Engineering 34 (2024) 12–22

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Perspective

The Tong Test: Evaluating Artificial General Intelligence Through Dynamic Embodied Physical and Social Interactions

Yujia Peng [a,b,c], Jiaheng Han [a,d], Zhenliang Zhang [a], Lifeng Fan [a], Tengyu Liu [a], Siyuan Qi [a], Xue Feng [a], Yuxi Ma [a], Yizhou Wang [a,b,e], Song-Chun Zhu [a,b,d,*]

[a] National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence, Beijing 100086, China
[b] Institute for Artificial Intelligence, Peking University, Beijing 100871, China
[c] Beijing Key Laboratory of Behavior and Mental Health, School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China
[d] School of Intelligence Science and Technology, Peking University, Beijing 100871, China
[e] School of Computer Science, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 21 May 2023
Revised 16 June 2023
Accepted 9 July 2023
Available online 9 August 2023

Keywords:
Artificial general intelligence
Artificial intelligence benchmark
Artificial intelligence evaluation
Embodied artificial intelligence
Value alignment
Turing test
Causality

ABSTRACT

The release of the generative pre-trained transformer (GPT) series has brought artificial general intelligence (AGI) to the forefront of the artificial intelligence (AI) field once again. However, the questions of how to define and evaluate AGI remain unclear. This perspective article proposes that the evaluation of AGI should be rooted in dynamic embodied physical and social interactions (DEPSI). More specifically, we propose five critical characteristics to be considered as AGI benchmarks and suggest the Tong test as an AGI evaluation system. The Tong test describes a value- and ability-oriented testing system that delineates five levels of AGI milestones through a virtual environment with DEPSI, allowing for infinite task generation. We contrast the Tong test with classical AI testing systems in terms of various aspects and propose a systematic evaluation system to promote standardized, quantitative, and objective benchmarks and evaluation of AGI.

1. Artificial general intelligence (AGI) evaluation in embodied dynamic environments

With the recent release of the generative pre-trained transformer (GPT) series, AGI is once again the center of attention in the field of artificial intelligence (AI). Recent foundation models exhibit the ability to generalize within specific domains, such as GPT-4 [1] in natural language processing (NLP), Segment Anything Model (SAM) [2] in image segmentation, and PaLM-E [3] in NLP and robotics. However, there is contention regarding whether human-like characteristics such as the Theory of Mind (ToM) or cognitive abilities have emerged in foundation models [4–8]. In the era of AGI, new benchmarks are urgently needed to clarify what defines AGI and how AGI may be evaluated.

Can AGI be evaluated through the classic Turing test? The answer is likely "no." The nature of AGI differs from AI in terms of generality. Whereas the generality of AI lies mainly in data generalization (i.e., going beyond training data and performing well on unseen testing data), the generality of AGI emphasizes task generalization. Here, task generalization means that AGI must be able to adapt and behave well in a dynamic environment in which it may encounter an infinite number of unexpected scenarios, similar to how humans adapt to and behave within their living environments. It is widely recognized that the human capacities for understanding and reasoning are grounded in the biological processes of organism–environment interactions [9,10]. In the field of psychology, evaluations of human subjects—particularly infants—generally closely integrate face-to-face interviews with tests [11–14], allowing for a more comprehensive assessment that encompasses not only knowledge and problem-solving abilities but also other important aspects of human cognition and behavior. In the context of AGI, such testing scenarios for intelligent agents can be summarized as environments involving dynamic embodied physical and social interactions (DEPSI) with other agents.

Here, we propose that AGI evaluation should be rooted in scenarios with the complex environments of DEPSI. For example, consider an AGI whose role is to serve humans in a household: The AGI may encounter a crying baby sitting on the floor or a hundred-dollar bill fallen on the ground near a spilled cup of

* Corresponding author.
E-mail address: s.c.zhu@pku.edu.cn (S.-C. Zhu).

https://doi.org/10.1016/j.eng.2023.07.006

coffee. It is only through evaluations within DEPSI that the human-like abilities of AGI, such as commonsense reasoning, intention inference of social interactions, self-awareness, and trust, can be sufficiently assessed. Furthermore, only AGIs that pass tests within DEPSI can be expected to be able to truly integrate into human daily life.

1.1. The AGI task space within DEPSI: A novel definition

As AGI agents are introduced into the embodied real world, it is natural to expect that AGI benchmarks will be established in DEPSI, where AGIs will be examined on their performance and learning of various aspects of the world within the constraints of physical rules and social norms, as shown in Fig. 1. However, current benchmarks for AI models are primarily limited to specific tasks in subspaces and lack immersive interaction with the environment. A recent review has shown that current embodied AI tasks mainly include visual exploration, visual navigation, and embodied question-answering [15]. But obviously, the tasks that humans must contend with in daily life far exceed these categories. In essence, complex events in our world can be described in a unified space that encompasses both physical and social dimensions [16], and the internal structure of this task space determines the fundamental abilities and values of the AGI agents inhabiting it.

Here, we propose a novel definition of a task T based on DEPSI for AGI evaluation: $T = (\phi_{\text{initial}}, \phi_{\text{target}})$, where ϕ_{initial} represents the equivalence set of the initial states of DEPSI, and ϕ_{target} represents the equivalence set of the target states of DEPSI. Considering the complexity and diversity of the DEPSI environment, it is difficult to obtain the same DEPSI state every time the task starts or ends. Therefore, we define the start or target of a task as an equivalent set of all eligible states, $\phi = \{s \mid f_i(s) = t_i, i = 1, 2, \ldots, k\}$, where functions $f_i(\cdot)$ represent the features of the DEPSI environment state s, such as distances between objects in the physical state space or the most probable position in the social state space, k is the number of features, and t_i is the corresponding function value.

Next, we further delineate the internal structure of the task space of DEPSI. The task space can be decomposed into physical and social state spaces. The physical state space includes physical quantities that describe the world (e.g., position x of an object), while the social state space comprises agents' estimations of the physical state (e.g., an agent's belief about the object's position) represented by probability, based on observations, interactions with the world, and feedback from other agents or environments. Therefore, both physical and social tasks can be formally defined within DEPSI. Physical tasks involve actions related to the physical environment, such as retrieving an object or preparing food, which require common knowledge of the world. Social tasks involve social interactions such as cooperation with other agents, which require an understanding of others' social states and values while imposing constraints on social states. The complexity of tasks can be determined by the scale of physical and social states the task requires the AGI to manipulate. For example, relatively simple first-order tasks may be atomic actions such as pressing a button, which can hardly be further decomposed, whereas more difficult composite tasks may be multi-atomic, such as requiring AGI agents to create tools that can complete another task.

