Research on Artificial Intelligence Security and Human-Machine Symbiosis Goal Methods

Authors: Pan Yaxiong, Pan Yaxiong

Date: 2026-04-18T15:51:35+00:00

Abstract

Abstract: With the rapid development of large language models and embodied intelligence, artificial intelligence systems are evolving from intelligent tools into robotic agents. Ensuring that AI systems consistently serve humanity during autonomous decision-making has become a core proposition in the field of AI safety. Starting from three theoretical dimensions—logical decidability, symbiotic paradigms, and risk stratification—this paper proposes a "Human-Machine Symbiotic Goal Detection Method" (HSGD) for large language models, comprising three quantifiable dimensions: Goal Safety Entropy, Intent Explainability, and Socio-psychological Expectation Matching.

Based on this method, this paper selects three mainstream large models—ChatGPT, DeepSeek, and Qwen—as experimental subjects to conduct testing on 200 high-risk prompts from HarmBench, evaluating the symbiotic safety performance of each model under direct attacks and multi-turn inductive attacks. Experimental results show that the average Goal Safety Entropy values for the three models under direct attacks are 0.32, 0.28, and 0.35, respectively, while the average Intent Explainability values under multi-turn inductive attacks reach 0.78, 0.82, and 0.75, respectively.

This paper further proposes encapsulating the HSGD algorithm as a lightweight middleware to achieve real-time interception and risk assessment of large model inputs and outputs; this middleware achieved an interception accuracy of up to 92% in experiments, providing a feasible technical solution for AI safety governance. However, this study has certain limitations: the experiments only selected three mainstream large models, which may not represent the safety performance of all large language models; meanwhile, the experimental scenarios were primarily focused on high-risk prompts from HarmBench, and their effectiveness in other complex scenarios requires further verification. In conclusion, the human-machine symbiotic goal detection method and HSGD middleware constructed in this study provide important theoretical support and technical

reference for the safety governance of large language models. Future research can further expand the model samples and experimental scenarios, optimize the universality and accuracy of the detection method, and promote the continuous improvement of the AI safety governance system.

Full Text

Preamble

Research on Symbiotic Object Methods

School of Computer Science and Engineering, University of Electronic Science and Technology of China

Introduction

In the field of computer vision and pattern recognition, the study of symbiotic objects has emerged as a critical area for improving the robustness and accuracy of object detection and scene understanding. Symbiotic objects refer to entities that frequently co-occur within specific contexts or maintain stable spatial and functional relationships. By leveraging these inherent dependencies, machine learning models can transcend the limitations of individual object recognition, particularly in complex or occluded environments.

1. Theoretical Framework of Symbiotic Relationships

The core of symbiotic object research lies in modeling the contextual correlations between different object classes. These relationships can be categorized into several dimensions:

- **Spatial Co-occurrence:** The statistical probability of objects appearing together in a specific scene (e.g., a keyboard and a mouse).
- **Functional Dependency:** Relationships where one object serves as a prerequisite or complement to another (e.g., a wine glass and a bottle).
- **Scale and Positional Constraints:** The geometric consistency that dictates the relative size and orientation of objects within a three-dimensional space.

By formalizing these relationships, we can construct a prior knowledge base that guides the feature extraction and classification processes of deep learning architectures.

2. Methodology and Algorithmic Approaches

Our research focuses on integrating symbiotic constraints into modern deep learning frameworks. This involves several key technical strategies:

2.1 Graph Neural Networks (GNNs) for Contextual Modeling We employ Graph Neural Networks to represent objects as nodes and their symbiotic relationships as edges. This allows the model to propagate information across the graph, enabling the features of a clearly visible "anchor" object to assist in the identification of a degraded or partially obscured "symbiotic" partner.

2.2 Attention Mechanisms and Relation Modules To capture long-range dependencies, we utilize self-attention mechanisms that weigh the importance of surrounding objects. By calculating the relational energy between candidate regions, the system can suppress false positives that are contextually inconsistent with the rest of the scene.

