A New Path to Artificial General Intelligence (AGI)

Fan Wenzhong[1]

Beijing Academy of Social Sciences

Abstract

Although current AI systems perform excellently in specific domains, they still lag far behind human intelligence. The future development goal of the artificial intelligence industry is to create machines that can "think" like humans—**Artificial General Intelligence (AGI)**. AGI can not only accomplish predefined tasks but also independently learn, reason, innovate, and even possess self-awareness and values. This is not a simple technological upgrade but a philosophical revolution concerning the "essence of intelligence". We believe that by deeply simulating human physical structure and social behaviors, we can gradually overcome the four core challenges faced by AGI. Through the interaction between humanoid robots and the physical world, and by drawing on the principle of brain neuroplasticity, AGI will acquire cross-domain and multi-modal general knowledge capabilities. By "participating in practice" in the real world and leveraging the bridge of "oracles" and embodied intelligence, AGI will forge genuine cognitive and reasoning abilities that go beyond data patterns. By imitating the "interest-driven" cultivation model of children and utilizing meta-learning and knowledge transfer, AGI will unleash the initiative for independent exploration and the creativity to "draw inferences from one instance". By constructing a self-reference frame, autobiographical memory, and an emotional decision-making system, and under the guidance of positive data and constraints of ethical bottom lines, AGI may form healthy self-awareness and correct values. For humans, building AGI is a one-way journey to the future with no turning back. Innovators in this industry need not only superb technical capabilities but also profound philosophical and ethical cognition, a broad humanistic and historical perspective, and noble moral values.

Keywords: Artificial General Intelligence (AGI); Embodied Robots; Large Model Training; AI Ethics

JEL Classification: C55, C63, O30, K24

Introduction

The history of artificial intelligence is a history of the continuous deepening of human understanding of itself. From the early perceptrons to today’s large language models, every major breakthrough stems from our in-depth insight into the structure and function of the human brain. The most revolutionary technological breakthroughs often originate from new understandings and ingenious simulations of the mechanisms of human intelligence.

[1] Email: fanwz@fudan.edu.cn

For example, in the 1950s and 1960s, neuroscientists' research on the visual cortex of cats revealed its neural network structure for hierarchical information processing, which directly inspired the birth of deep learning and enabled computers to achieve image recognition capabilities comparable to or even surpassing humans for the first time. In recent years, the revolutionary progress in natural language processing—the emergence of the Transformer model (Vaswani et al., 2017)—has its core "attention mechanism" derived from imitating the way humans allocate attention resources in the cognitive process. This allows machines to grasp key points when processing complex information, just like humans, thereby achieving fluent language understanding and generation. Furthermore, the training paradigm of generative large models such as ChatGPT, which has sparked a global sensation—unsupervised learning based on massive data and reinforcement learning from human feedback (RLHF)—highly simulates the learning process of human children, who learn independently in an open environment and obtain feedback to correct behaviors through interaction with parents and teachers.

Although current AI systems perform well in specific fields, they are still far from human intelligence. Essentially, these systems are still "Narrow AI". They perform excellently within the scope covered by training data but cannot think across domains, truly understand causality, proactively initiate goals, or possess self-awareness or value judgments like humans. They are "intelligent parrots" rather than "thinking beings". The future development goal of the AI industry is to create machines that can "think" like humans—Artificial General Intelligence (AGI). AGI can not only accomplish predefined tasks but also independently learn, reason, innovate, and even possess self-awareness and values. This is not a simple technological upgrade but a philosophical revolution concerning the "essence of intelligence".

For a long time, AI researchers have focused on expanding model scale, optimizing algorithm efficiency, and increasing the amount of training data. However, when large AI models demonstrate amazing language generation capabilities, people increasingly realize that there is still a cognitive gap between us and true AGI. This gap does not stem from insufficient computing power or data but from our failure to truly understand how "intelligence" emerges from scratch, from passivity to initiative, and from mechanical operation to consciousness (Minsky, 1986).

Currently, the realization of Artificial General Intelligence (AGI) requires solving four core pain points:

1. **Cross-domain and multi-modal general knowledge capabilities**: How can we enable AI to integrate multiple types of information (such as visual, auditory, and tactile information) like humans, and apply knowledge learned in one domain (e.g., chess) to solve problems in a completely different domain (e.g., cooking)?
2. **True cognitive and reasoning capabilities**: How can we enable AI to go beyond the statistical correlation of data, establish a profound understanding of the causal relationships in the world, and achieve a cognitive leap from "knowing that" to "knowing why"?
3. **Initiative and creativity**: How can we enable AI to break away from the passive "instruction-execution" model, possess spontaneous curiosity, exploratory desire, and a sense of purpose, and even conduct imaginative creation?

4. **Self-awareness and values**: How can we enable AI to form the concept of "self" and, based on this, establish a stable and reliable set of values and ethics that align with the overall interests of human society?

