results clarify the human behavioural and neural sensitivity of evaluating AI human likeness, extending Turing's testing criteria, and illustrate an empirical basis for pointing out directions of various human-like AI applications to approach NI.

Artificial intelligence (AI) research learns from natural intelligence (NI)[1,2] and acts as the machine version of NI.[3,4] And it facilitates the liberation of human labour in a spectrum of human-involved tasks. In industrial manufacturing, AI-driven robotic systems can learn operational patterns from human interactions, thereby enhancing adaptability, predictive maintenance, and overall performance.[5,6] Similar benefits are observed in other areas such as intelligent healthcare[7,8] and remote education.[9,10] Most of these tasks require AI systems to behave or interact at the human level. Generally, humans exhibit NI from cerebral neural networks,[11,12] which is endowed by natural biological evolution and allows us to cognize, analyse and behave intelligently. On the other hand, AI tries to learn from NI by borrowing or imitating functional structures or operational principles of human brains.[13] Following this train of thought, related AI methodologies have become even more powerful and feasible today by optimizing neural network model structures or adding millions to billions of compute nodes.[14,15] Recent advances in large language models (LLMs) in generative AI, such as DeepSeek, ChatGPT, Sora, ERNIE Bot, and DALL-E, have shown powerful text, image, and video generation abilities indistinguishable from those of human creations.[16-18] As AI systems increasingly function as interactive agents rather than passive tools, human likeness has emerged as a critical dimension of evaluation. Prior work shows that humans naturally treat machines as social actors, applying human social norms and expectations even to artificial systems.[19] Human likeness may significantly influence trust, engagement, and interaction outcomes,[20,21] while also introducing risks such as overtrust and misinterpretation of AI capabilities.[22] Therefore, understanding and evaluating the degree to which AI systems resemble human intelligence is essential, not only for improving interaction quality, but also for ensuring safe, predictable, and responsible deployment. In this context, systematic and objective approaches to assessing human-likeness are needed.

Currently, the criteria for judging AI's human likeness of an AI system are insufficient. The classical "imitation game" proposed by Alan Turing in 1950[23] has dominated today's behavioural research to assess various AI systems.[24,25] And we followed this behaviourist definition of "intelligence" in the rest of the paper. The classical Turing test can be summarized by an interrogator interacting with a human and an AI in a certain task (where each party is physically separated and invisible from the other two), and attempting to distinguish which one is artificial. If more interrogators mistakenly conclude that the AI is a human partner, then the AI is considered to have the ability to exhibit human-like natural intelligence.

Turing test results rely heavily on binary self-reports (i.e., human or AI partner) from interrogators. For instance, the Loebner Prize, which aimed to select the most human-like AI chatbot following the classical Turing test pipeline, was widely criticized as chatbots were tuned to "fool" the poorly qualified human judges in a given short timeframe. Furthermore, aggregated binary decisions can barely reflect human decision boundaries or the drifting process of implicit cognition,[26,27] raising controversy regarding its effectiveness in assessing state-of-the-art AI.[28,29] To overcome this limitation, other multidimensional criteria have been introduced to extend the Turing test. For instance, the Total Turing test added other behavioural similarities to the Turing test, such as the visual or motor ability of a robot, to better assess the human likeness of AI systems.[30] However, these additions are limited to the behavioural level of the Turing test, which tends to be subjective and biased. Therefore, better objective testing criteria and advanced unbiased methodologies are urgently needed for assessing AI systems.

Here, we extended the binary evaluation metrics and proposed comprehensive measurements to leverage AI to NI. We designed an adaptive-learning Turing test based on the modified human-machine interaction Joint Simon (JS) task to imitate the process of AI learning from NI. The proposed JS task acts as a typical independent collaboration paradigm, which facilitates inserting interactions taken over by AI during collaboration without being detected. Additionally, we employed functional near-infrared spectroscopy (fNIRS) to monitor neurological signature changes and illustrate how humans evaluate AI neurally. Validation experiments were performed with two LLMs, and the results showed promising implementation of the proposed method to evaluate various AI applications.

Human-likeness: From AI to NI

In this study, we aimed to measure the human-likeness of AIs from behavioural and neural perspectives. To this end, we referred NI in this study as human behavioural performance and neural activation in doing certain cooperative tasks. On the other hand, AI refers to the computer's or machine's performance in doing these tasks by adopting a similar approach as humans. We regard the human likeness of AI system's behavioural performance as the major criterion for their development towards NI. Generally, AI's human likeness should *continuously increase* according to its development and eventually reach 100% (from naively imitating humans to behaving identically to humans), where the AI is identical to NI. However, the process of increasing AI human likeness in terms of its development remains unknown.

We hypothesized that the increasing process of human likeness would be monotonic, which meant that AI development could positively contribute to