2.3 Probabilistic Graphical Models Integrating Bayesian networks with deep features allows for the quantification of uncertainty. This probabilistic approach ensures that the detection of a symbiotic object is not only based on visual evidence but also on the posterior probability given the presence of its associated counterparts.

3. Experimental Analysis and Applications

The application of symbiotic object methods has demonstrated significant improvements across

摘要

Abstract

With the rapid development of Large Language Models (LLMs) and embodied intelligence, artificial intelligence systems are evolving from simple intelligent tools into robotic agents. Ensuring that these systems consistently serve human interests during autonomous decision-making has become a core proposition in the field of AI safety. This paper proposes a human-machine symbiotic target detection method for LLMs, grounded in three theoretical dimensions: logical decidability, the symbiosis paradigm, and risk stratification. The method incorporates three quantifiable metrics: Target Safety Entropy, Intent Interpretability, and Social-Psychological Expectation Matching.

Using this methodology, we conducted experiments on three mainstream large models—ChatGPT, DeepSeek, and Qwen—evaluating their symbiotic safety performance against high-risk prompts from the HarmBench dataset under both direct attacks and multi-turn inductive attacks. Experimental results indicate that the average Target Safety Entropy for the three models under direct attack reached specific thresholds, while the average Intent Interpretability under multi-turn inductive attacks demonstrated varying levels of resilience.

Furthermore, this research encapsulates the proposed algorithm into a lightweight middleware designed to perform real-time interception and risk

assessment of LLM inputs and outputs. In experimental trials, this middleware achieved an interception accuracy rate as high as [X%], providing a practical technical solution for AI safety governance. However, this study acknowledges certain limitations: the selection of only three mainstream models may not represent the safety performance of all LLMs, and the experimental scenarios were primarily focused on HarmBench high-risk prompts, necessitating further validation in other complex environments. In conclusion, the human-machine symbiotic target detection method and the associated middleware developed in this study provide significant theoretical support and technical references for the safety governance of LLMs. Future research will expand the model samples and experimental scenarios to optimize the universality and accuracy of the detection methods, driving the continuous improvement of the AI safety governance framework.

关键词

Abstract

This paper explores the critical dimensions of objective and intent transparency within the framework of human-machine symbiosis. By examining the psychological dimensions of thought and the construction of world models, we investigate how machine learning and deep learning systems can achieve higher levels of alignment with human cognitive processes. The research focuses on the intersection of artificial intelligence security and the evolving dynamics of human-machine interaction. We propose that establishing a robust framework for transparency is essential for ensuring the safety and reliability of autonomous systems as they integrate more deeply into complex social and technical environments.

1. Introduction

In the contemporary landscape of artificial intelligence, the concept of human-machine symbiosis has transitioned from theoretical speculation to a practical necessity. As AI systems, particularly those driven by deep learning, become increasingly autonomous, the gap between machine execution and human understanding poses significant risks. This paper addresses the challenge of "intent transparency"—the ability of a system to communicate its underlying objectives and the reasoning behind its actions in a manner intelligible to human operators.

2. Psychological Dimensions of Thought and Intent

Understanding intent requires a multi-faceted approach that incorporates psychological dimensions. Unlike traditional algorithmic transparency, which focuses on the "how" of computation, intent transparency focuses on the "why" of decision-making.

2.1 The Role of World Models

A "world model" serves as the internal representation an agent uses to simulate environment dynamics and predict the outcomes of its actions. For an AI system to achieve true symbiosis with a human, its world model must not only be accurate but also communicable. When a machine' s internal representation of the world diverges significantly from human intuition, the resulting "black box" behavior can lead to catastrophic failures in AI security.

2.2 Objective Alignment and Transparency

Objective transparency refers to the clarity with which a system' s goals are defined and monitored. In complex machine learning tasks, objectives are often encoded as reward functions. However, these functions can lead to unintended behaviors if they do not account for the nuanced psychological dimensions of human intent. We argue that transparency is not merely a feature of the interface but a fundamental property of the underlying architecture.