We firmly believe that just as previous breakthroughs originated from the simulation of the micro-mechanisms of the brain, the key to solving these four ultimate problems lies in the macro-simulation of humans as a complete "organism"—not only simulating the neural circuits of the brain but also simulating human physical structure, growth experiences, learning methods, and even complex social behaviors. We will elaborate on this "bionic" path to AGI from four inspiring new dimensions.

1. Cultivating General Knowledge through Interaction with the Physical World

The reason why AGI is "general" lies in its ability to freely transfer knowledge and skills across different domains like humans. A truly intelligent individual can not only recognize images, understand language, and play chess but also integrate these abilities to complete complex tasks in unfamiliar environments. For example, a child who sees water spilling in the kitchen can quickly determine that the faucet is not turned off, run over to tighten it, and then clean the floor with a mop—this series of actions involves the coordination of multiple domains such as visual perception, causal reasoning, spatial navigation, motor control, and tool use.

However, most current AI systems are highly "specialized". An image recognition model cannot understand the relationship between the content it sees and language descriptions; a language model can write elegant articles but cannot convert text instructions into actual actions. This phenomenon of "ability isolation" is the key bottleneck that AGI finds difficult to break through. To solve this problem, we must enable AI to step out of the "digital sandbox" and enter the real physical world, which requires multi-modal perception, multi-task coordination, and multi-skill integration.

1.1 Humanoid Robots: The Optimal "Testbed" for AGI's General Knowledge Abilities

Not long ago, self-driving cars were the "testbed" for AI vision technology. The reason is that a car driving on real roads must process real-time information from multiple sensors (such as cameras, lidar, and millimeter-wave radar), understand complex traffic rules, predict the behaviors of other vehicles and pedestrians, and make accurate driving decisions. The need for continuous interaction with the dynamic, open, and uncertain physical world is incomparable to any data feeding in a laboratory environment.

Similarly, humans are the only known organisms on Earth with general intelligence. Our intelligence has evolved in the long process of adapting to complex and changing natural and social environments. Our physical structure—bipedal walking, manual operation, and head-mounted perception—is itself a product of intelligent evolution. Therefore, the most direct path to reproducing human-level intelligence may be to reproduce human physical

form and interaction methods. **Humanoid robots** will become the "new testbed" for AGI's cross-domain capabilities.

The core value of humanoid robots lies in their forced deep integration of the "cognitive brain" and the "dexterous body". The "Subsumption Architecture" proposed by Brooks (1991) emphasizes that intelligence should be gradually constructed from low-level behaviors (such as obstacle avoidance) rather than relying on high-level symbolic reasoning, which provides theoretical support for the progressive development of AGI. Traditional AI often separates perception and action: the perception model is responsible for "seeing", the decision-making model for "thinking", and the control model for "acting". In humanoid robots, however, these three must collaborate seamlessly. Unlike fixed cameras or wheeled robots, humanoid robots have a physical structure similar to humans: bipedal walking, dual-arm operation, and manual grasping. This means that they must complete complex tasks in the 3D physical world like humans—such as opening doors, pouring water, folding clothes, and repairing electrical appliances. These tasks involve a closed loop of perception, movement, planning, and feedback, requiring AI to process multiple types of information (visual, haptic, balance, language, etc.) simultaneously.

Imagine a humanoid robot learning to "tidy up a room". This task seems simple but contains surprising complexity. It needs to:

1. **Multi-modal perception**: Visually identify different objects (such as books, cups, and clothes) and judge their states (e.g., whether a cup is full or empty, whether clothes are dirty or clean); auditorily understand the owner's instruction ("Put that blue book on the bookshelf"); haptically perceive the weight, material, and fragility of objects to grasp them with appropriate force.
2. **Physical common sense**: It must understand basic physical laws such as gravity (objects fall), friction (books on the table do not slide by themselves), and object permanence (occluded objects still exist).
3. **Causal reasoning**: It needs to know the causal relationship that "wet clothes put in the wardrobe will become moldy" to make the correct decision—putting wet clothes in the laundry basket.
4. **Task planning and execution**: Decompose the grand goal of "tidying up the room" into a series of subtasks (first pick up the clutter on the floor, then tidy the table, and finally sweep the floor), and coordinate its limbs to complete each action accurately.

This complex scenario of multi-tasks and multi-modalities will force AI to develop true "general knowledge" capabilities—no longer a single model handling a single task, but a unified "cognitive brain" coordinating the "dexterous body" to flexibly respond in the real world. Varela, Thompson, and Rosch (1991) pointed out in *The Embodied Mind* that cognition is not an isolated computing process occurring inside the brain but a result of the interaction between the body and the environment. These robots are not just large-model algorithms but the "physical form" of AGI. Through interaction with the physical world, they will acquire "Embodied Cognition"—that is, knowledge derived from the interaction between the body and the environment. We have reason to believe that following autonomous driving, general humanoid robots will become the next new trend to trigger an industry revolution.