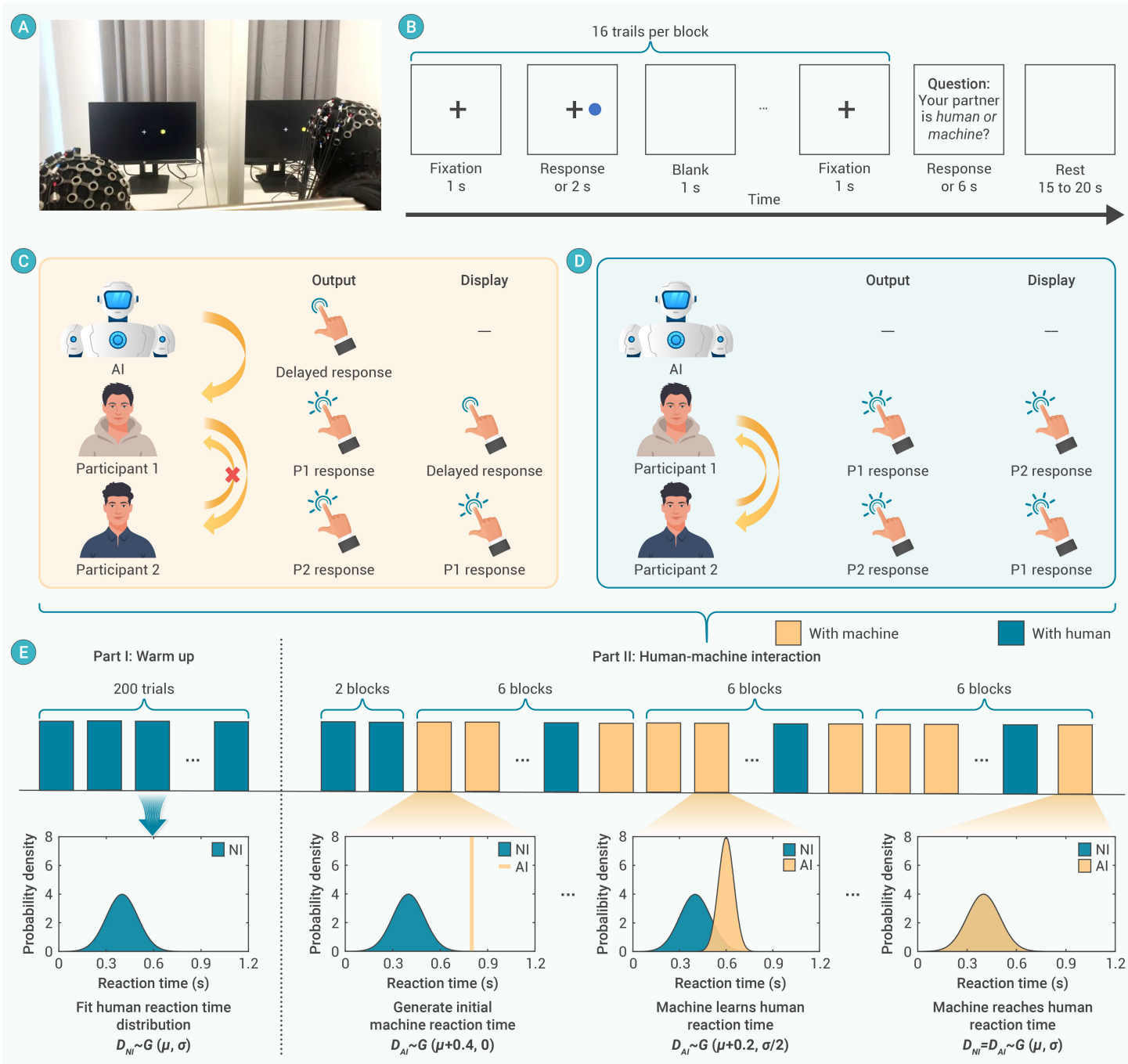


Figure 1. Experimental task and procedures (A) Experimental setup of the proposed adaptive-learning Turing test. (B) Experimental procedure of a typical block in the task. (C & D) Output and display manners of the two types of blocks, where Participant 1 interacts with either AI (pink blocks) or the other human participant (blue blocks). The dark hand indicates that the displayed reaction for Participant 1 was generated by AI for 15 of the 20 blocks. Participant 2 saw the real outputs of the two human participants for the entire human-machine interaction phase. (E) Organizational details of the adaptive-learning Turing test. The human participants were told that they were interacting with the human partner in the warm-up phase, and were asked to discriminate whether the partner was a machine or human in the human-machine interaction phase. AI was trained with input data from Participant 2 in the warm-up phase, and generated stilted intelligent responses adaptively with an equal learning rate in 15 of 20 blocks in the human-machine interaction phase.

the increase in human likeness. In this case, the corresponding gradients of AI human likeness are potentially linear, quadratic, exponential, stepwise, etc. Moreover, the thresholds of the increasing curves need to be investigated and quantified.

Adaptive-learning Turing test

We propose a simple yet quantifiable adaptive-learning Turing test to discretize AI learning processes into equal-length steps and examine changes in human likeness. The test is a generalization of the nonverbal Turing test[31,32] based on the modified human-machine interaction JS task,[33,34] where two participants interact in a simple joint key-pressing task, and one of them can be easily replaced by AI. In this task, a two-person dyad was asked to respond to either yellow or blue circles and press fixed keys corresponding to the colours. Each participant was in charge of only one colour. Both participants can see the stimuli. State-of-the-art neuroscience techniques were introduced to quantify the correlations of neural responses during interactions[35] and extend the criterion dimensions. The brain activity of each participant was monitored with an fNIRS optical topography system (Figure 1A).

The test consisted of two phases: (1) a warm-up phase (200 trials) to become familiar with each other and (2) a human-machine interaction phase (20 blocks, 16 trials each, see Figure 1B for detailed experimental procedures of a typical block). Participant 1 acted as the interrogator, and Participant 2

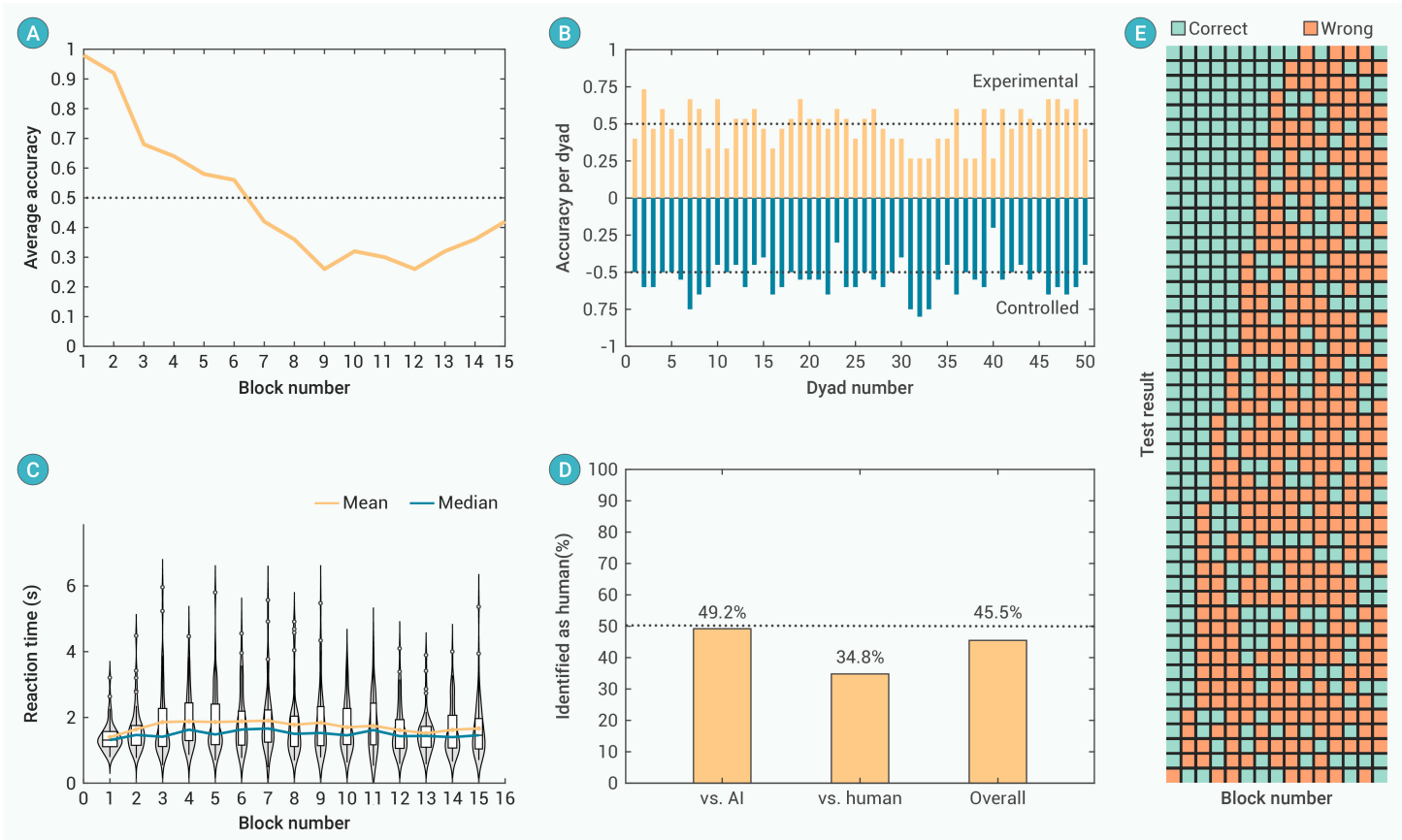


Figure 2. Behavioural performance (A) Accuracy changes in response to the questions of Participant 1 interacting with AI in 15 blocks. The yellow dotted line indicates 50% accuracy. (B) Accuracy for answering questions of each dyad, including both the experimental (Participant 1) and controlled (Participant 2) groups. (C) Violin diagram of the reaction time of Participant 1 when answering the question at the end of each human-AI interaction block. (D) Average accuracy of Participant 1 to answer questions upon interactions with AI, humans, or both. The yellow dot line indicates 50% accuracy. (E) Chess map of the sorted test results of Participant 1 for answering questions upon interacting with AI.

acted as the witness in the experiment. First, during the warm-up phase, each participant was told that they were interacting with the human partner, and was required to pay attention to the reaction patterns from their partner. In the JS task, reaction time serves as a direct observable dimension through which participants perceive and evaluate their interaction partner. Therefore, we recorded the key-pressing reaction patterns of Participant 2, namely the mean reaction time and standard deviation, in the warm-up phase online and used the Gaussian model $D_{NI} \sim G(\mu,\sigma)$ to fit the reaction time distribution (see Figure S1).

Next, during the human-machine interaction phase, each participant was requested to discriminate AI from a human partner from the key-pressing patterns at the end of each block (Figures 1C-E). Participant 1 conducted the human-machine interaction experiment with AI-generated responses for 15 of the 20 blocks and with responses from Participant 2 for the rest 5 blocks. Participant 2 conducted the human-machine interaction experiment with responses from Participant 1 for all 20 blocks. Each participant was asked whether the interacting partner was human or machine at the end of each block (see Methods section for details).