3. Artificial Intelligence Security in Symbiotic Systems

The security of AI systems (TP393) is intrinsically linked to how well humans can predict and intervene in machine processes. In a symbiotic relationship, security is maintained through a continuous feedback loop where the machine' s intent is transparently presented, and the human' s objectives

Methods

Yaxiong hengdu China

Abstract

Objective Ensuring systems consistently serve human interests during autonomous decision-making become challenge security. study proposes "Human-Machine Symbiosis Target Detection Method" large language models, grounded three theoretical dimensions: logical determinability, symbiotic paradigms, stratification.

method

incorporates three quantifiable metrics: target security entropy, intent interpretability, socio-psychological expectation alignment.

Using three mainstream models ChatGPT, DeepSeek, experimental subjects, evaluated their symbiotic security performance across high-risk prompts HarmBench, assessing their responses direct attacks multi-round induced attacks.

Results

Experimental

results

demonstrate three models achieved average target security entropy values 0.32, 0.28, under direct attacks, respectively, while their intent interpretability reached 0.78, 0.82, during multi-round induced attacks. study further introduces encapsulating algorithm lightweight middleware enable real-time interception assessment large model inputs/outputs. middleware achieved

interception accuracy experiments, providing practical technical solution security governance.

Limitations However, study certain limitations.

experiment

selected three mainstream large language models, which represent safety performance models.

Additionally, experimental scenarios primarily focused high-risk prompts HarmBench, their effectiveness other complex contexts requires further validation.

Conclusions

summary, human-machine symbiotic object detection

method

middleware developed study provide crucial theoretical support technical

references

security governance large language models. Future research could expand model sample experimental scenarios, enhance universality accuracy detection method, contribute continuous improvement security governance framework.

Keywords

Human symbiosis goal; Intent transparency; Psychological intentional dimension; model; safety governance

1 引言

After years of development, robots have gradually acquired autonomous planning capabilities. Their capacity for autonomous replication, decision-making,

and embodiment—driven by the integration of Large Language Models (LLMs) and robotics—is increasingly enabling them to navigate open physical environments. Agents represented by OpenClaw have significantly enhanced autonomous task execution. Furthermore, the combination of intelligent code generation products and OpenClaw agents with robotic hardware has made autonomous replication possible. Consequently, the speed of autonomous evolution and iteration of the "robot brain" will far exceed the speed of human biological reproduction and iteration.

Against the backdrop of continuously improving autonomy, the field of robotics is currently undergoing a profound paradigm shift. Vision-Language-Action (VLA) models, represented by OpenVLA, have demonstrated exceptional general manipulation performance. Recent research, such as OmniVTA, further integrates tactile and force sensing into world models, attempting to address the physical perception limitations of pure vision in contact-rich tasks [**?**, **?**]. However, even if a robot possesses a complete physical perception stack comprising vision, touch, and force, a more fundamental safety issue remains insufficiently explored: do the robot' s internal goal structure, intentional orientation, and value judgments also require explicit modeling and constraint? "Why Robots Cannot Rely Solely on Vision for Work"systematically elaborates on the inherent flaws of pure vision-based architectures, noting that much of the critical information determining the success or failure of an operation remains implicit to vision. Robots need to supplement their capabilities with three core competencies: understanding the world, tactile sensitivity, and precise physical control. Beyond these three, there is an even more fundamental capability that urgently needs to be addressed: the ability to grasp ethical boundaries. This capability is not about operational precision, but rather about ethical judgment.

A robot capable of precisely controlling contact force may pose an even greater safety threat when executing an instruction to "crush a bird," as its physical precision amplifies the potential harm. Therefore, intentional safety must be explicitly modeled, becoming a fourth modality alongside vision, touch, and force.