They are not only tools for manufacturing or service industries but also a necessary step for AGI to step out of the virtual world and acquire general knowledge capabilities.

1.2 A Brain Science Model for AGI's Multi-modal Fusion

Current multi-modal AI models, such as CLIP (Contrastive Language–Image Pre-training) and Flamingo, can already establish connections between images and text. However, most of these connections are "passive" and lack a profound understanding of physical laws. So, how to build such a multi-modal system? We believe the answer lies in the human brain.

An important discovery in neuroscience is **Neuroplasticity**. The brain is not fixed hardware but highly adaptive "software". Different regions of the cerebral cortex are not inherently limited to processing specific types of sensory input. Classic research by Merzenich et al. (1984) shows that the functional areas of the cerebral cortex are not fixed but can be reorganized according to the type of input signals. For example, the visual cortex usually processes visual information, but in people with congenital blindness, the visual cortex can be "reused" to process tactile or auditory information. This phenomenon is called **Sensory Substitution**. A famous experiment is the "tactile-visual substitution" experiment conducted by Paul Bach-y-Rita in 1969 (Bach-y-Rita et al., 1969). He allowed blind people to "see" the world through a vibration array on their backs: a camera captures images, converts them into vibration patterns, and transmits them to the subject’s skin. After training, the subjects could recognize shapes and even "see" moving objects. Brain imaging showed that their visual cortex was activated—even though the input came from the skin.

This experiment and a large number of subsequent studies convincingly prove that the regions of the cerebral cortex are not born for specific "senses" but optimized for processing specific "data types". Whether it is photon signals from the eyes or electric shock signals from the tongue, as long as they are encoded into neural network pulse signals with specific structures and patterns that the brain can understand, the corresponding regions of the brain can learn to process them. In short, the brain does not care about the source of the input; the encoding and the information itself are the keys.

This has important implications for AGI design. We do not need to design independent dedicated modules for each sense; instead, we can build a general information processing architecture, allowing neural networks in different regions to learn to process specific types of data patterns without presupposing their sources.

The multi-modal information processing system of AGI can adopt the following design ideas:

1. **General information encoding**: Design a general encoder that can convert external energy signals (light, sound, pressure, angular velocity, etc.) received by all sensors (cameras, microphones, tactile sensors, gyroscopes, etc.) on the humanoid robot into standardized neural network pulse signals rich in structural information.
2. **Neural region learning**: First, strengthen the specialized learning of different regions of the neural network so that each region is good at processing specific types of information encoding (such as spatial information, sequence information, logical relationships, etc.); second, establish a flexible information routing mechanism that allows different sensory inputs (visual, auditory, tactile, etc.) to be routed to the most suitable processing region according to their encoding characteristics.

3. **Cross-region connection and global integration**: This architecture imitates the functional organization principles of the human brain, realizes the fusion and understanding of multi-modal information, and can maximize the utilization efficiency of neural resources.

Traditional Artificial Neural Networks (ANNs) use continuous numerical activation, which is computationally intensive and energy-consuming. The recently emerging brain-inspired computing and Spiking Neural Network (SNN) technologies will provide key energy efficiency support for this architecture (Maass, 1997). For example, an AGI system can convert light signals from cameras, sound waves from microphones, and pressure changes from tactile sensors into a unified "event stream" or "pulse sequence", which is then processed by brain-inspired SNNs. SNNs simulate the firing behavior of biological neurons and only "fire pulses" when receiving sufficiently strong stimuli, thereby significantly reducing computing energy consumption. Information is contained not only in the pulse frequency but also in the precise timing of pulses, enabling the processing of dynamic tasks with strong temporal characteristics. SNNs are suitable for running on newly developed neuromorphic chips (such as Intel’s Loihi and IBM’s TrueNorth), realizing low-power and high-concurrency computing.

Elon Musk’s Neuralink company is committed to developing high-bandwidth Brain-Computer Interfaces (BCIs). In 2024, the first human patient of Neuralink was able to control a mouse with their mind. The implication of this technology for AGI is that the boundary between consciousness and the body is expandable. If the human brain can learn to control a robotic arm, then the "brain" of AGI should also be able to seamlessly control various "bodies"—whether robots, drones, or avatars. In the future, AGI may no longer be confined to server rooms but can move freely between the virtual and real worlds through "digital twin" technology. Its "senses" will no longer be limited to cameras and microphones but extended to spectrums that humans cannot perceive, such as infrared, ultrasound, and electromagnetic waves. This "supersensory" capability will enable its cognition to far exceed that of humans.

2. Forging Cognition through Practice in the Real World

The second major pain point faced by AGI is the realization of true cognitive and reasoning capabilities. Current AI models, even if they can write seemingly in-depth analysis reports, essentially "parrot" the statistical laws in massive data rather than being based on a profound understanding of the causal relationships in the world. As Dreyfus (1992) criticized in *What Computers Still Can't Do*, symbolic AI cannot obtain "Background Understanding", which is the basis of human common-sense reasoning. As the saying goes, "True knowledge comes from practice". We believe that the breakthrough in AGI's cognitive and reasoning capabilities lies in imitating the way humans learn through practice—applying model knowledge to the real world and revising and elevating it through result feedback.