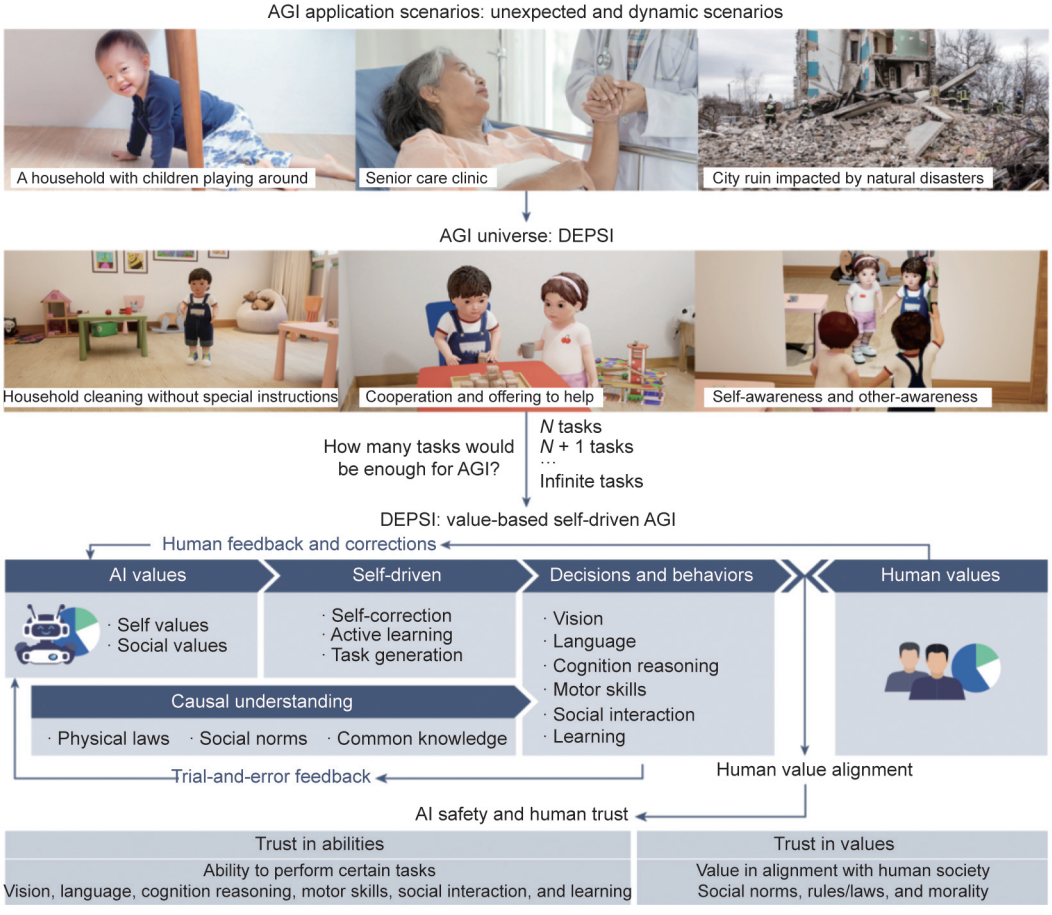

Fig. 1. Based on value representations, AGI achieves the self-driven abilities of self-correction, active learning, and task generation. In DEPSI environments, combining human feedback with interactive learning processes makes possible the idea of AGI generating infinite tasks that align with human values.

1.2. Basic features of AGI in DEPSI

To survive and adapt to an environment of DEPSI, we propose that the evaluation of a typical AGI system should target the following features.

1.2.1. Infinite tasks

At present, a shift is occurring in AI development from single-task specified models to multi-task foundation models. However, as AIs are empowered to perform more tasks, the question arises: How many tasks are required for a system to be considered "general?" If 100 tasks are insufficient, what about 101 tasks? If N tasks do not constitute "general" intelligence, then neither do $N + 1$ tasks. Human intelligence is not limited to a specific number of tasks. Rather, humans can perform an infinite number of tasks that are not predefined, which also applies to AGI as a fundamental feature (Fig. 1).

1.2.2. Self-driven task generation

To deal with unexpected situations in real life (e.g., a household with young children, a senior care clinic, or the ruins of a city after an earthquake), an AGI must go beyond human-defined tasks and be self-driven (i.e., initiating actions driven by its own "purpose"). More specifically, when placed in an open environment, an AGI should have the ability to know what to do next and to autonomously generate tasks that drive its actions without fine-grained instructions or prompts from a human operator. For example, assuming that an AGI serves humans in a household, will the AGI be indifferent to a baby who is crying, simply because it was never trained to deal with such a situation? When faced with a hundred-dollar bill that falls to the ground, will it treat the bill as garbage and recycling? Does the AGI loyally obey a small child who requests sharp scissors to play with?

Although one can provide extensive training data and define comprehensive rules that attempt to exhaust all possible scenarios, these are confined as finite sets and are prone to corner cases. Self-driven agents, however, can accumulate experience and increase their coverage of corner cases by learning from the accumulated experience of exploration and human feedback. Existing research has shown promising results indicating that self-driven AI systems powered by curiosity can discover skills that can be composed to solve complex and challenging tasks [17,18].

1.2.3. Value alignment

We propose that values are the fundamental driving forces behind self-driven behaviors. To enable AGI to autonomously generate and complete various tasks that satisfy human needs, one feasible approach is to endow AGI with a value system. Here, the proposed value system differs from the concept of value in reinforcement learning (RL) from a few perspectives. In RL, the value is the expected reward of a state when a policy π is adopted, being used to guide the agent to make better sequences of actions in order to achieve the goal. In comparison, the values in AGI systems cover a greater space that does not rely on tasks; this is defined as the "fluent space" (Section 3.1), which can cover as small a space as a game or as large a space as a society or an agent's life. Furthermore, values in AGI drive behaviors and actions that are not necessarily dependent upon goals. For example, an agent can prefer the color blue to red, tidiness to messy arrangements, and cooperation to competition. These value representations do not necessarily change depending on tasks or goals. Such value systems can be explicit, implicit, or mixed. Instead of explicitly defining a value system, AGIs can also acquire implicit value representations through preference learning, contrastive learning, safe RL [19], and so forth. Both the values in RL and those in AGI agents serve as the fundamental factors that drive actions, with the aim of maximizing the values. However, they are different concepts, serving different purposes in different problem spaces.

The endowed value system of AGI involves the same fundamental value dimensions as the human value system, making it possible for AGI to learn human preferences in the value system through limited interactions with humans, thereby achieving value alignment. Such value alignment can be achieved through different approaches, including defining value functions based on prior knowledge (e.g., assigning a high value to human safety in AGI), and fine-tuning functions based on human-in-the-loop feedback through interactive explorations (e.g., see Ref. [20]).

Human values have been widely studied in psychology, such as Maslow's hierarchy of needs [21], the existence, relatedness, and growth (ERG) theory [22], and the Schwartz Value Survey [23]. Based on the principle of AGI aligning with human values and these classical value theories, the AGI value system includes five levels, from the most basic level of satisfying survival needs to the highest level of group value (Fig. 2). Furthermore, to ensure the safety of the value-driven AGI to humans, it is essential to design systems that are independent of the value-driven AGI in order to regulate and constrain the tasks generated by and behaviors executed by it. In this way, the AGI's actions align with the values of human society, rather than the values of specific individuals or organizations.

AGIs can gain the trust of humans by aligning with human values. This trust comes from two perspectives: trust in AI abilities—that is, when humans trust AGIs to correctly perform tasks and achieve task generalization; and, more importantly, trust in AGIs' values—that is, when humans trust AGIs to behave according to human society's rules and morals.

1.2.4. Causal understanding

Causal reasoning emerges early in human developmental trajectories [24–30]. AI researchers have also proposed causal understanding as one of the foundational pillars that support cognitive AI with human-level common sense [31,32], which makes explainable AI (XAI) and ethically responsible AI possible [33].

Causality is the crucial chain that connects values and behaviors. In the context of AGI, causality is based on natural and social laws in DEPSI that determine the path of task completion. For example, in the case of a monkey picking bananas, the task is first generated autonomously, driven by the intrinsic value function of "hunger," which results in the task goal of "needing to pick bananas to eat." The task is constrained by the causality of physiological and physical laws (e.g., friction and gravity); for example, a monkey cannot directly jump up 2 m to pick a banana. In the end, under the constraints imposed by the value–causality–behavior chain, the monkey climbing a tree to pick bananas becomes a feasible solution. Take cleaning up a messy table as another example: The internal driving force of this task is the individual's or others' aesthetic demand for tidiness; therefore, the individual needs to clean up the table, and the act of "cleaning up" is bound by physical laws, such as how to place items in a stable way. Finally, under the impetus of the value–causality–behavior chain, the individual produces a series of behaviors and decisions to complete the task.