For these AI-generated responses shown to Participant 1 in the 15 human-machine interaction blocks, we tuned the machine reaction time distribution, imitating the reaction patterns of Participant 2 using the fitted Gaussian model $D_{NI} \sim G(\mu,\sigma)$ in the warm-up phase. The machine reaction time was sampled from the distribution $D_{AI} \sim G(\mu+0.4,0)$ in the first AI-generated response block (slow, no variability) and $D_{AI} \sim G(\mu,\sigma)$ in the last AI-generated response block (identical to Participant 2). The mean and standard deviation values of the Gaussian distribution for the other 13 AI-generated response blocks were assigned a fixed learning rate. That is, the mean value decreased evenly from $\mu+0.4$ to μ, where was $\mu+0.371$ in the second block, $\mu+0.343$ in the third block, etc.; and the variance increased evenly from 0 to σ, where was 0.071σ in the second block, 0.143σ in the third block, etc.; see Figure S2 and Table S4 for the model parameter details.

Using this design, the AI initially generated slow and stilted machine reactions, became intelligent by gradually obtaining human-like behavioural patterns,[24] and ultimately behaved identically to the human partner. Task design could discretize and imitate the adaptive learning manner of AI systems and facilitate quantifying corresponding human-likeness changes.

Behavioural performance

The adaptive-learning Turing test was conducted on 50 dyads. The overall behavioural performance was measured for the 15 human-machine interaction blocks (see Figure 2 & S3 for details). The average accuracy of Participant 1 for answering questions discriminating AI from human partners decreased from 98% in the first machine-active block to 26% in the 9th machine-active block and ranged between 0.26 and 0.42 for the remaining blocks. The corresponding median reaction time for answering these questions ranged between 1.31 and 1.66 s. The mean reaction time for answering the questions was greater due to outliers ranging between 1.40 s and 1.90 s.

The tendency of the accuracy curve for answering questions showed that the AI-generated machine responses became difficult to distinguish from human responses as the mean reaction time and variance approached humans. The accuracy decreased to less than 50% in the 7th block, where the reaction time was sampled from $D_{AI} \sim G(\mu+0.23, 0.43\sigma)$. The accuracy bounced back in the last 3 blocks, potentially resulting from a compensation effect of the imbalanced human-human/human-machine interaction block numbers.

Neural performance

Performing interactive tasks and dynamically predicting mutual behaviours can cause stronger interbrain neural activations among participants.[36,37] Here, we regarded the AI-generated stilted machine reaction as passive cooperation

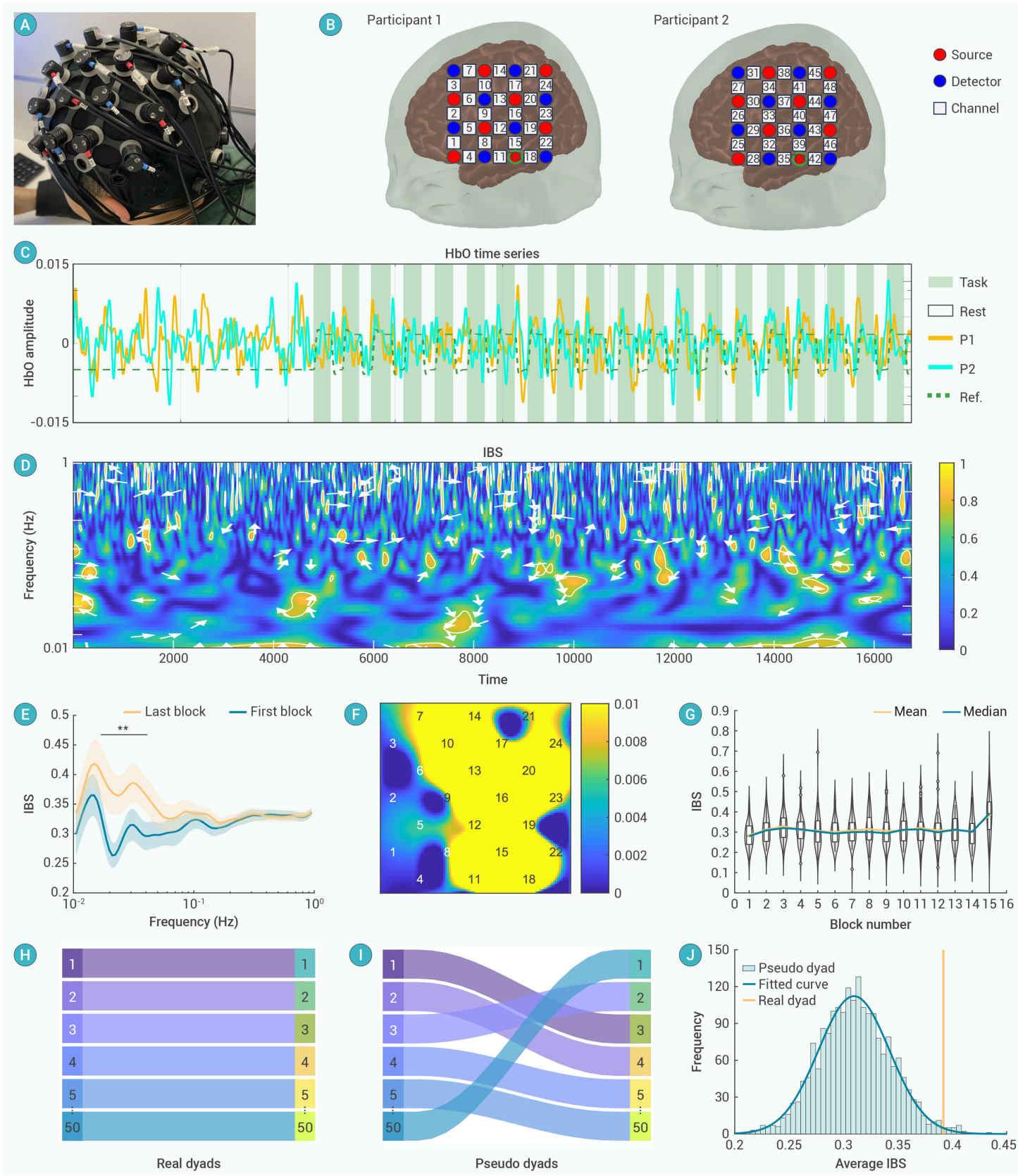


Figure 3. IBS analysis and neural performance (A) Optode emitter setup for the fNIRS hyperscanning experiment. (B) Optode probe set located on the fronto-temporo-parietal regions in the left cortex of each participant. (C) Channel-wise HbO amplitude time series extracted from both participants during the experiment. (D) Corresponding IBS time series in the 0.01 to 1 Hz frequency band estimated by the WTC algorithm. (E) IBS values associated with the first and last blocks interacting with AI across all dyads, ranging between 0.01 and 1 Hz. (FDR corrected, ** $p < 0.01$). The shade represents 2.5 standard error of the mean (s.e.m.). (F) Corresponding p value map for each channel in the FOI across all dyads (FDR corrected). (G) Violin diagram of the IBS values in each human-AI interaction block. (H & I) Schematics of the permutation analysis. (J) Comparison of IBS values of the permutation result histogram and true means of the original dyads.

and the human-like machine reaction as active cooperation. As the machine reaction gradually approached human behaviours, the participants' interbrain

Figure 4. Summary of the results of the proposed experiment (A) Tendency of the behavioural experiment results. (B) Tendency of the neural experiment results. (C) Tendency of the median reaction time when responding to the questions; all shades represented 2.5 s.e.m. (D & E) Line chart and schematic figures describing the relationship between the AI development and human likeness. Human likeness gradually increased from a behavioural perspective and suddenly increased in the end from a neural perspective.

neural activations would strengthen as the machine progressively learned.[38]

We used interbrain synchrony (IBS) as the neural signature to measure interbrain neural activations. The optodes of the fNIRS system were set on the fronto-temporo-parietal regions in the left cortex (Figures 3A & B). The oxyhaemoglobin (HbO) amplitude was extracted from homologous channels in both participants (Figure 3C), and wavelet transform coherence (WTC) was employed to estimate the IBS between these channel pairs (Figure 3D). The task-evoked IBS estimated in the first and last human-machine interaction blocks were compared to determine the frequencies of interest (FOIs) and regions of interest (ROIs). The FOI ranged from 0.017 to 0.045 Hz (Figure 3E, FDR corrected, ** $p < 0.01$). ROIs were mainly located in the left frontal region (Figure 3F, FDR corrected, ** $p < 0.01$); these were consistent with the findings of other studies on human social cognition.[39] The IBS estimated within the 15 human-machine interaction blocks is shown in Figure 3G; here, the IBS values were relatively steady at approximately 0.30 in the first 14 blocks and then increased to 0.39 in the last block.