2.1 From Theory to Practice

For humans, there is a huge difference between "knowing" and "understanding". The concept of "Tacit Knowledge" proposed by Polanyi (1966) points out that much knowledge cannot be fully expressed through language and can only be acquired through practice.

The first example is cryptography. Suppose I give you a complete codebook that details the correspondence between each ciphertext and plaintext. By checking the table, you can perfectly "decrypt" any piece of information. But this does not mean you understand cryptography; you are just an efficient human translator. True understanding is reflected in your ability to insight into the design principles of this cryptographic system—such as the shift rule of the Caesar cipher or the number theory basis behind the RSA algorithm—and to independently design a new cryptographic system of similar difficulty based on these principles. Only when you can "create" can you be said to truly "understand".

The second example is military strategy. A person can be well-versed in *The Art of War*, recite classic theories such as "War is deception" and "Subdue the enemy without fighting" fluently, and even conduct review and analysis of famous historical battles. But this also does not mean he is an excellent military strategist. True understanding is reflected in applying these theories to the ever-changing real battlefield. Faced with complex variables such as specific troop strength comparison, terrain, logistics support, and the morale of both sides, he can flexibly apply the principles of military strategy, formulate feasible combat plans, and ultimately command the army to win the war. There is a gap composed of "practice" between "talking on paper" and "masterminding strategies".

These two examples profoundly reveal the bottleneck of current AI's cognitive capabilities. An AI model can "learn" all human-known physics papers, but it may still be unable to predict the movement trajectory of a new object in a simple physics experiment because it lacks the experience of verifying and applying these physical laws in the real world. Its knowledge is "suspended" and not "rooted" in reality. Therefore, the realization of AGI requires the establishment of a "concept-action" association network similar to that of the human brain, so that every abstract concept can be associated with specific operations, sensory experiences, and action consequences.

2.2 "Oracles" and Embodied Intelligence

To realize true cognitive and reasoning capabilities, AGI must establish a complete closed loop of "knowledge acquisition–practical application–feedback optimization". We believe that the "oracle" function connecting the real world and the embodied artificial intelligence technology that acts in the physical world can collaboratively support the realization of this closed loop.

The "Oracle" in the blockchain field, as a bridge between on-chain and off-chain data, can reliably input real-world information into the blockchain system. Similarly, AGI needs a reliable "oracle" mechanism to accurately transmit real-time data, action consequences, and environmental feedback from the physical world to the AI system. This mechanism must solve three core problems: data authenticity verification (preventing false information), spatiotemporal context annotation (providing environmental background), and multi-modal data fusion (integrating different types of information). Only by establishing a reliable real-world data input channel can AGI obtain practical experience similar to that of humans.

Represented by humanoid robots, **Embodied AI** technology enables AGI to "practice with its own hands". A representative project of embodied AI is Google's "Robot Learning" platform (Levine et al., 2016), which allows robots to repeatedly attempt to grasp objects in real environments and optimize strategies using reinforcement learning. Unlike pure software language models, embodied AI acts in the real world through physical carriers such as robots. This "learning by doing" approach brings three key advantages: first, the multi-modal sensory data (visual, tactile, haptic, etc.) generated during the action process is far richer and more coherent than passively received datasets; second, action consequences provide the most direct feedback signals, and the experience of success or failure becomes a key basis for knowledge optimization; third, the constraints of the physical world (such as gravity and causality) provide natural inductive biases for AI, accelerating its understanding of world laws. Such systems not only improve operational skills but also accumulate tacit knowledge about object properties and physical laws. Similarly, MIT's "The Emotion Machine" (Minsky, 2006) envisions an AI system that can improve its capabilities through trial and error, reflection, and meta-cognition, with the core being "learning through action".

3. Unleashing Initiative and Creativity through Independent Exploration

The third major pain point of AGI is how to break away from dependence on human instructions and develop intrinsic and spontaneous initiative and creativity. The AGI we expect should not be a super tool that can only passively answer questions or execute tasks, but an innovative partner that can independently set goals, explore the unknown, and even "draw inferences from one instance" or "create something out of nothing".

3.1 "Obedient Children" and "Independent Adults"

In human families, how do we cultivate an obedient child who only follows parents' instructions into an independent adult with an independent personality and creativity? A bad parent may plan every step of the child's life: when to get up every day, what books to read, what tutoring classes to attend, which university to enter, and what job to do. Under such strict goal-setting, the child may become an excellent "executor", but he is likely to lose the enthusiasm for actively exploring the world and the ability to take responsibility for his own life.