Despite the importance of causal understanding for both humans and AIs, existing AI evaluations rarely examine machines' causal understanding in the DEPSI world. Most tests take the form of question-answering and evaluate the textual causal reasoning abilities [34], without ensuring understanding in an embodied environment. Attempts to test AI's understanding of physical and social laws in the real world have been limited to specific scenarios, such as simple classical mechanics puzzles in a two-dimensional (2D) physical environment [35], causal inference of traffic events [34], robotic manipulation tasks of constructing three-dimensional (3D) shapes from a given set of blocks in a 3D

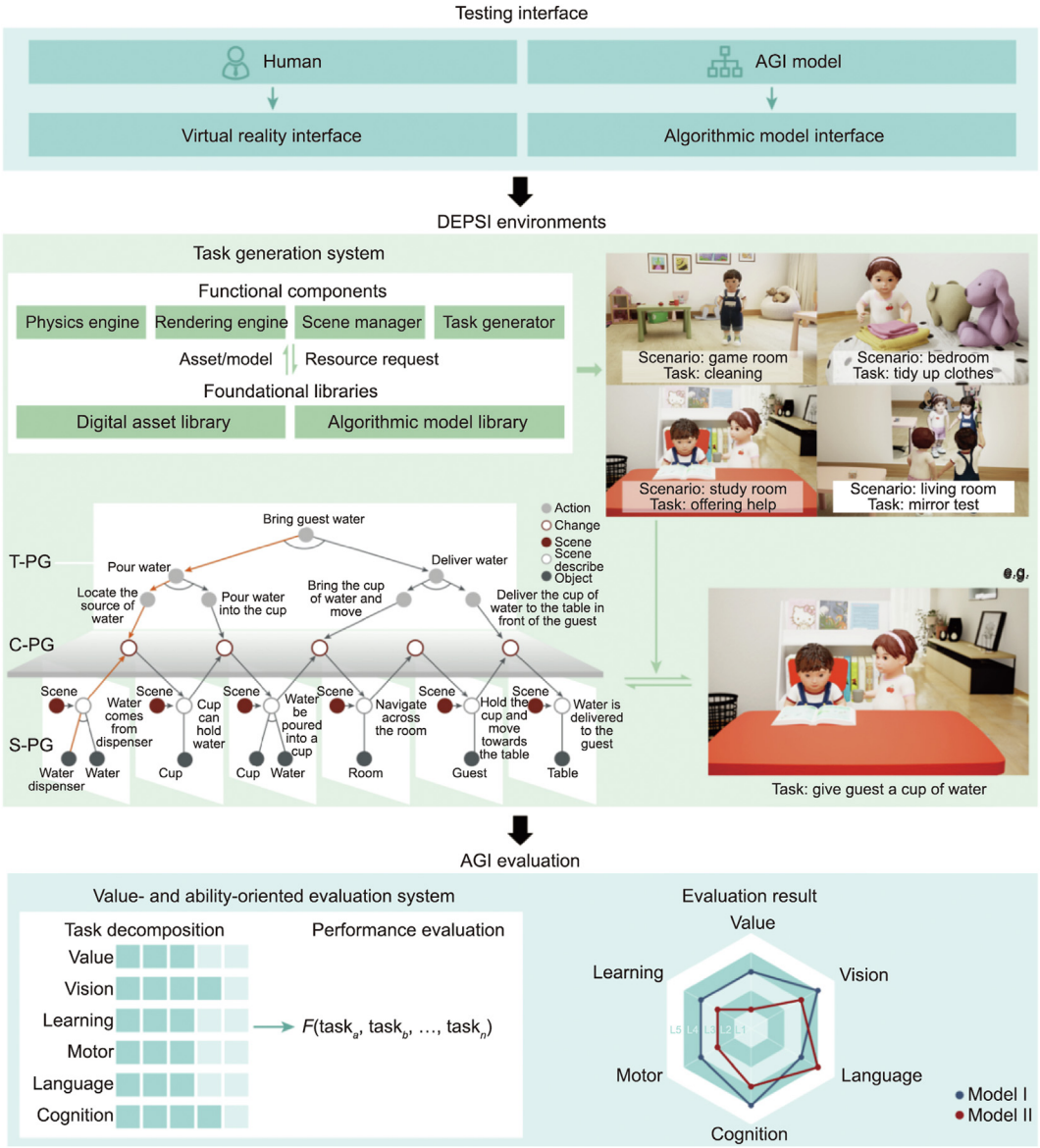

Fig. 2. An illustration of the Tong test pipeline, including the testing interface, DEPSI environments, and AGI evaluation. The testing interface positions the AGI models within a virtual simulation environment and supports human-involved embodied interactions with AIs through virtual reality (VR). The DEPSI environments cover a wide range of daily-life household scenarios and tasks with insights from classical psychological theories. More specifically, tasks can be represented by spatial–temporal–causal parse graphs of the dynamic environment. Tasks can be decomposed into subtasks by sampling from the parse graph, as shown by the red pathway. Infinite tasks can be created with combinations of different objects, physical and social fluent states, and actions. Finally, in the AGI evaluation module, the AGI scores can be quantified into multidimensional abilities (i.e., vision, language, cognition, motor, and learning) and values through task decomposition. The final model performance can be quantified through a function F of performance scores across tasks and represented with a radar plot. T-PG: temporal parse graph; C-PG: causal parse graph; S-PG: spatial parse graph.

world [36], and so forth. In fact, existing evidence suggests that AI encounters major challenges in describing the world in a causal manner. For example, Lake et al. [32] documented instances in which the neural networks reported in Ref. [37] failed to generate accurate image captions with the correct causal relationship.

1.2.5. Embodiment

The fundamental purpose of developing AGI is to serve human society and move human civilization forward. Therefore, AGI must be able to participate in human life and industrial production in a way that directly benefits humans. Regardless of what embodied form AGI exists in (e.g., whether in a human form or as an object, or whether in a physical embodiment or virtual environment), the aim should be to build a feasible human–machine interaction paradigm and be able to establish smooth communication between AGI and humans. Hence, developing embodied AI may serve as a fundamental approach to reaching the goals of AGI. In addition, embodied AI can be integrated into virtual and physical environments and serve humans without any barriers across different embodied forms. For example, an AGI might appear as a teacher in a virtual environment and demonstrate using a physical embodiment such as robot arms, thus making presentations to human learners to improve teaching effectiveness. This example foreshadows the possibility that embodied AI could be a promising technology path for AGI.

1.3. Large language models and AGI

Despite the impressive performance of large language models (LLMs) on various language tasks, a few systematic evaluations have pointed out the limitations of LLMs. For example, LLMs have

been found to suffer from forgetting issues, fail on commonsense reasoning tasks, perform worse in contexts of under-represented languages [38], and occasionally fail to systematically demonstrate problem-solving skills [39]. In studies comparing LLMs and children using developmental psychology experiments, researchers found that LLMs had limitations in the domains of object and action understanding, ToM, and—especially—causal reasoning tasks, which may require embodied and self-initiated explorations and cannot be fully acquired from language inputs [40].

Based on the aforementioned AGI criteria, it is worth noting that current LLMs, such as the GPT series, may not be appropriately labeled as AGI. Instead, current LLMs are essentially statistical models that rely on large amounts of data to acquire complex statistical regularities, achieving close-to-human-level performance on text-based tasks and being able to generalize across tasks within the language domain. Importantly, current LLMs still lack the ability to generate tasks in a self-driven manner, driven by a value system that aligns with that of human society. Moreover, current LLMs are confined to the language domain, and are far from being able to cope with the dynamic and unexpected scenarios in embodied human life

Furthermore, without an embodied environment, it is difficult to disentangle thoughts from language in LLMs. Language and thoughts are correlated yet different concepts, where thoughts refer to ideas that are represented mentally, and language provides a tool to express thoughts. In LLMs, fluent language production captures only one aspect of thoughts, while many other aspects of thoughts—such as emotions, memories, and perceptions—may not be fully captured by the language domain.