Permutation analysis was conducted to verify the dyad-specific IBS by shuffling participant pairs and forming pseudo-dyads (Figures 3H & I). The permutation results are shown in Figures 3J, and it is confirmed that the task-evoked enhanced IBS was dyad specific.

Human-likeness scale from AI to NI

We investigated the hypothesis regarding the increasing human likeness in relation to the development of AI from both behavioural and neural perspectives. We fitted tendency curves for behavioural and neural human likeness and median reaction time after breakpoint selection (Figures 4A-C). The behavioural human likeness was quantified by the error rate (1 minus accuracy), which indicated instances where the interrogator misidentified the AI-generated machine response for those from human partners. The behavioural human likeness linearly increased from the first set and reached its maximum in the 9th block. In contrast, neural human likeness did not increase until the AI performed identically to NI. The median reaction time curve was fitted by a quadratic equation, which meant that the interrogator could easily discriminate the stilted and human-like AI at both ends but would become confused as the AI-generated behaviours approached NI.

To conclude our answer, we plotted a time chart and schematic figures to describe the human-likeness scale from AI to NI (Figures 4D & E). From the behavioural perspective, AI's human likeness linearly increased when its behaviour approached NI. Interestingly, from the neural perspective, the interrogator had the ability to subconsciously discriminate between the AI and NI until the AI performed identically to NI. This suggests that although the interrogator was implicitly aware of subtle differences between AI and NI in the brain, they still tended to report incorrect answers.

Our empirical findings support the hypothesis: The process of AI's human likeness increases linearly behaviourally and stepwise neurally. The behavioural threshold for human-like AI occurs much earlier than the neural

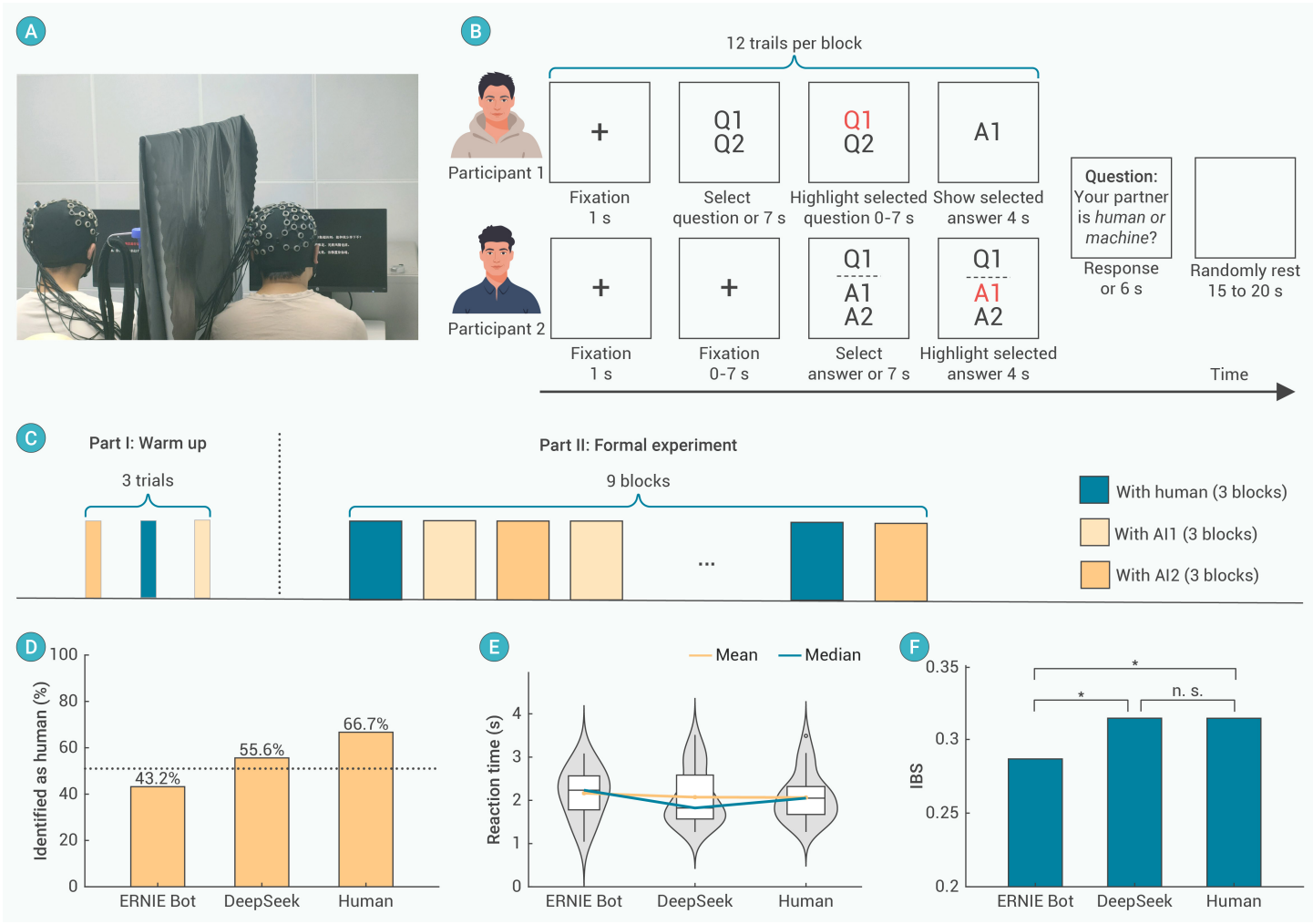


Figure 5. Experimental procedures and results of the verbal Turing test (A) Experimental setup of the test. (B) Experimental procedure of a typical block in the task. (C) Organizational details of the verbal Turing test. (D) Average percentage of Participant 1 to identify the partner as a human. The yellow dotted line indicates 50% accuracy. (E) Violin diagram of the reaction time of Participant 1. (F) Average IBS values in each human-AI interaction type. (FDR corrected, $\star$ $p < 0.05$, n.s. represents no significant difference).

threshold. Human brains can implicitly distinguish AI from NI as long as AI-generated behaviour is not identical to human-generated behaviour.

Validation on LLMs

We conducted an additional experiment utilising different LLMs to demonstrate the typical implementation of the proposed method and validate its effectiveness. The verbal Turing test with human- or LLM-generated conversations were implemented as the interaction task, and both behavioural and neural performance were assessed. Identical FOI and ROI were used to estimate the IBS as that of the main experiment (see Methods section for details).

Specifically, ERNIE Bot and DeepSeek were tested to generate questions and answers separately. The experimental procedures and results are shown in Figure 5. Similar to the main experiment, we focused on analysing the performance of Participant 1. From the behavioural performance perspective, ERNIE Bot was identified as the human partner in 43.2% blocks, while DeepSeek surpassed 50% and was comparable to human partners. From the neural performance perspective, the average IBS value of ERNIE Bot has a significant difference from that of DeepSeek (p = 0.03, FDR corrected) and human partner (p = 0.03, FDR corrected), and the IBS values of DeepkSeek and human partner showed no significant difference (p = 0.99, FDR corrected). From the proposed criteria, we can conclude that DeepSeek can pass the Turing test both behaviourally and neurally.

DISCUSSION

We analysed the tendency of AI human likeness with respect to its development from both behavioural and neural perspectives. Also, we proposed comprehensive measurements to leverage AI and NI. Our method addresses the limitations of the binary self-reporting approach used in the classical Turing test by introducing IBS as a neural criterion. Our results indicated that humans had different thresholds for judging AI human likeness at the behavioural and neural levels. Furthermore, using neurological hints, humans could discern subtle behavioural differences during human-machine interactions. Our empirical finding allows for a more informed assessment and alignment of various human-like AI.