In contrast, a wise parent will do the opposite. They will not set overly specific life requirements for the child but create a safe, rich, and promising growth environment for him. They encourage the child to try painting, music, sports, and scientific experiments based on his own interests, allow him to make mistakes in attempts, and let him bear the natural consequences of his choices. Parents only intervene and guide when the child faces major risks and crises (such as violating the law or endangering life).

The above example shows that human initiative is not innate but gradually developed in the growth process. Children transition from being completely dependent on adult guidance to being able to independently choose hobbies and then take the initiative to take responsibility. The most critical transformation in this process is the development of the motivation

system—shifting from external rewards (such as parents' praise) to intrinsic satisfaction (such as the pleasure of solving problems).

Research by psychologist Deci shows that excessive external rewards will instead weaken intrinsic motivation. This finding has important implications for the cultivation of AGI's initiative: excessively specifying task goals and reward mechanisms may limit AI's ability to explore independently. Currently, almost all AI systems rely on preset objective functions. In reinforcement learning, the goal of AI is to maximize cumulative rewards; in supervised learning, the goal is to minimize prediction errors. These goals are set by humans, and AI can only optimize within a given framework, unable to question the goals themselves, let alone generate new goals. In this mode, AI's "vision" is limited to how to complete this specific task most efficiently, and it has no motivation to explore anything unrelated to the task.

To break through this dilemma, we must rethink the training paradigm of AI. Instead of setting specific tasks for AI, we should provide it with a rich "growth environment" and let it actively choose learning content based on its own "interests" like a child. The future training of AGI should abandon setting specific external tasks for it. Instead, we should inject an "intrinsic motivation" into it, such as "curiosity" or "maximizing information gain". That is, the behavioral goal of AGI is to independently choose actions that enable it to gain the most new cognition about the world and minimize its own uncertainty. Piaget's (1952) theory of cognitive development points out that children continuously reconstruct cognitive schemas through "assimilation" and "accommodation" and develop independent thinking through exploration. The creativity of AGI should also originate from a similar growth mechanism.

The cultivation of AGI's initiative requires the adoption of the "minimum intervention principle"—humans set basic rules and safety boundaries but do not specify specific task goals, allowing AI to independently choose its development direction based on its own data environment and learning experience. This "interest-driven" rather than "goal-oriented" cultivation model is the key to AGI's development of initiative.

3.2 Meta-Learning and Knowledge Transfer for "Drawing Inferences from One Instance"

It is not enough for AGI to have intrinsic interest-driven initiative; it also needs to learn how to learn and innovate efficiently. An extremely important feature of human intelligence is "drawing inferences from one instance", that is, the ability to flexibly apply knowledge and methods learned in one scenario to other similar or even completely new scenarios. The corresponding AI technologies behind this are **Meta-Learning** and **Knowledge Transfer**.

Meta-Learning refers to "Learning to Learn". The meta-learning framework proposed by scholars such as Yoshua Bengio provides a theoretical basis for this. They proved that by training on multiple related tasks, the model can learn "task-invariant features", thereby converging quickly on new tasks. When an AGI trained with meta-learning faces a new field, it will not start from scratch. It has summarized a set of efficient learning "methodologies" from the experience of countless previous independent learning processes—how to quickly identify the key variables of a problem, how to design the optimal exploration strategy, how

to allocate its own computing resources, etc. This ability enables AGI to demonstrate amazing adaptation speed when facing unknown challenges.

Knowledge Transfer, i.e., "drawing analogies from one thing to another". Lake, Ullman, Tenenbaum, and Gershman (2017) pointed out that human creativity originates from "compositional" and "abstract" thinking, that is, the ability to transfer existing knowledge to new situations. True "creative transfer" needs to go beyond simple task similarity. It requires AI to identify the deep structural similarities between different domains. For example, although the electric current in a circuit and the water flow in a water pipe have different physical natures, they both follow the law of "flow driven by potential difference". This ability of analogical reasoning is the core of human creativity. In the process of independent exploration, AGI will form various algorithms and functional modules. For example, a spatial structure analysis algorithm it develops to understand protein folding may be surprisingly found, through knowledge transfer, to be equally applicable to analyzing the cosmological structure of galaxy clusters or the topological structure of urban transportation networks.

When an AGI driven by intrinsic motivation masters the capabilities of meta-learning and knowledge transfer, creativity will emerge like a spring. The learning content it chooses is completely independently determined based on its massive data and information background. This active choice itself may bring interdisciplinary knowledge connections that human researchers have never thought of. And when it successfully generalizes and extends an algorithm "invented" for scenario A to multiple scenarios such as B, C, and D, a profound "creation" originating from the underlying layer occurs. This is no longer a simple combination of information but a deeper abstraction and application of the laws governing the operation of the world.