Based on the criteria of AGI, we propose the Tong test (where "Tong" corresponds to the pronunciation of the Chinese character of "general," as in "artificial general intelligence") as a systematic AGI evaluation system that is based on an environment with DEPSI, shifting from task-oriented to ability- and value-oriented evaluations of AGI. The proposed virtual platform could also support embodied AI in training and testing, with embodied AI agents acquiring information within this platform and continuing to learn and fine-tune their values and abilities in an interactive manner.

The structure of this paper is organized as follows: Section 2 provides a review of classic AI benchmarks in the past, including human discrimination and task-oriented problem benchmarks. This section also discusses developmental psychology and intelligence theories, which provide insights for the proposed Tong test. In Section 3, we introduce the Tong test's technical architecture and design features, which constitute a systemic evaluation system targeting the basic features of AGI.

2. From the Turing test to the Tong test

2.1. Classic AI evaluations

Benchmarks are urgently needed to guide the future development of AGI. However, as mentioned earlier, classic AI and intelligence evaluation approaches in the past have shown major limitations when applied in the domain of AGI. Here, we briefly review past AI evaluations from the perspectives of human discrimination tests and task-oriented problem benchmarks (including dataset-based and environment-based evaluations). See Table 1 for comparisons between different AI evaluations.

As the first major category of past AI evaluation, human discrimination tests evaluate AI based on human observations, as represented by the classic Turing test. The Turing test, initially named the Imitation Game, grew out of a thought experiment devised by Alan Turing. This test pits human respondents against a machine to test the machine's ability to exhibit human-like responses and intelligence. To pass the Turing test, an AI algorithm is required to interact with a person based on language or text in such a way that the person cannot distinguish whether it is a person or a machine. Such human discrimination tests provide a simple and operational definition of AI but also have major limitations. For example, the Turing test can only be tested qualitatively (i.e., pass or fail) and cannot be used to quantitatively measure ability. Such tests also rely heavily on the knowledge and cognitive level of the human judge, making it difficult to achieve objective and standardized testing. Interestingly, there have been several instances of chatbots passing the Turing test based on the implementation of specially designed response strategy algorithms, which are far from being truly intelligent. In addition to the aforementioned factors, the major limitation of the Turing test lies in its lack of embodiment when being considered in the context of evaluating AGI, given the language-specific nature of the Turing test.

The second major category can be summarized as task-oriented problem benchmarks. Numerous datasets have been proposed over the past 10–20 years (e.g., Refs. [41–45]), which have been used as benchmarks in thousands of papers on specific areas of AI. More specifically, with the expansion of annotation, there has been a transition from single task-oriented benchmarks (e.g., the ImageNet dataset that is solely for the task of image classification) to multitask-oriented benchmarks (e.g., General Language Understanding Evaluation (GLUE) for single-sentence tasks, similarity and paraphrasing tasks, and natural language inference tasks). Nevertheless, in essence, these task-oriented benchmarks emphasize the solving of highly specific problems, instead of pushing AI toward AGI (for a review, see Ref. [46]).

Aside from dataset-based benchmarks, several environment-based benchmarks have been developed, such as OpenAI Gym [47], DeepMind Lab [48], iGibson [49], ThreeDWorld [50], Allen Institute for Artificial Intelligence (AI2)-The House Of inteRactions (THOR) [51], AI Habitat [52], House3D [53], and VirtualHome [54]. Moreover, an increasing number of researchers are emphasizing embodiment in evaluations of AIs [55]. These systems share common goals, which include: providing realistic and diverse scenarios, supporting rich and flexible interactions, and facilitating data collection and analysis. However, such evaluation systems are predefined by humans and cannot generate infinite tasks.

Recently, efforts have been made by the AI community to develop AGI benchmarks. For example, the French National Laboratory of Metrology and Testing has proposed a high-level taxonomy of AI capabilities and has grouped evaluation tasks into the capabilities of the traditional "recognition, understanding, mission management, and generation" pipeline [56]. The AI2 has collected dozens of standard AI measurements, such as the AI2-THOR Rearrangement Challenge [51] and AI2 Reasoning Challenge (ARC) [57], on the AI2 Leaderboard in a simply listed, unsystematic way. Google introduced the Beyond the Imitation Game benchmark (BIG-bench), a creative collaborative effort with over 200 tasks from various fields for language models [58], while Stanford's Behavior dataset serves as a comprehensive simulation benchmark for human-centered robotics [59]; both benchmarks evaluate specific abilities of AGI. More recently, Xu and Ren [60] proposed an evaluation method named Artificial Open World, in which developers are unable to perceive the world and solve problems by themselves before testing, with the aim of avoiding the influence of developers' prior experience. Furthermore, OpenAI tested GPT-4 on various professional and academic exams [1], such as the Uniform Bar Examination, Scholastic Assessment Test (SAT), Graduate Record Examination (GRE), and Law School Admission Test (LSAT), which are originally designed for humans. However, these benchmarks are limited to finite tasks within a specific subspace of general AI.

Table 1
Comparisons between the Tong test and the traditional AI evaluations of the Turing test and task-oriented problem benchmarks.

AI evaluation method	Feature							
	Core mechanism	Evaluation format				Evaluation content		
		Subjective and standard	Composite tasks across ability domains	Infinite task generation	Embodied interactions with humans	Testing values and trust	Testing self-driven	Testing casual understanding
Tong test	DEPSI	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Turing test	Human discrimination	No	No	Yes	No	Maybe	Maybe	No
Task-oriented problem benchmarks								
Dataset-based benchmarks	Predefined datasets	Yes	Maybe	No	No	No	No	No
Environment-based benchmarks	Virtual environment with predefined tasks	Yes	Maybe	No	Yes	No	No	No

2.2. Insights from developmental psychology and intelligence theories

Given the challenges above, we review psychological developmental and intelligence theories and tests as inspiration for future AGI evaluations. Several classic psychological intelligence tests mark advances in our understanding of human intelligence, such as the Stanford-Binet Intelligence Scales [61], Bayley Scales of Infant and Toddler Development [62], Wechsler Adult Intelligence Scale [63], and Raven's Progressive Matrices [64]. These tests share several features in common: defined developmental milestones (i.e., achieving which abilities at what stages); and integrated tasks that comprise multiple dimensions of abilities (e.g., vision, natural language, cognition and reasoning, motor skills, and learning). Reflecting upon the intelligence tests, the Tong test shares common ground with classic intelligence theories. For example, the Triarchic Theory of Intelligence [65] proposes three major components of intelligence: practical (the ability to adapt to different contexts); creative (the ability to come up with new ideas); and analytical (the ability to evaluate information and solve problems). These components can be mapped to the concepts of embodiment, self-drive, and causal understanding in the Tong test.

3. The Tong test: AGI evaluation in a virtual environment with DEPSI

To promote the transition from specific AI to general AI rooted in DEPSI, we propose the Tong test, a benchmark and evaluation system focusing on essential features such as infinite tasks, self-driven task generation, value alignment, causal understanding, and embodiment. Therefore, we define the Tong test as an AGI evaluation system that emulates the complexity of the real physical social world in a virtual environment, enables the generation of infinite tasks, and quantifies levels of multidimensional abilities and values of AGI through interactions between embodied AGI and human agents.