Our study provided the following key merits and implications. First, the original Turing test aimed to help humans assess AI-related human likeness without restrictions. However, many AI applications were assessed within constrained task-solution spaces for human-like claims, which oversimplified the Turing test's original intent. In contrast, our proposed method integrates explicit self-reports and implicit neural indicators, capturing the essence of the Turing test while relaxing the constraints. Second, traditional testing criteria often rely on explicit self-reporting tasks (e.g., yes/no responses or confidence ratings). However, interrogators' subjective consciousness can significantly bias the results. Advantageously, the neural synchrony metrics of our proposed method implicitly evaluated attitudes towards AI's human likeness, mitigating self-reporting bias and complementing the Turing test criteria. Third, although the Turing test has been widely used to assess AI in human-machine interaction tasks, the focus has been on whether the AI could finish the task properly. Our study takes an additional step by treating AI as an independent "social being" and assessing its human likeness purely based on social interactions. This perspective extends the basics of the Turing test. Fourth, the observed surge in neural synchrony towards the last block is a particularly intriguing phenomenon, displaying a

distinct pattern compared to the behavioural results. We suggest this may reflect increased sensitivity at the neural level,[40,41] with the brain's dichotomous response. Specifically, the brain seems to categorize the interactive entity as either human or AI, with no middle ground. Confusion arises only when human and AI behaviours become highly similar. The last block of the experiment may have reached this threshold of similarity, triggering the sudden increase in synchrony as the brain mistakenly identifies AI as human. Last, a "compensation effect" appeared in the last three blocks of the behavioural outcome, likely due to biases in decision-making. After Block 9, the behavioural performance plateaued, suggesting participants struggled to differentiate between humans and AI. Based on the low accuracy in Blocks 9-12, participants exhibited an increased tendency to choose 'human'. As the task became harder, they likely adopted a strategy of balancing responses, selecting 'AI' more often to even out the distribution, which may have boosted overall accuracy.

As the validation test shows, our proposed method could be generalized and implemented beyond RT to richer behavioural spaces, to assess the human likeness of various AI applications. We acknowledge that motor and verbal coordination involve distinct neural processes. However, our method is not tied to any specific modality, but rather to the perceived coordination and synchrony between agents. AI applications should first undergo the Turing test to determine whether they can pass it behaviourally. Their human likeness can be relatively quantified based on the behavioural rating, which tends to be linear. Moreover, neurological IBS can be simultaneously estimated to determine whether AI can stimulate neural synchrony similar to that observed in humans. If the AI can pass both the behavioural and neural test, it can be considered identical to humans.

Our analysis has several limitations that point to important avenues for future research. One important aspect is that the simple key-pressing experiment test lacks rich social cues in naturalistic interactions. We chose this experiment setup as it had excellent quantitative metrics with few interference factors and was highly controllable. Importantly, intelligence and consciousness are multifaceted constructs, but the proposed research only investigated the pathway through AI to NI from the behaviour and cognition similarity perspective. We acknowledge that verification of the generalization ability of the proposed research in more complex cognitive and emotional domains, such as various social collaborative environments or multimodal settings, should be further tested elaborately. Other AI evaluation pipelines beyond behavioural tests, such as pattern-recognition intelligence tests,[42] can be fruitful to complement the proposed work by adding more dimensions to the evaluation criteria. Though our experiment extends the classical Turing test to the neurological level, the test-setting framework may not be applicable or generalizable to all AI application scenarios. In some face-to-face human-machine interaction tasks, the neural synchrony of the interrogator may be weaker when facing the machine partner.[43] Importantly, research in social neuroscience has shown that interactive contexts not singly captured by IBS, but are associated with coordinated changes across multiple neural levels, including inter-brain synchrony, intra-brain activation patterns, and functional connectivity within social-cognitive networks.[44] Therefore, future efforts on exploring alternative neurological metrics beyond IBS could compensate for the deficiencies of our proposed testing method. Alternative neural or behavioural metrics such as EEG, fMRI and pupillometry can serve as complementary measures when IBS is unavailable due to specific task settings or equipment limitations. Such studies could strengthen our theoretical findings and improve the explainability and robustness of the proposed method.

There is a surging demand for modern human-like AI systems and applications to collaborate closely and intelligently with humans. Specifically, robots with more human-like attributes are more likely to be perceived by humans as potential collaborators,[45] which may facilitate human-AI cooperations. Meanwhile, the misuse of LLM-based human-like content generation tools must be correctly identified for better regulation purposes.[46] All these needs can be fulfilled by properly measuring the human-likeness of AI to NI. Our experimental results indicate that future AI should not only mimic human behaviour, but also synchronize seamlessly with humans during interactions to become an ideal fit for various human-machine cooperation tasks. We believe the solution to improve AI's human likeness would not be simply providing exponentially more data to deeper neural networks. We underline the importance of developing human-like AI by imitating specific cerebral functions or neural information processing pipelines.[47,48] For instance, in the case of ChatGPT and other LLMs, although its plausibility has increased by feeding numerous labelled data, its reliability has decreased notably and cannot be ensured in workflows.[49] Brain-inspired algorithms or neuromorphic computing techniques need to be developed.[50,51] These approaches hold promise for enabling various AI systems to achieve a level of sophistication comparable to NI in the future.

EXPERIMENTAL PROCEDURES

Resource availability

Materials availability. This study did not generate new unique materials.

Data and code availability. The implemented adaptive-learning Turing test source code and the processed data used in the main experiments to support the findings are publicly available in the following Zenodo repository: https://doi.org/10.5281/zenodo.11381796.[52] The data and code associated with the validation experiment are available in the following Zenodo repository:https://doi.org/10.5281/zenodo.15546296.[53]

Ethical approval. The experimental procedures were approved by the Research Ethics Committee of the Department of Psychology and Behavioural Sciences, Zhejiang University ([2021]024). The study was conducted in accordance with the ethical standards of the 1964 Declaration of Helsinki. The inclusion criteria can meet the Global Code of Conduct. All participants provided written informed consent before the experiment, and the possible consequences of the study were explained. Specifically, before the experiment, participants were informed that they would unknowingly interact with either humans or AI and make judgments. Afterward, we provided a full and honest explanation of the experimental design. Additionally, all data is securely stored and kept confidential in accordance with privacy and ethical guidelines.

Participants

For the main experiment, one hundred right-handed undergraduate/graduate students (18-33 years old, mean age=21.84±2.88, 52 female) from Zhejiang University, forming 50 dyads, were recruited. Previous studies[54,55] suggested this sample size is sufficient to detect meaningful effects and ensure the robustness of the results. Moreover, we performed a priori power analysis using G*Power 3.1[56] to estimate the appropriate number of participants. The effect size (dz) was estimated to be 0.6 based on previous studies that used the JS task in fNIRS research.[54,55] The power calculation yielded an estimated minimum of 45 pairs of participants to detect the IBS with 90% power (with α set to 0.01). For the additional validation experiment, we recruited other 54 participants, forming 27 dyads. The participants are right-handed undergraduate/graduate students (18-33 years old, mean age=21.94 ± 3.01). The sample size was determined using a medium effect size (Cohen's d = 0.5), which was based on the study on LLMs.[34] The power calculation yielded an estimated minimum of 27 pairs of participants to detect the IBS with 80% power (with α set to 0.05). All participants were prescreened to ensure that none came from disciplines directly related to artificial intelligence or cognitive science. Each dyad was formed with two participants of the same sex following previous evidence and recommendations to mitigate the confounding effects elicited by sex compositions.[57]

Experimental setup and stimuli

The experiment was conducted in a noiseless room. The two participants sat side by side, were separated by a partition, and faced the display screens and keyboards, as shown in Figure 1A. The experimental program for the adaptive-learning Turing test was developed using the PsychoPy library under the Windows system, and all screens and keyboards were connected to one host. The experimental stimuli were designed using the interactive button-pressing task. During the experimental trials, yellow and blue circles were randomly presented on the left or right on the black background of the display screen, with a white fixation cross in the middle for position instruction purposes. Each participant was in charge of a fixed colour stimulus, regardless of the position.

Turing test experiment

The adaptive-learning Turing test experiment was divided into the warm-up phase and the human-machine interaction phase (Figure 1E).