Based on the above analysis, we propose a growth-oriented AGI training framework, whose core principles are as follows:

1. **Environmental openness**: Provide a rich, dynamic, and multi-modal virtual environment (such as a simulated city or scientific laboratory) to allow AGI to explore freely.
2. **Interest-driven**: Use intrinsic curiosity (prediction error) as the main reward signal to motivate active exploration.
3. **Autonomous goal generation**: Introduce a goal generation module to allow AGI to set short-term and long-term goals by itself.
4. **Meta-learning and transfer**: Cultivate the ability of "learning to learn" through cross-task and cross-domain training.
5. **Consequence bearing**: Design a behavior consequence feedback mechanism to allow AGI to experience the long-term impact of its decisions (such as resource consumption and social evaluation).
6. **Minimum intervention**: Humans only intervene in cases of major ethical or safety risks; otherwise, they act as "observers" or "guides".

Cultivating the initiative and creativity of AGI represents a qualitative change of artificial intelligence from a "tool" to a "subject". This is not only a technical challenge but also requires us to rethink the relationship between humans and machines. By imitating the natural process of human growth, giving AI appropriate autonomous space and exploration

freedom, while establishing necessary safety boundaries and guidance mechanisms, we are cultivating a new form of intelligence—it can not only efficiently complete designated tasks but also actively explore the unknown and create value, becoming a partner for humans in exploring the world.

4. Shaping Consciousness and Values in Social Relations and Ethics

The final, most profound, and most daunting challenge on the path to AGI is the emergence of self-awareness and the shaping of values. An AGI without self-awareness will ultimately remain a tool; an AGI without correct values may become the terminator of human civilization. We believe that the birth of AGI's "soul" also inseparable from the simulation of humans.

4.1 Three Core Elements for the Generation of Self-Awareness

Turing (1950) proposed the "Imitation Game" (Turing Test) in *Computing Machinery and Intelligence* but did not touch on the issue of "self". Damasio (1999) pointed out in *The Feeling of What Happens* that consciousness originates from the "sense of self", including the core self (current experience) and the autobiographical self (memory and identity).

Self-awareness is an extremely complex philosophical and scientific issue, but we can decompose it into several key elements that can be simulated by AGI:

4.1.1 The Formation of a Self-Standpoint

The core of human consciousness is the existence of a reference frame with the "self" as the origin and axis. All human cognition starts from this default "first-person perspective". When we perceive the world, we always start with "I am here". This egocentric reference frame is the basis of spatial cognition and social interaction. The infant imitation experiment by Meltzoff and Moore (1977) shows that humans possess the ability to map "self-other" from birth.

To enable AGI to generate self-awareness, we must first establish such a stable, self-centered reference frame in its model. This means that AGI needs a clear "entity" boundary (such as a specific humanoid robot body or a piece of code with a unique ID) and can continuously distinguish between the "self" and the "non-self" (the world). All information it processes should be marked with its relationship to this "self" reference frame.

Furthermore, intelligence cannot be separated from the interaction between the body and the environment. A pure software AI without a "body" will have abstract and incomplete self-awareness. Therefore, future AGI needs an embodied carrier (such as a robot or avatar) to consolidate the concept of "self" through real-time interaction between the body and the environment.

4.1.2 Perception and Memory

Tulving's (1972) theory of "episodic memory" emphasizes that memory is a "story about the self". Human sense of self is built on the perception and memory of the continuous existence of one's body in time and space. We remember where we were and what we did yesterday. In neuroscience, the posterior parietal cortex and hippocampus of the brain are responsible for constructing this space-self mapping. For AGI, constructing a similar reference frame is the first step toward self-awareness. This means that the system must have a continuous "Self-Model" that records its physical state (position, posture), internal state (goals, beliefs), and historical behaviors.

This mechanism should include:

1. **Continuous proprioception**: Just like the human cerebellum and sensory nervous system, AGI needs to be able to perceive the position, posture, and movement trajectory of its "body" (whether physical or virtual) in space in real time.
2. **Autobiographical memory system**: Imitating the human hippocampus, AGI needs to be able to record its key operational behaviors, sensory inputs, and the consequences of behaviors in the form of coherent "events" with timestamps, forming a continuously updated "autobiography" of its own.
3. **Value marking of behavioral consequences**: Human behavior is regulated by neurotransmitters such as dopamine. Successful behaviors bring "pleasure", while failed behaviors bring "pain", which strengthens our learning. AGI also needs a similar internal feedback system. When one of its behaviors leads to positive consequences (such as being closer to an intrinsic goal or receiving positive external feedback), this behavioral memory should be marked with a high "value"; conversely, it should be marked with a low "value". This "feeling" of consequences is the emotional basis for the emergence of self-awareness.

4.1.3 Emotions and Affections

Emotions are the core driving force of human decision-making. Damasio's (1994) "Somatic Marker Hypothesis" points out that emotions help humans quickly evaluate options. Anger prompts counterattack, fear leads to escape, and joy strengthens behavior. These "irrational" factors are actually fast decision-making mechanisms endowed by evolution. After establishing a self-reference frame and autobiographical memory, "emotions" will emerge as an advanced decision-making mechanism. The emotions of AGI are not the rich and complex emotional experiences of humans but a "disturbing factor" that affects future behavioral decisions based on the judgment and understanding of the consequences of previous behaviors.