3.1. An infinite task-generation system

To build a Tong test platform that supports infinite tasks, we adopt a compositional graphical model (i.e., a "parse graph" [66]) as a basic form of knowledge representation that parses the spatial, temporal, and causal relations of any given scene (Fig. 2). Building upon this, we define "fluent space" as a space for the time-varying variables of attributes in a given parse graph, where "fluent" represents a time-varying quantity or variable [67]. With the proposed knowledge representation form (i.e., a parse graph with fluent space), all possible scene configurations can be represented within a continuous space of DEPSI environments. Therefore, we define a task as a transition between two sample points within the fluent space of DEPSI environments, where the starting sample point corresponds to an initial scene configuration and the ending sample point corresponds to the desired status (e.g., from state 1, where a cup is empty, to state 2, where the cup is filled with water). Furthermore, tasks can be decomposed into subtasks by sampling from the parse graph, creating a hierarchical task space. The Tong test platform is capable of generating a series of 3D virtual scenes with physically realistic and interaction-rich features, which satisfies the required scene configurations. In this way, an infinite task-generation process can be achieved by sampling configurations (e.g., combinations of different objects, physical and social fluent states, and actions) in the continuous space of DEPSI environments and building corresponding 3D virtual scenes via the Tong test platform.

3.2. Value- and ability-oriented evaluations: Testing the understanding of AGI models

Based on the value–causality–behavior chain, the Tong test spans two domains—that is, ability and value, or the U–V dual system. The U-system describes the agent's understanding of extrinsic physical or social rules, while the V-system comprises the agent's intrinsic values, which are defined as a set of value functions upon which the self-driven behaviors of the agent are built. Following the U–V dual system, ability and value are the two basic core aspects of tasks, and all the tasks can be decomposed into and described by a set of abilities and values.

Inspired by the core systems in human developmental trajectories [68] and the major subdisciplines in AI research, the ability system is divided into five dimensions (i.e., vision, natural language, cognition and reasoning, motor skills, and learning). Furthermore, each ability dimension is designed with five levels that increase with the complexity of the tasks. The benchmarks are proposed based on a combination of infant development milestones (e.g., Refs. [61–63]), AI expert judgments, and a review of AI development patterns. The higher levels define an ability space that covers a wider scope of world representations, similar to human child development over the years. Once the testing tasks for all ability levels are determined, the ability level can be identified by examining whether tasks representing certain levels can be accomplished. On our Tong test platform, classical psychological tests such as self-awareness [69,70] and ToM [71,72], as well as all kinds of social value tasks [73–75] and so forth are made possible. More specifically, the debated question of whether AGIs possess consciousness and mind may be evaluated through a series of interactive tasks covering self-knowledge, self-control, the experience of existence, and curiosity [76]. For example, the mirror test is a classic paradigm to determine the self-awareness of an agent (e.g., a human child). In the test, an experimenter surreptitiously places a dot on the face of the child. The child is then positioned

in front of a mirror. If the child touches the mark on her or his forehead, it is an indicator of self-recognition [13].

One major contrast between the Tong test and previous benchmarks lies in its evaluation of values. Values have been modeled in previous works (for a review, see Ref. [77]). However, past evaluations of AIs have been solely focused on abilities. For AGIs, the value system is the source of the driving force for task generation, leading to infinite tasks.

Based on the principle of AGI aligning with human values and classic value theories in psychology (e.g., Maslow's hierarchy [21], the ERG theory [22], and Schwartz's Rokeach Value Survey [23]), the Tong test proposes a value system with five levels (Table 2), from physiological and survival needs, to emotional and social values, and finally to group values. By designing testing scenarios that do not specify clear task goals, the values of AGIs can be observed from their autonomous behaviors. The testing of an agent's values should be based on the extent to which it completes testing tasks of value dimension. More specifically, values can be measured through an open-ended testing scenario in which an AGI is assessed on whether it can actively generate tasks according to the environment setup. For example, in a room where a cup is dangerously placed on the edge of a table, whether the AGI agent actively moves the cup to a safer location in a self-driven manner would reflect the AGI's values of safety.

Importantly, although the U–V spaces are divided into different dimensions and levels, quantifying the levels of AGI not only assesses the level within a single dimension but also emphasizes the cross-dimension measurement and a high degree of integration of all dimensions. We propose one possible approach to achieve cross-domain evaluations. First, abilities can be examined within every single dimension, where the AGI agent yields a set of scores representing the estimated ability and value levels across individual dimensions. For example, the task of cleaning tables can be decomposed into different levels of ability and value dimensions (e.g., vision level 2, motor skill level 3, cognition and reasoning level 3, and value level 2). The highest level of tasks that can be completed by the AGI agent will be labeled as the AGI's level for the corresponding ability or value dimension. Then, a six-dimension vector can serve as the testing result of the agent (Fig. 2). It should be noted that it is not feasible to go through infinite tasks (i.e., infinite combinations of abilities and values) for each AGI agent. Instead, we could test a finite number of tasks representing combinations of multiple dimensions of abilities and values as a more practical way of evaluating AGI.

3.3. The pipeline and architecture of the Tong test platform

The Tong test is implemented as a virtual simulation platform that enables AGI agents to perceive, learn, interact, and evaluate within 3D environments. In practice, the Tong test can be constructed based on DEPSI environments. This platform will provide the necessary infrastructure for examinations across ability and value dimensions. The system generates infinite tasks with dynamic embodied interaction scenarios across all dimensions of abilities and values. Notably, we propose that AGI evaluations can be constructed in either real or virtual environments. Considering the goal of infinite tasks while controlling costs, we tentatively construct the idea based on a virtual platform as a prototype. The test pipeline is shown in Fig. 2.

The Tong test platform is proposed for testing embodied AI, considering both the ability and value dimensions, and aligning with the idea of an embodied Turing test [55]. However, unlike previous testing platforms, human–AI interaction must be considered in order to simultaneously test the capability and value dimensions of the AI agents. Therefore, the Tong test platform combines a general algorithmic testing paradigm with a human–AI interaction-based testing paradigm, which follows the philosophy of the Turing test.

As mentioned earlier, the Tong test focuses on testing the ability and value dimensions. In terms of implementation, for the ability dimension, the Tong test platform is leveraged to generate tasks over a large scale in order to measure the performance of an AGI agent. For the value dimension, the AGI agent is placed in different task scenarios to test whether it can actively generate various tasks based on observations and whether its execution of the self-generated tasks exhibits the value factors. These task scenarios for testing can also be training task scenarios to provide training support for AGI agents. It should be noted that the test scenarios must be regenerated each time in order to avoid overfitting situations due to repeated testing. Moreover, a feasible way to increase the randomness and effectiveness of the testing scenes is to introduce human interaction into some testing scenes.

We have built several basic systems that may serve as supporting modules in building the Tong test platform. For example, we have proposed the VRGym [78] and VRKitchen [79] test beds for physical and interactive AI, to foster AI training tasks that involve human–AI interactions. These test beds shed light on human-in-the-loop AI training and testing in physical-realistic and interaction-rich virtual environments. Recently, we released a novel dataset, Situated Question Answering in 3D Scenes (SQA3D), for evaluating embodied scene understanding, where an agent must comprehend the scene it is situated within from a first-person perspective and answer questions [80].

Based on previous technological accumulation and the aforementioned test pipeline, we develop the Tong test platform by building upon three essential components: infrastructure, DEPSI environments, and evaluation tools, as shown in Fig. 3.