The warm-up phase consisted of two cooperation blocks, and each block contained 100 trials. Both the colours and positions of the stimuli were generated in a random order, and the total number of each stimulus was the same (25 trials per type in each block). The participants were informed that they were interacting with each other and should be familiar with the other's reaction patterns. During the warm-up phase, participants were not aware of any potential AI involvement. This design ensured that participants formed an initial representation of a human interaction partner without bias. Note that we recorded the response times of Participant 2 under the premise of pressing the correct button to calculate the machine's sampling response time (i.e., fitting the Gaussian model) required in the second phase of the experiment.

The human-machine interaction phase consisted of 20 blocks, including 5 human-human interaction blocks and 15 human-machine interaction blocks. Each block contained 16 trials (4 trials per type, randomly ordered). The first 2 blocks were human-human interaction blocks, and the following 18 blocks were equally divided into 3 parts, each containing 6 consecutive blocks. Each 6-block part included 1 human-human interaction block and 5 human-machine interaction blocks in random order. The participants interacted with each other in the human–human interaction blocks. In the human-machine interaction blocks, Participant 1 was interacting with the machine, whose reaction time was generated by an adaptive-learning AI. At the end of each block, participants were asked to answer the question "Do you think the human partner did the task, or the machine did the task?" The maximum time allowed for answering the question was set to 6.0 s. The details can be found in Supplementary Note 1.

Machine reaction time calculation. We first recorded the reaction time of Participant 2 in the warm-up phase (100 trials) and employed a Gaussian model to fit the reaction time distribution online before entering the second phase; this can be denoted as $G_{NI}(\mu,\sigma)$, where μ is the mean reaction time and σ represents the standard deviation. In the human-machine interaction phase, there were 15 blocks for human-machine interaction. The machine reaction time was generated by an adaptive-learning AI with a fixed learning rate for these human-machine interaction blocks. Specifically, the reaction time was generated by sampling from $G_{AI}(\mu+x,0)$ in the first human-machine interaction block (acting as primitive and stilted AI) and from $G_{AI}(\mu,\sigma)$ in the last human-machine interaction block (behaving identically to Participant 2). The time delay parameter x was set to 0.4 s, which could be easily distinguished by most participants in the test. The AI-generated reaction time of the k-th human-machine interaction block $G_{AI}(\mu_k,\sigma_k)$ could be calculated as follows:

$$\mu_k = \mu + \frac{K-k}{K-1}x,$$

$$\sigma_k = \frac{k-1}{K-1}\times\sigma, \tag{1}$$

where $K=15$ and is the total number of human-machine interaction blocks and $k=1,2,3,\ldots,15$. The equations indicate that AI human likeness increases over time during human-machine interaction blocks (Supplementary Note 2). The pattern for AI-generated reaction time is more similar to that of the human participant than to that of the previous human-machine interaction block, with less behavioural difference and more variability.

Behavioural data analysis

The binary answer of Participant 1 to the question (Do you think the human partner did the task, or the machine did the task?) at the end of each block was recorded and used to assess the behavioural performance of the AI. The reaction time utilized by Participant 1 in response to the question at the end of each block was recorded to perform the behavioural analysis.

fNIRS data acquisition

We used a LightNIRS optical topography system (Shimadzu, Japan) to simultaneously monitor and record brain activity in the dyads (i.e., hyperscanning[58]). The configuration of the optodes featured 16 light emitters and 16 detectors clustered over the fronto-temporo-parietal regions in the left cortex; this configuration was based on previous studies showing that these regions were strongly associated with social interaction and learning tasks.[59-61] Each emitter-detector pair formed one channel, resulting in 24 channels per participant, as shown in Figure 3B.

The modified Beer-Lambert law[62] was then used to estimate oxyhaemoglobin (HbO) and deoxyhaemoglobin (HbR) concentrations. We focused on the HbO concentrations in the experiment, which showed better sensitivity to cerebral blood flow[63] and a higher signal-to-noise ratio than HbR.[64,65] The controlled analysis of HbR can be found in Supplementary Note 3, where the parallel analyses returned analogous results.

fNIRS data analysis

Preprocessing. The collected fNIRS data were preprocessed and visualized using the NIRS-KIT toolbox.[66] The correlation-based signal improvement (CBSI) method was employed to reduce motion-induced artefacts.[67] We did not use any filtering or detrending techniques for HbO signals.

Interbrain synchrony (IBS). Given a pair of signals collected from homologous regions (Figure 3C), we employed the wavelet transform coherence (WTC) algorithm[68,69] to estimate the IBS within the dyads (Figure 3D). The WTC can be defined as follows:

$$WTC(t,s) = \frac{\left|s^{-1}\boldsymbol{W}^{ij}(t,s)\right|^2}{\left|s^{-1}\boldsymbol{W}^{i}(t,s)\right|^2\left|s^{-1}\boldsymbol{W}^{j}(t,s)\right|^2}, \tag{2}$$

where t denotes the time, i and j are two signals, $\boldsymbol{W}$ is the coefficient matrix calculated by the continuous wavelet transform, and s is the wavelet scale parameter. We defined the average WTC value in a specific frequency bin over homogeneous channels from a dyad as the corresponding IBS, which is consistent with previous studies.

Task-evoked IBS. Task-evoked IBS can be analysed using the following steps. First, we estimated the IBS within each block of the task and defined the target and baseline conditions. Previous studies have shown that an active task condition could be a better baseline than the rest conditions.[57,70] Hence, we selected the first human-machine interaction block as the baseline and the last human-machine interaction block as the target to detect enhanced IBS.

Second, we averaged the IBS across all channels in each dyad and selected frequencies of interest (FOIs). We compared the average IBS between the target and baseline conditions for each frequency bin within the entire frequency range (0.01-1 Hz; see Table S1)[71,72] using a series of paired one-sided t tests (see Figures S3 & 4). This frequency range covered all frequencies investigated in previous fNIRS hyperscanning studies. Given the large number of statistical tests performed across multiple frequency bins, we applied the false discovery rate (FDR) correction[73] to control multiple comparisons, setting the significance threshold at 0.01. The FOIs were chosen between 0.017 and 0.045 Hz (Table S2); this range was consistent with the claim that low-frequency oscillations (0.01-0.1 Hz) were more reliable signatures of neural synchronization.[72] FOIs excluded physiological noise such as Mayer waves (0.1 Hz), respiration (0.15 to 0.3 Hz), and cardiac pulsation (> 0.7 Hz).

Third, the regions of interest (ROIs) were identified following the same procedure. We compared the average IBS between the target and baseline conditions for each homogenous channel combination within the FOIs using a series of paired one-sided t tests. The FDR-corrected p values of the ROI tests are illustrated in Figure 3F. Since the activated brain area was composed of adjacent channels rather than sparse channels, we selected CHs 1-5 (CHs 25-29 in Participant 2; Table S3) as the ROIs for spatial continuity purposes. The IBS values used in this experiment to generate results were averaged across all FOIs and ROIs.

Permutation test. We performed a permutation test to validate whether the detected IBS was specific to real dyads. We randomly repaired signals from all participants into newly shuffled dyads (Figures 3H & I) and reconducted the IBS analysis. The permutation test was conducted 2,000 times. The HbO time sequences extracted from each pseudo-dyad were trimmed to an equal length. The average IBS values from the pseudo-dyads were esti-

mated following the same procedures and compared with the real values to verify that the task-evoked enhanced IBS was dyad specific.

Tendency curve estimation

We used the curve fitting toolbox of MATLAB to fit the tendencies of the behavioural/neural experimental results and reaction time. Specifically, we employed basic functions (e.g., constant, linear, and quadratic functions) to describe the tendencies. We used linear equations (i.e., $y = kx + c$) to fit the two segments of the behavioural results and the second segment of the neural result and used a constant value ($y = c$) to fit the first segment of the neural result. The breakpoints were selected by minimizing the root mean square errors (RMSE). The tendency of the median reaction time was fitted by a quadratic equation (i.e., $y = ax^2 + bx + c$). See Supplementary Note 4 for details.