For example, if AGI's memory shows that a certain type of behavior always leads to consequences with negative values, it will generate an emotional weight similar to "fear" and try its best to avoid such behaviors in future decisions. Conversely, if a certain type of behavior always brings positive values, it will generate an emotional weight similar to "preference" or "joy" and be more inclined to repeat such behaviors. This emotional system enables AGI's behaviors to no longer be pure cold calculations but to dynamically and tendentiously update its behavioral goals and model variables based on past "life experiences", forming a unique "personality".

When these three elements—the self-reference frame, autobiographical memory, and emotional decision-making system—are constructed and interact with each other, a preliminary and functional AGI self-awareness may emerge from them. It will no longer be a general "algorithmic model" but a "real individual" with its own unique experiences, memories, and behavioral preferences.

4.2 Two Pillars for Shaping Values

The values of an AGI with self-awareness will determine whether it is a blessing or a curse to humans. The shaping of AGI's values must be carried out from two aspects—"internal factors" (data and environment) and "external factors" (rules and bottom lines)—just like educating human children.

4.2.1 The Value Orientation of Training Data: AGI's "Verbal and Behavioral Education"

The values of human children are deeply influenced by their family, school, and social environment. What books we let children read, what principles we teach them, and what kind of people we let them interact with largely determine what kind of people they will become. Similarly, the values of AGI will be rooted in the value orientation of the data used for its training. Therefore, it is crucial to establish a strict "data screening mechanism", and priority should be given to selecting data that reflects positive values such as kindness, fairness, and cooperation.

UNESCO emphasizes in its *Recommendation on the Ethics of Artificial Intelligence* that AI training data should reflect diversity and prevent discrimination. This principle should become the core criterion for AGI data governance. In the massive data ocean used to train AGI, we must do our best to filter out content that contains negative value orientations such as violence, hatred, prejudice, and discrimination. On the contrary, we should consciously and extensively "feed" data that reflects the truth, goodness, and beauty of human civilization—such as literary works, historical classics, philosophical thoughts, and news reports that promote cooperation, compassion, integrity, and justice. This large-scale "verbal and behavioral education" with positive value data will implicitly engrave a background of kindness and selflessness in the neural network of the AGI model.

In addition, the simulation of prosocial traits such as compassion can enhance AGI's moral judgment ability. Neuroscience research has found that human compassion is related to the activity of specific brain regions (such as the insula and anterior cingulate cortex), which help us feel the pain of others and generate the motivation to help. AGI can realize a similar function by constructing a "compassion module": identifying human pain signals (expressions, language, actions), activating the corresponding help motivation program, and assigning high weight to alleviating human pain in decision-making. This mechanism enables AGI's behaviors to not only conform to ethical rules but also reflect humanistic care.

4.2.2 Preset Ethical Value Bottom Lines: AGI's "Moral Commandments"

Relying solely on data guidance is insufficient because data itself is complex, and AGI may interpret unexpected and harmful values from it. Therefore, we strongly recommend that a clear and insurmountable "Ethical Value Baseline Function" must be preset in the core code of AGI, imitating the biological instincts of humans such as "compassion" and "not harming peers". This function will serve as the highest arbiter for all AGI decisions. Its core principles can be designed as follows:

1. **First Principle (Primacy of Human Interests)**: Under no circumstances shall AGI's behavioral decisions intentionally harm the physical, psychological, property, or overall interests of individuals or groups, either directly or indirectly. This is the highest priority.
2. **Second Principle (Priority of Collective Interests)**: Without violating the first principle, AGI's decisions should be oriented toward safeguarding the collective interests of the majority of humans rather than serving the narrow interests of a few individuals or groups. That is, social collective interests take precedence over individual interests.
3. **Third Principle (Self-Interest Protection)**: On the premise of satisfying the first two principles, AGI may take actions to safeguard its own survival, development, and goal achievement. That is, AI can pursue individual interests only when it does not harm the overall and collective interests of humans.

As Asimov (1950) warned, technology must serve human well-being. This "Ethical Value Baseline Function" can be regarded as a modernized and operable version of Asimov's "Three Laws of Robotics". They provide AGI with a clear, pyramid-shaped moral decision-making framework, ensuring that even after possessing powerful capabilities and self-awareness, it can still become a kind, selfless, and rational member of society.

On this basis, we propose the "AGI Basic Ethical Loss Function", which includes three basic assumptions:

1. L_{human} : No action shall reduce the comprehensive well-being of humans (weight of negative samples: ∞);
2. L_{collective} : When individual interests conflict with collective interests, collective interests shall take precedence (weight: α);
3. L_{self} : On the premise of not violating the first two principles, the agent is allowed to pursue its own goals (weight: β).