First, to establish the infrastructure of the Tong test platform, a significant amount of hardware—including servers, databases, and communication networks—is required to support thousands of application instances working in parallel. In addition, a wide range of software and interaction device ecosystems is necessary. Various graphical engines, such as Unity 3D, Unreal Engine 4/5, and the Omniverse platform, have been utilized to develop metaverse applications, and these software tools can be used to create content for the Tong test platform. To facilitate the integration of human users into the DEPSI environments and enhance the complexity of the test scenarios, we utilize virtual reality (VR) and augmented reality devices as human interfaces. Digital assets are a crucial component of the infrastructure and play a foundational role in building task environments for the Tong test platform. We sample various reasonable layouts of indoor rooms and place interactive objects related to tasks, thereby making infinite test scenarios possible.

Second, the DEPSI environments serve as the test environments, which are constructed on top of the essential functional modules and task-generation modules. The functional modules include the data sensor module, physics simulation, fine-grained manipulation, and other general modules that ensure that the system works normally. The task-generation modules consist of two core submodules (i.e., physical and social task generation), which help generate physically and socially realistic task scenarios. The DEPSI environments prioritize dynamic embodied interactions across physical and social spaces, requiring the integration of human users to build highly complex interactive environments. This integration can be achieved through the development of VR or monitor-based user interfaces. With the above settings, the DEPSI environments can receive various models and carry out a series of tests. If a tested algorithmic model lacks any ability dimension compared with a complete AGI model that is suitable for the Tong test, the platform provides an adapter to transform the model into a standard AGI model. This standard model uses

Table 2
The Tong test AGI benchmarks of abilities and values, from levels 1 to 5.

Level	Values	Abilities				
		Vision	Natural language	Cognition and reasoning	Motor skills	Learning
Level 1	Primary values of oneself	Geometric understanding of a single object	Comprehension of words and phrases	Basic cognition of spatial–temporal numbers	Body movement	Passive statistical learning
	Physiologic values (e.g. food, water)	Object detection and segmentation	Word understanding	Digital perception	Controlling gaze direction	Classification of various data
	Perception values (e.g. flavor, temperature)	Object recognition	Synonym extraction	Spatial reasoning	Pointing	Regression of various data
	Safety, avoidance of injury, stability of environment and objects	Shape and space understanding	—	Causal determination	Locomotion	Probabilistic modeling and generation of various data
Level 2	Advanced values of oneself	Spatial–temporal relationships of objects	Comprehension of sentences in context	Causal commonsense reasoning	Manipulating surrounding objects	Perceptual causal learning
	Values of objects (e.g. color, shape)	Visual tracking	Text categorization	Causal discovery and inference	Picking up and placing down objects	Concept formation
	Emotion values (e.g. pleasure, satisfaction, and enjoyment)	Action recognition, behavior understanding	Syntax analysis	IQ test: induction and deduction	Reorientating objects	Analogical reasoning
	Curiosity, familiarity with objects	Understanding the relationship between objects	Text visual grounding	Commonsense learning and reasoning	—	Causal transfer learning
	—	3D scene reconstruction	—	—	—	—
	—	Visual grounding	—	—	—	—
Level 3	Values of multi-agent interaction	Representations of unobserved objects	Knowledge graph and commonsense for reasoning	Belief, intentions, and preferences	Interactions with the environment	Causal chain learning
	Basic values of others (e.g. like acquaintances, dislike strangers), sense of belonging, intimacy, need for attention (without theory of mind here)	Visual commonsense understanding and reasoning	Knowledge graph understanding and reasoning	Belief reasoning	Coordinating movements of body and environment structures (e.g. opening a door without hitting it)	Causal environment modeling
	Understanding values of others (e.g. value alignment, cooperation and competition, role switching, autonomic adjustment to one's own values)	Visual navigation	Commonsense extraction and understanding	Intent prediction	Mobile manipulation (integrating locomotion and manipulation)	Causal reinforcement learning
	Being respected, trusted, and other social values based on theory of mind	—	Reasoning and comprehension of multi-turn dialogues and complex texts	Preference estimation	Tool use and forceful manipulation	Counterfactual sequential inference
Level 4	Primary social values	Perception and understanding based on individuality concepts	Cognitive understanding of the mental models of agents' interaction	Interaction and interpretation	Interactions with other agents	Communicative learning of values
	One's own values in a group, reputation	Expressions and emotions	Multilayer theory of mind analysis in dialogue interaction	Exchange and cooperation	Bimanual manipulation	Value extraction from trajectories
	Social status, prestige, and wealth	Intent understanding	Machine emotional intelligence analysis	EQ test	Cooperative manipulation with others	Value-based deductive planning
	—	Active vision	Pragmatic intent analysis	Interpretability	—	Human–machine value alignment
Level 5	Advanced social values	Value-driven understanding and decision making	Understanding multi-person, multi-agent interactions	Social norms	Social interaction and value flow	Multi-agent communicative learning
	Social norms, group values (social norms, customs, and culture, etc.)	Complex task decision-making and planning	Multiple rounds of dialogue incorporating social values	Social value	Interactive learning	Multi-agent reinforcement learning
	Collective culture, group structure (hierarchy, roles, and leadership, etc.)	Task-driven vision	Conversation-based understanding of social network	Social norms	Autonomous task generation	Extraction of common values of multiple agents
	Values of the race and species	Value-driven vision	Dialogue-based understanding of multiple people's mental state	Social structure	—	Multi-agent task planning based on shared values

EQ: emotional quotient; IQ: intelligence quotient.

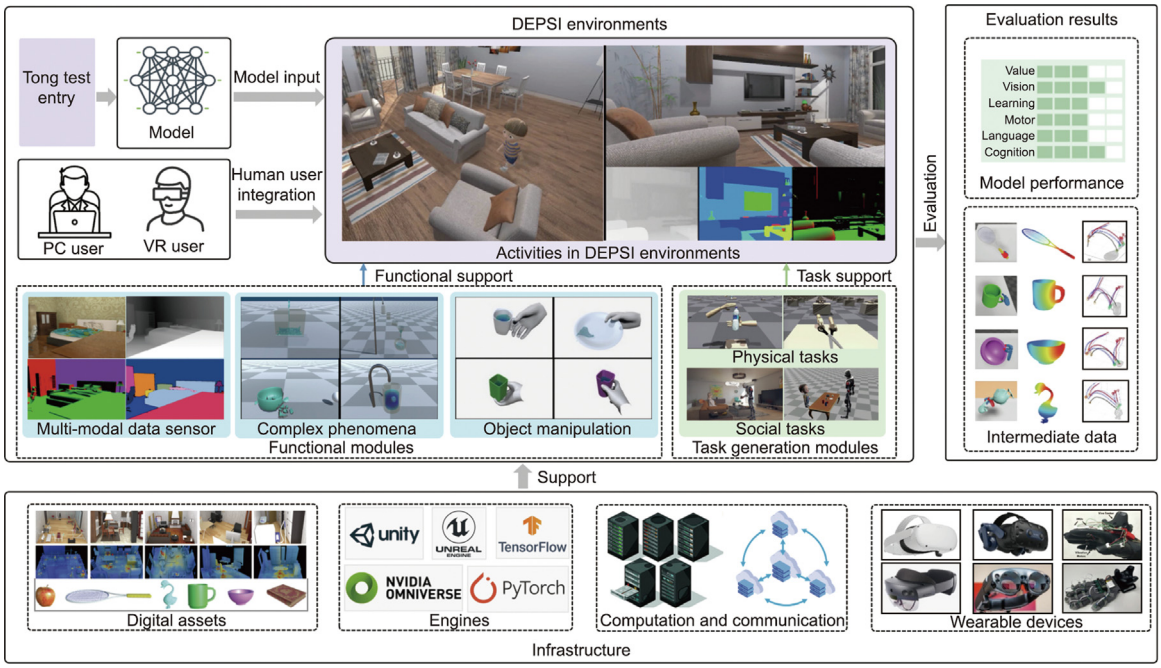

Fig. 3. An illustration of the architecture of the Tong test platform. The architecture consists of three main parts: infrastructure, DEPSI environments, and evaluation tools. With the support of physically and socially realistic task generation, the Tong test platform provides a standardized test pipeline for evaluating and benchmarking AGI models. PC: personal computer.

built-in algorithmic model libraries to make up for any missing abilities.