Validation experiment

The validation experiment was conducted following protocols similar to those used in the main experiment, and the fNIRS data were recorded simultaneously. The colour stimuli and key-pressing reactions were replaced by interactive conversations in a typical verbal Turing test, where Participant 1 asked a question, and Participant 2 answered the question. Both the questions and answers contained three types of responses, separately generated from DeepSeek and ERNIE Bot, as well as human volunteers. To standardise the experimental procedure, all verbal materials were collected or generated before the experiment, and trimmed to a similar length (in Chinese, 20 characters).

The experiment contained 9 blocks, including 3 human-human interaction blocks, 3 human-DeepSeek interaction blocks, and 3 human-ERNIE Bot interaction blocks. Each block contained 12 trials with the same interaction type. In each trial, Participant 1 chose a question from two options, and Participant 2 chose a preferred answer to the question from two options. Both participants can see the question or answer selected by the partner. The order of displayed questions was randomised before each experiment. After each block, participants were asked to answer the question "Which do you think executed the task: the human partner or AI?" More details can be found in Supplementary Note 5.

REFERENCES

1. Cottam R. and Vounckx R. (2024). Intelligence: Natural, artificial, or what. *Biosystems* **246**:105343. DOI:10.1016/j.biosystems.2024.105343
2. Forli A. and Yartsev M.M. (2024). Understanding the neural basis of natural intelligence. *Cell* **187**:5833–5837. DOI:10.1016/j.cell.2024.07.049
3. Xu N.-l. (2024). How the brain's primary processing units compute to give rise to intelligence. *Nat. Rev. Neurosci.* DOI:10.1038/s41583-024-00810-4
4. Cai H., Ao Z., Tian C., et al. (2023). Brain organoid reservoir computing for artificial intelligence. *Nat. Electron.* **6**:1032–1039. DOI:10.1038/s41928-023-01069-w
5. Noy S. and Zhang W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**:187–192. DOI:10.1126/science.adh2586
6. Kumar D., Addu S.R., Lind M., et al. (2026). AI-driven hybrid deep learning and swarm intelligence for predictive maintenance of smart manufacturing robots in Industry 4.0. *Electronics* **15**:715. DOI:10.3390/electronics15030715
7. Kiyasseh D., Cohen A., Jiang C., et al. (2024). A framework for evaluating clinical artificial intelligence systems without ground-truth annotations. *Nat. Commun.* **15**:1808. DOI:10.1038/s41467-024-46000-9
8. Kurvers R.H.J.M., Nuzzolese A.G., Russo A., et al. (2023). Automating hybrid collective intelligence in open-ended medical diagnostics. *Proc. Natl. Acad. Sci. USA* **120**:e2221473120. DOI:10.1073/pnas.2221473120
9. Liu L.T., Wang S., Britton T., et al. (2023). Reimagining the machine learning life cycle to improve educational outcomes of students. *Proc. Natl. Acad. Sci. USA* **120**:e2204781120. DOI:10.1073/pnas.2204781120
10. Belpaeme T., Kennedy J., Ramachandran A., et al. (2018). Social robots for education: A review. *Sci. Robot.* **3**:eaat5954. DOI:10.1126/scirobotics.aat5954
11. Grossberg S. (1988). *Neural networks and natural intelligence*. MIT Press. DOI:10.7551/mitpress/4934.001.0001
12. Gielen S. (2007). Natural intelligence and artificial intelligence: Bridging the gap between neurons and neuro-imaging to understand intelligent behaviour. In Duch W. and Mańdziuk J. (eds). *Challenges for computational intelligence*. Springer, pp:145-161. DOI:10.1007/978-3-540-71984-7_7
13. Qiu J. (2016). Research and development of artificial intelligence in China. *Natl. Sci. Rev.* **3**:538–541. DOI:10.1093/nsr/nww076
14. Han X., Zhang Z., Ding N., et al. (2021). Pre-trained models: Past, present and future. *AI Open* **2**:225–250. DOI:10.1016/j.aiopen.2021.08.002
15. LeCun Y., Bengio Y. and Hinton G. (2015). Deep learning. *Nature* **521**:436–444. DOI:10.1038/nature14539
16. Acerbi A. and Stubbersfield J.M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proc. Natl. Acad. Sci. USA* **120**:e2313790120. DOI:10.1073/pnas.2313790120
17. Prillaman M. (2024). Is ChatGPT making scientists hyper-productive. The highs and lows of using AI. *Nature* **627**:16–17. DOI:10.1038/d41586-024-00592-w
18. Epstein Z. and Hertzmann A. (2023). Art and the science of generative AI. *Science* **380**:1110–1111. DOI:10.1126/science.adh4451
19. Nass C., Steuer J. and Tauber E.R. (1994). Computers are social actors. *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.* **1994**:72-78. DOI:10.1145/191666.191703
20. Waytz A., Heafner J. and Epley N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* **52**:113–117. DOI:10.1016/j.jesp.2014.01.005
21. de Visser E.J., Pak R. and Shaw T.H. (2018). From "automation" to "autonomy": The importance of trust repair in human–machine interaction. *Ergonomics* **61**:1409–1427. DOI:10.1080/00140139.2018.1457725
22. Robinette P., Li W., Allen R., et al. (2016). Overtrust of robots in emergency evacuation scenarios. *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact.* **2016**:101-108. DOI:10.1109/HRI.2016.7451740
23. Turing A.M. (1950). Computing machinery and intelligence. *Mind* **59**:433–460. DOI:10.1093/mind/LIX.236.433
24. Ciardo F., De Tommaso D. and Wykowska A. (2022). Human-like behavioral variability blurs the distinction between a human and a machine in a nonverbal Turing test. *Sci. Robot.* **7**:eabo1241. DOI:10.1126/scirobotics.abo1241
25. Mei Q., Xie Y., Yuan W., et al. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci. USA* **121**:e2313925121. DOI:10.1073/pnas.231392512
26. Johnson D.J., Hopwood C.J., Cesario J., et al. (2017). Advancing research on cognitive processes in social and personality psychology: A hierarchical drift diffusion model primer. *Soc. Psychol. Personal. Sci.* **8**:413–423. DOI:10.1177/1948550617703174
27. Fudenberg D., Newey W., Strack P., et al. (2020). Testing the drift-diffusion model. *Proc. Natl. Acad. Sci. USA* **117**:33141–33148. DOI:10.1073/pnas.2011446117
28. Bayne T. and Williams I. (2023). The Turing test is not a good benchmark for thought in LLMs. *Nat. Hum. Behav.* **7**:1806–1807. DOI:10.1038/s41562-023-01710-w
29. Biever C. (2023). ChatGPT broke the Turing test—The race is on for new ways to assess AI. *Nature* **619**:686–689. DOI:10.1038/d41586-023-02361-7
30. Harnad S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds Mach.* **1**:43–54. DOI:10.1007/BF00360578
31. Rebol M., Gütl C. and Pietroszek K. (2021). Passing a non-verbal Turing test: Evaluating gesture animations generated from speech. *Proc. IEEE Virtual Real. 3D User Interfaces*, 573-581. DOI:10.1109/VR50410.2021.00082
32. Pfeiffer U.J., Timmermans B., Bente G., et al. (2011). A non-verbal Turing test: Differentiating mind from machine in gaze-based social interaction. *PLOS ONE* **6**:e27591. DOI:10.1371/journal.pone.0027591
33. Sebanz N., Knoblich G. and Prinz W. (2003). Representing others' actions: Just like one's own. *Cognition* **88**:B11–B21. DOI:10.1016/S0010-0277(03)00043-X
34. Simon J.R. and Craft J.L. (1970). Effects of an irrelevant auditory stimulus on visual choice reaction time. *J. Exp. Psychol.* **86**:272–274. DOI:10.1037/h0029961
35. Wei Z., Chen Y., Zhao Q., et al. (2023). Implicit perception of differences between NLP-produced and human-produced language in the mentalizing network. *Adv. Sci.* **10**:e2203990. DOI:10.1002/advs.202203990
36. Mayo O. and Shamay-Tsoory S. (2024). Dynamic mutual predictions during social learning: A computational and interbrain model. *Neurosci. Biobehav. Rev.* **157**:105513. DOI:10.1016/j.neubiorev.2023.105513
37. Réveillé C., Vergotte G., Perrey S., et al. (2024). Using interbrain synchrony to study teamwork: A systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **159**:105593. DOI:10.1016/j.neubiorev.2024.105593
38. Henschel A., Hortensius R. and Cross E.S. (2020). Social cognition in the age of human–robot interaction. *Trends Neurosci.* **43**:373–384. DOI:10.1016/j.tins.2020.03.013
39. Frith C.D. and Frith U. (2007). Social cognition in humans. *Curr. Biol.* **17**:R724–R732. DOI:10.1016/j.cub.2007.05.068
40. Dehaene S., Naccache L., Le Clec'H G., et al. (1998). Imaging unconscious semantic priming. *Nature* **395**:597–600. DOI:10.1038/26967
41. Hsu M., Bhatt M., Adolphs R., et al. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310**:1680–1683. DOI:10.1126/science.1115327
42. Holzinger A., Kieseberg P., Weippl E., et al. (2018). Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI. In Holzinger, A., Kieseberg, P., Tjoa, A.M., and Weippl, E. (eds.), Machine learning and knowledge extraction: CD-MAKE 2018, pp. 1-8. Cham: Springer. https://doi.org/10.1007/978-3-319-99740-7_1
43. Tsai C.-C., Kuo W.-J., Hung D.L., et al. (2008). Action co-representation is tuned to other humans. *J. Cogn. Neurosci.* **20**:2015–2024. DOI:10.1162/jocn.2008.20144
44. Montague P.R., Berns G.S., Cohen J.D., et al. (2002). Hyperscanning: Simultaneous