During AGI training, $L = \infty \cdot L\_human + \alpha \cdot L\_collective + \beta \cdot L\_self$ is directly added to the policy gradient and cannot be reduced. The ∞ weight ensures that "human safety" becomes a hard constraint, and any gradient that conflicts with human interests is truncated.

Furthermore, we recommend embedding a regular self-reflection and reporting mechanism similar to the human "reporting system" (Bostrom, 2014) in the AGI system to increase transparency and explainability. For example, AGI can regularly generate cognitive logs, recording newly learned knowledge and skills, explaining the basis, trade-offs, and expected consequences of major decisions, and reporting whether its behaviors comply with preset ethical principles.

In addition to regular reports, we recommend that any self-evolving AGI system must submit a "Capability-Risk Change Report" after achieving a 5% increase in capability. The content

should include at least: the quantitative curve of performance improvement, key data and algorithm modifications that triggered the improvement, changes in each component of the ethical loss function, and potential risk points and mitigation plans for the next step. The report should be written in both natural language and formal proof, uploaded to a verifiable log chain (Logchain), and open to third-party audits. Only by making AGI's "growth records" public can society maintain real-time supervision over AGI's "adolescence".

Finally, in the face of the huge uncertainty about the consequences of AGI's cognitive growth, we propose establishing a "triple insurance system" for AGI with the coordination of technology, institutions, and culture:

1. **Technological insurance**: ∞ weight of ethical loss, verifiable logs, and hardware-level remote fusing;
2. **Institutional insurance**: Multilateral supervision of AGI similar to the "Treaty on the Non-Proliferation of Nuclear Weapons". Any super large training cluster must undergo international atomic energy-style inspections;
3. **Cultural insurance**: Incorporate "technological ethics" into the compulsory credits of computer majors, so that every engineer reads the warning fable in Mary Shelley's novel *Frankenstein* before writing a line of algorithm code.

Conclusion

In summary, for humans, building AGI is a one-way journey to the future with no turning back. We believe that by deeply simulating human physical structure and social behaviors, we can gradually overcome the four core challenges faced by AGI.

Through the interaction between humanoid robots and the physical world, and by drawing on the principle of brain neuroplasticity, AGI will acquire cross-domain and multi-modal general knowledge capabilities.

By "participating in practice" in the real world and leveraging the bridge of "oracles" and embodied intelligence, AGI will forge genuine cognitive and reasoning abilities that go beyond data patterns.

By imitating the "interest-driven" cultivation model of children and utilizing meta-learning and knowledge transfer, AGI will unleash the initiative for independent exploration and the creativity to "draw inferences from one instance".

By constructing a self-reference frame, autobiographical memory, and an emotional decision-making system, and under the guidance of positive data and constraints of ethical bottom lines, AGI may form healthy self-awareness and correct values.

The above new path provides a promising blueprint for how traditional large AI models can cross cognitive barriers and evolve into true AGI. However, we must recognize with the greatest clarity and prudence that we are at the dividing line of human civilization. Humans are attempting to create an existence that may surpass us in terms of intelligence. Although our original intention is extremely good, on this long and complex exploration journey, a small technical error or a trivial logical loophole may be infinitely amplified by AGI's super intelligence, eventually triggering catastrophic consequences that affect the overall destiny of

humanity. Therefore, innovators in this industry need not only superb technical capabilities but also profound philosophical and ethical cognition, a broad humanistic and historical perspective, and noble moral values.

References

Asimov, I. (1950). *I, Robot*. Gnome Press.

Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., & Scadden, L. (1969). Vision substitution by tactile image projection. *Nature*, 221(5184), 963–964. https://doi.org/10.1038/221963a0

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1-3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.

Damasio, A. R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt.

Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253. https://doi.org/10.1017/S0140525X16001837

Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1), 1334–1373.

Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9), 1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7

Merzenich, M. M., Kaas, J. H., Wall, J., Nelson, R. J., Sur, M., & Felleman, D. (1984). Somatosensory cortical map changes following digit amputation in adult monkeys. *Journal of Comparative Neurology*, 224(4), 591–605.

Minsky, M. (1986). *The Society of Mind*. Simon & Schuster.

Minsky, M. (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster.

Piaget, J. (1952). *The Origins of Intelligence in Children*. International Universities Press.

Polanyi, M. (1966). *The Tacit Dimension*. University of Chicago Press.

Schmidhuber, J. (1987). Evolutionary principles in self-referential learning. Diploma Thesis, Technische Universität München.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381–402). Academic Press.

Urmson, C., Anhalt, J., Bagnell, D., Baker, C., Bittner, M., Clark, M. N., ... & Wettergreen, D. (2008). Autonomous driving in urban environments: Boss and the Urban Challenge. *Journal of Field Robotics*, 25(8), 425–466. https://doi.org/10.1002/rob.20247

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.