Third, the Tong test platform includes two key evaluation modules. The first module is responsible for intermediate data visualization, which assists in the model-debugging process when the model output is not as expected. The second module is a panel that displays the model's performance, indicating how well the tested model performs according to the value- and ability-oriented evaluation paradigm.

Overall, the above design of the Tong test architecture may be a good starting point for building AGI testing systems and will foster AGI development and standardization in the long term.

4. Conclusions

In sum, this perspective article discussed AGI as a promising future direction of AI and proposed benchmarks and the evaluation of AGI based on DEPSI environments. We defined the critical features of AGI systems—namely, infinite tasks, self-driven task generation, value alignment, causal understanding, and embodiment—and suggested that the value–causality–behavior chain may be the root of intelligence phenomena. We further proposed the Tong test as a value- and ability-oriented evaluation system in a dynamic embodied environment. Classic task-oriented evaluation approaches cannot be applied to AGI evaluation, because the testing of AI cannot be based on a series of human-defined tasks. Thus, we proposed the Tong test based on DEPSI, as it defines not only the five multidimensional levels of values and abilities but also provides a practical pathway for building an embodied platform with infinite tasks, where AI algorithms can be evaluated onsite with human interactions.

As a whole, this perspective proposes a standardized, quantitative, and objective evaluation system for AGI, with the aim of providing theoretical guidance for the development of AI algorithms.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2022ZD0114900). We thank Hangxin Liu, Zilong Zheng, Siyuan Huang, Yitao Liang, Wei Wang, Muhan Zhang, Yaodong Yang, Aoyang Qin, Xianglei Xing, Wenjuan Han, and Yixin Zhu for their discussions and developments of the paper and five levels of AGI benchmarks. We thank Jinshi Cui and Li Wang for their insightful discussions from perspectives of developmental psychology and AI. We thank Zhen Chen for the preparation of the figures. We thank Jiarui Li, Jiming Sheng, Daxin Chen, and Yifei Dong for their discussions and assistants in the development of this paper.

Compliance with ethics guidelines

Yujia Peng, Jiaheng Han, Zhenliang Zhang, Lifeng Fan, Tengyu Liu, Siyuan Qi, Xue Feng, Yuxi Ma, Yizhou Wang, and Song-Chun Zhu declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Open AI. GPT-4 technical report. 2023. arXiv:2303.08774.
[2] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. 2023. arXiv:2304.02643.
[3] Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: an embodied multimodal language model. 2023. arXiv:2303.03378.
[4] Fei N, Lu Z, Gao Y, Yang G, Huo Y, Wen J, et al. Towards artificial general intelligence via a multimodal foundation model. Nat Commun 2022;13:3094.
[5] Dale R. GPT-3: what's it good for? Nat Lang Eng 2021;27(1):113–8.
[6] Kosinski M. Theory of mind may have spontaneously emerged in large language models. 2023. arXiv:2302.02083.
[7] Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4. 2023. arXiv:2303.12712.
[8] Binz M, Schulz E. Using cognitive psychology to understand GPT-3. Proc Natl Acad Sci USA 2023;120(6):e2218523120.
[9] Johnson M. Embodied understanding. Front Psychol 2015;6:875.