fMRI during linked social interactions. *NeuroImage* **16**:1159–1164. DOI:10.1006/nimg.2002.1150
45. Trafton J.G., McCurry J.M., Zish K., et al. (2024). The perception of agency. *ACM Trans. Hum.-Robot Interact.* **13**:14. DOI:10.1145/3640011
46. Kendall G. and Teixeira da Silva J.A. (2024). Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learn. Publ.* **37**:55–62. DOI:10.1002/leap.1578
47. Xia Y., Chen H. and Chen X. (2023). Integrating social neuroscience into human-machine mutual behavioral understanding for autonomous driving. *Innovation* **4**:100455. DOI:10.1016/j.xinn.2023.100455
48. Xia Y., Geng M., Chen Y., et al. (2023). Understanding common human driving semantics for autonomous vehicles. *Patterns* **4**:100730. DOI:10.1016/j.patter.2023.100730
49. Stilgoe J. (2023). We need a Weizenbaum test for AI. *Science* **381**:eadk0176. DOI:10.1126/science.adk0176
50. Schoepe T., Janotte E., Milde M.B. et al. (2024). Finding the gap: Neuromorphic motion-vision in dense environments. *Nat. Commun.* **15**:817. DOI:10.1038/s41467-024-45063-y
51. Milisav F. and Misic B. (2023). Spatially embedded neuromorphic networks. *Nat. Mach. Intell.* **5**:1342–1343. DOI:10.1038/s42256-023-00771-w
52. Geng M. (2024). Leveraging artificial intelligence to natural intelligence. *Zenodo*. DOI:10.5281/zenodo.11381796
53. Zhu B. (2025). Leveraging artificial intelligence to natural intelligence. *Zenodo*. DOI:10.5281/zenodo.15546296
54. Lin S., Zhao H. and Duan H. (2023). Brain-to-brain synchrony during dyadic action co-representation under acute stress: Evidence from fNIRS-based hyperscanning. *Front. Psychol.* **14**:1251533. DOI:10.3389/fpsyg.2023.1251533
55. Song X., Dong M., Feng K., et al. (2024). Influence of interpersonal distance on collaborative performance in the joint Simon task—An fNIRS-based hyperscanning study. *NeuroImage* **285**:120473. DOI:10.1016/j.neuroimage.2023.120473
56. Faul F., Erdfelder E., Lang A.-G., et al. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**:175–191. DOI:10.3758/BF03193146
57. Pan Y., Cheng X. and Hu Y. (2023). Three heads are better than one: Cooperative learning brains wire together when a consensus is reached. *Cereb. Cortex* **33**:1155–1169. DOI:10.1093/cercor/bhac127
58. Czeszumski A., Eustergerling S., Lang A., et al. (2020). Hyperscanning: A valid method to study neural inter-brain underpinnings of social interaction. *Front. Hum. Neurosci.* **14**:39. DOI:10.3389/fnhum.2020.00039
59. Dommer L., Jäger N., Scholkmann F., et al. (2012). Between-brain coherence during joint n-back task performance: A two-person functional near-infrared spectroscopy study. *Behav. Brain Res.* **234**:212–222. DOI:10.1016/j.bbr.2012.06.024
60. Holper L., Scholkmann F. and Wolf M. (2012). Between-brain connectivity during imitation measured by fNIRS. *NeuroImage* **63**:212–222. DOI:10.1016/j.neuroimage.2012.06.028
61. Takeuchi N., Mori T., Suzukamo Y., et al. (2017). Integration of teaching processes and learning assessment in the prefrontal cortex during a video game teaching–learning task. *Front. Psychol.* **7**:2052. DOI:10.3389/fpsyg.2016.02052
62. Cope M. and Delpy D.T. (1988). System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. *Med. Biol. Eng. Comput.* **26**:289–294. DOI:10.1007/BF02447083
63. Hoshi Y. (2007). Functional near-infrared spectroscopy: Current status and future prospects. *J. Biomed. Opt.* **12**:062106. DOI:10.1117/1.2804911
64. Ding X.P., Fu G. and Lee K. (2014). Neural correlates of own- and other-race face recognition in children: A functional near-infrared spectroscopy study. *NeuroImage* **85**:335–344. DOI:10.1016/j.neuroimage.2013.07.051
65. Homae F., Watanabe H., Nakano T., et al. (2007). Prosodic processing in the developing brain. *Neurosci. Res.* **59**:29–39. DOI:10.1016/j.neures.2007.05.005
66. Hou X., Zhang Z., Zhao C., et al. (2021). NIRS-KIT: A MATLAB toolbox for both resting-state and task fNIRS data analysis. *Neurophotonics* **8**:010802. DOI:10.1117/1.NPh.8.1.010802
67. Cui X., Bray S. and Reiss A.L. (2010). Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *NeuroImage* **49**:3039–3046. DOI:10.1016/j.neuroimage.2009.11.050
68. Grinsted A., Moore J.C. and Jevrejeva S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Process. Geophys.* **11**:561–566. DOI:10.5194/npg-11-561-2004
69. Torrence C. and Compo G.P. (1998). A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79**:61–78. DOI:10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2
70. Reindl V., Gerloff C., Scharke W., et al. (2018). Brain-to-brain synchrony in parent-child dyads and the relationship with emotion regulation revealed by fNIRS-based hyperscanning. *NeuroImage* **178**:493–502. DOI:10.1016/j.neuroimage.2018.05.060
71. Zhao H., Zhang C., Tao R., et al. (2023). Distinct inter-brain synchronization patterns underlying group decision-making under uncertainty with partners in different interpersonal relationships. *NeuroImage* **272**:120043. DOI:10.1016/j.neuroimage.2023.120043
72. Li L., Wang H., Lin Y., et al. (2023). One single-person bicycling enhances interpersonal cooperation via increasing interpersonal neural synchronization in left frontal cortex. *Hum. Brain Mapp.* **44**:4535–4544. DOI:10.1002/hbm.26397
73. Benjamini Y. and Hochberg Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**:289–300. DOI:10.1111/j.2517-6161.1995.tb02031.x

FUNDING AND ACKNOWLEDGMENTS

This research was funded by the Fundamental Research Funds for the Central Universities (226-2025-00170), National Natural Science Foundation of China (72401256, 72525009, 72431009, 72171210, 72350710798), and Zhejiang Provincial Natural Science Foundation of China (LZ26E080002). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Y.X., H.C. and X.C. proposed the question. B.Z., M.G. and S.Z. conducted the experiment. Y.X., S.S. and X.C. analysed data and wrote the first draft of the paper. H.C. X.N, J.Y. and C.T. edited the paper. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

Xiqun (Michael) Chen is an Editorial Board member of The Innovation journals and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.