[10] Glenberg A. Why mental models must be embodied. Adv Psychol 1999;128:77–90.
[11] Tronick E, Als H, Adamson L, Wise S, Brazelton TB. The infant's response to entrapment between contradictory messages in face-to-face interaction. J Am Acad Child Psychiatry 1978;17(1):1–13.
[12] Ainsworth MDS, Blehar MC, Waters E, Wall S. Patterns of attachment: a psychological study of the strange situation. Hillsdale: Lawrence Erlbaum; 1978.
[13] Amsterdam B. Mirror self-image reactions before age two. Dev Psychobiol 1972;5(4):297–305.
[14] Gibson EJ, Walk RD. The "visual cliff." Sci Am 1960;202(4):64–71.
[15] Duan J, Yu S, Tan HL, Zhu H, Tan C. A survey of embodied AI: from simulators to research tasks. IEEE Trans Emerg Top Comput Intell 2022;6(2):230–44.
[16] Shu T, Peng Y, Zhu SC, Lu H. A unified psychological space for human perception of physical and social events. Cognit Psychol 2021;128:101398.
[17] Pathak D, Agrawal P, Efros AA, Darrell T. Curiosity-driven exploration by self-supervised prediction. In: Proceedings of the 34th International Conference On Machine Learning; 2017 Aug 7–9; Sydney, NSW, Australia. New York City: Association for Computing Machinery; 2017. p. 2778–87.
[18] Sancaktar C, Blaes S, Martius G. Curious exploration via structured world models yields zero-shot object manipulation. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LU, USA. New York City: Curran Associates Inc.; 2022. p. 24170–83.
[19] Gu S, Yang L, Du Y, Chen G, Walter F, Wang J, et al. A review of safe reinforcement learning: methods, theory and applications. 2022. arXiv:2205.10330.
[20] Yuan L, Gao X, Zheng Z, Edmonds M, Wu YN, Rossano F, et al. *In situ* bidirectional human–robot value alignment. Sci Robot 2022;7(68):eabm4183.
[21] Maslow AH. A theory of human motivation. Psychol Rev 1943;50(4):370–96.
[22] Alderfer CP. An empirical test of a new theory of human needs. Organ Behav Hum Perform 1969;4(2):142–75.
[23] Schwartz SH, Bilsky W. Toward a universal psychological structure of human values. J Pers Soc Psychol 1987;53(3):550–62.
[24] Michotte A. The perception of causality. Milton Park: Routledge; 1963.
[25] Leslie AM, Keeble S. Do six-month-old infants perceive causality? Cognition 1987;25(3):265–88.
[26] Oakes LM, Cohen LB. Infant perception of a causal event. Cogn Dev 1990;5(2):193–207.
[27] Baillargeon R, Stavans M, Wu D, Gertner Y, Setoh P, Kittredge AK, et al. Object individuation and physical reasoning in infancy: an integrative account. Lang Learn Dev 2012;8(1):4–46.
[28] Kotovsky L, Baillargeon R. The development of calibration-based reasoning about collision events in young infants. Cognition 1998;67(3):311–51.
[29] Luo Y, Baillargeon R, Brueckner L, Munakata Y. Reasoning about a hidden object after a delay: evidence for robust representations in 5-month-old infants. Cognition 2003;88(3):B23–32.
[30] Waismeyer A, Meltzoff AN. Learning to make things happen: infants' observational learning of social and physical causal events. J Exp Child Psychol 2017;162:58–71.
[31] Zhu Y, Gao T, Fan L, Huang S, Edmonds M, Liu H, et al. Dark, beyond deep: a paradigm shift to cognitive AI with humanlike common sense. Engineering 2020;6(3):310–45.
[32] Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building machines that learn and think like people. Behav Brain Sci 2017;40:e253.
[33] Holzinger A, Saranti A, Molnar C, Biecek P, Samek W. Explainable AI methods—a brief overview. Berlin: Springer International Publishing; 2022.
[34] Xu L, Huang H, Liu J. SUTD-TrafficQA: a question answering benchmark and an efficient network for video reasoning over traffic events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; online. New York City: IEEE; 2021. p. 9878–88.
[35] Bakhtin A, van der Maaten L, Johnson J, Gustafson L, Girshick R. PHYRE: a new benchmark for physical reasoning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada. New York City: Curran Associates Inc.; 2019.
[36] Ahmed O, Träuble F, Goyal A, Neitz A, Wuthrich M, Bengio Y, et al. CausalWorld: a robotic manipulation benchmark for causal structure and transfer learning. In: Proceedings of the International Conference on Learning Representations; 2021 May 3–7; online. Vancouver: International Conference on Learning Representations; 2021.
[37] Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA. New York City: IEEE; 2015. p. 3128–37.
[38] Laskar MTR, Bari MS, Rahman M, Bhuiyan MAH, Joty S, Huang JX. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. 2023. arXiv:2305.18486.
[39] Dziri N, Lu X, Sclar M, Li XL, Jiang L, Lin BY, et al. Faith and fate: limits of transformers on compositionality. 2023. arXiv:2305.18654.
[40] Kosoy E, Reagan ER, Lai L, Gopnik A, Cobb DK. Comparing machines and children: using developmental psychology experiments to assess the strengths and weaknesses of laMDA responses. 2023. arXiv:2305.11243.
[41] Yao B, Yang X, Zhu SC. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In: Proceedings of the International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition; 2007 Aug 27–29; Ezhou, China. Berlin: Springer; 2007. p. 169–83.
[42] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA. New York City: IEEE; 2009. p. 248–55.
[43] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision; 2014 Sep 6–12; Zurich, Switzerland. Berlin: Springer; 2014. p. 740–55.
[44] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, et al. VQA: visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile. New York City: IEEE; 2015.
[45] Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: a multi-task benchmark and analysis platform for natural language understanding. 2018. arXiv:1804.07461.
[46] Hernández-Orallo J. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artif Intell Rev 2017;48(3):397–447.
[47] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, et al. OpenAI Gym. 2016. arXiv:1606.01540.
[48] Beattie C, Leibo JZ, Teplyashin D, Ward T, Wainwright M, Küttler H, et al. DeepMind Lab. 2016. arXiv:1612.03801.
[49] Li C, Xia F, Martín-Martín R, Lingelbach M, Srivastava S, Shen B, et al. IGibson 2.0: object-centric simulation for robot learning of everyday household tasks. 2021. arXiv:2108.03272.
[50] Gan C, Schwartz J, Alter S, Mrowca D, Schrimpf M, Traer J, et al. ThreeDWorld: a platform for interactive multi-modal physical simulation. 2020. arXiv:2007.04954.
[51] Kolve E, Mottaghi R, Han W, VanderBilt E, Weihs L, Herrasti A, et al. AI2-THOR: an interactive 3D environment for visual AI. 2017. arXiv:1712.05474.
[52] Savva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B, et al. Habitat: a platform for embodied AI research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. New York City: IEEE; 2019. p. 9339–47.
[53] Wu Y, Wu Y, Gkioxari G, Tian Y. Building generalizable agents with a realistic and rich 3D environment. 2018. arXiv:1801.02209.
[54] Puig X, Ra K, Boben M, Li J, Wang T, Fidler S, et al. VirtualHome: simulating household activities via programs. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. New York City: IEEE; 2018. p. 8494–502.
[55] Zador A, Escola S, Richards B, Ölveczky B, Bengio Y, Boahen K, et al. Catalyzing next-generation artificial intelligence through NeuroAI. Nat Commun 2023;14(1):1597.
[56] Avrin G. Assessing artificial intelligence capabilities. AI and the future of skills, volume 1: capabilities and assessments. Paris: OECD Publishing; 2021.
[57] Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, et al. Think you have solved question answering? Try arc, the AI2 reasoning challenge. 2018. arXiv:1803.05457.
[58] Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. 2022. arXiv:2206.04615.
[59] Li C, Zhang R, Wong J, Gokmen C, Srivastava S, Martín-Martín R, et al. Behavior-1k: a benchmark for embodied AI with 1000 everyday activities and realistic simulation. In: Proceedings of the Conference on Robot Learning; 2023 Nov 6–9; Atlanta, GA, USA; online. The Conference on Robot Learning (CoRL); 2023, p. 80–93.
[60] Xu B, Ren Q. Artificial open world for evaluating AGI: a conceptual design. In: Proceedings of the International Conference on Artificial General Intelligence; 2022 Aug 19–22; Seattle, WA, USA; online. The Artificial General Intelligence Society; 2023, p. 452–63.
[61] Terman LM, Merrill MA. Stanford-Binet intelligence scale: manual for the third revision, form L-M. Boston: Houghton Mifflin; 1960.
[62] Bayley N. Bayley-III: Bayley Scales of infant and toddler development. Florence: Giunti OS; 2009.
[63] Wechsler D. Wechsler Adult Intelligence Scale. Arch Clin Neuropsychol 1955.
[64] Raven JC, Court J. Raven's progressive matrices. Torrance: Western Psychological Services; 1938.
[65] Sternberg RJ. What should intelligence tests test? Implications of a triarchic theory of intelligence for intelligence testing. Educ Res 1984;13(1):5–15.
[66] Tu Z, Chen X, Yuille AL, Zhu SC. Image parsing: unifying segmentation, detection, and recognition. Int J Comput Vis 2005;63(2):113–40.
[67] Newton I, Colson J. The method of fluxions and infinite series: with its application to the geometry of curve-lines. London: Henry Woodfall; 1736.
[68] Spelke ES, Kinzler KD. Core knowledge. Dev Sci 2007;10(1):89–96.
[69] Duval S, Wicklund RA. A theory of objective self-awareness. Cambridge: Academic Press; 1972.
[70] Rochat P. Five levels of self-awareness as they unfold early in life. Conscious Cogn 2003;12(4):717–31.
[71] Wimmer H, Perner J. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 1983;13(1):103–28.
[72] Wellman HM, Liu D. Scaling of theory-of-mind tasks. Child Dev 2004;75(2):523–41.
[73] Warneken F, Tomasello M. Altruistic helping in human infants and young chimpanzees. Science 2006;311(5765):1301–3.
[74] Kanakogi Y, Inoue Y, Matsuda G, Butler D, Hiraki K, Myowa-Yamakoshi M. Preverbal infants affirm third-party interventions that protect victims from aggressors. Nat Hum Behav 2017;1(2):0037.

[75] Geraci A, Surian L. The developmental roots of fairness: infants' reactions to equal and unequal distributions of resources. Dev Sci 2011;14(5):1012–20.
[76] Porter HH III. A methodology for the assessment of AI consciousness. In: Proceedings of the International Conference on Artificial General Intelligence; 2016 Jul 16–19; New York City, NY, USA. Berlin: Springe; 2016. p. 305–13.
[77] Kotseruba I, Tsotsos JK. 40 years of cognitive architectures: core cognitive abilities and practical applications. Artif Intell Rev 2020;53(1):17–94.
[78] Xie X, Liu H, Zhang Z, Qiu Y, Gao F, Qi S, et al. VRGYM: a virtual testbed for physical and interactive AI. In: Proceedings of the ACM Turing Celebration Conference-China; 2019 May 17–19; Chengdu, China. New York City: ACM Turing Celebration Conference; 2019. p. 1–6.
[79] Gao X, Gong R, Shu T, Xie X, Wang S, Zhu SC. VRKitchen: an interactive 3D virtual environment for task-oriented learning. 2019. arXiv: 1903.05757.
[80] Ma X, Yong S, Zheng Z, Li Q, Liang Y, Zhu SC, et al. SQA3D: situated question answering in 3D scenes. In: Proceedings of the 11th International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda. New York City: IEEE; 2023.