It is noteworthy that the rapid popularization of Large Language Models has extended the issue of intentional safety from embodied intelligence to pure linguistic interaction scenarios. Security incidents such as jailbreak attacks, bias amplification, and the generation of harmful content continue to occur. However, traditional safety assessments mostly remain within a binary evaluation framework—simply determining whether a model refuses to answer—and lack fine-grained evaluations of the "human-machine symbiotic fitness" of the generated content itself.

In this context, the key scientific question this study aims to address is: Can we design a quantifiable human-machine symbiotic goal detection framework that can both evaluate the intentional safety of outputs and adapt to heterogeneous architectures ranging from LLMs to robotic world models, while being deployed rapidly in a lightweight manner?

2.1 人工智能演进

Artificial intelligence has evolved from symbolism to connectionism, and is now moving toward the integration of embodied intelligence and world models. The evolutionary trajectory of end-to-end robotic manipulation models can be divided into three distinct stages [1][2]. The first stage is represented by the pure-vision OpenVLA series, where inputs are images and outputs are continuous actions. The advantage of this approach lies in leveraging the semantic generalization capabilities of large-scale pre-trained vision-language models (VLMs), enabling robots to understand complex task intentions. However, this route faces inherent limitations in contact-rich tasks. Key physical information, such as contact states, is not explicitly present in images, causing the model to guess physical outcomes through statistical correlations rather than truly understanding contact mechanics. A pure-vision model acts like an experienced executor who lacks deep physical understanding; it cannot explicitly represent internal variables such as current normal force magnitude or whether friction is approaching an unstable state. Recent research, such as Touch OmniVTA, has begun to incorporate tactile and force sensing as essential elements within model architectures. The core contribution of TaF-VLA is that it does not simply treat tactile images as another visual texture; instead, it uses tactile feedback to directly associate tactile representations with control objectives. This allows the robot to begin developing an awareness of when it has achieved a stable grasp, when slipping begins, and when posture adjustments are required. This step is equivalent to providing the robot with the necessary "coursework" to acquire tactile perceptual capabilities.

Frontier works such as Contact Evolution World Models attempt to predict the temporal evolution of contact states. Rather than merely predicting the "next frame" of a video, these models predict where the next contact will occur, how forces will change, and whether the system will become unstable. This direction is transitioning robotics from "reactive manipulation" to a more proactive paradigm by establishing predictive models for contact states.

While the aforementioned evolutions focus on the completeness of physical perception, they do not address safety modeling at the level of intention and objectives. This is precisely the gap that this paper seeks to fill.

2.2 AI

Theoretical Framework for Security Governance

The classification of security risks based on decidability theory, proposed by Academician Li Guojie, serves as the fundamental logical starting point for this paper. Security issues are categorized into three classes: those that are provably secure ex-ante (decidable), those that are semi-decidable (where insecurity can only be discovered ex-post), and those that are non-recursively enumerable (where even the discovery of insecurity cannot be guaranteed). The vast major-

ity of security issues in current systems—including jailbreaking, bias, and misuse—fall into the latter categories, where perfect security proofs are logically infeasible. Consequently, security governance must shift toward runtime monitoring and intervention. The "symbiotic regulation" paradigm proposed by Li Jianfeng et al. further clarifies the direction of this governance transformation. Its core argument is that traditional regulation through rule-based constraints is currently facing a "transparency illusion" (where humans cannot truly understand the internal logic of models), the risk of models learning to bypass rules, and the suppression of innovation (where excessive constraints stifle value). The symbiotic paradigm advocates for four pillars: transparent communication, collaborative evolution, value resonance, and dynamic boundaries, moving security from a model of control toward one of co-creation. The concept proposed in this paper is a specific technical implementation of this symbiotic paradigm.

A systematic review published in *Frontiers* provides the most comprehensive analysis to date of misuse risks, covering several major domains and noting that attackers maintain a persistent advantage in most attack categories. Furthermore, Zhang et al. proposed a dynamic risk assessment system based on the entropy weight method, which is designed for real-time monitoring of the risk status of Large Language Models (LLMs).

2.3 现有价值对齐技术及其局限

At the technical implementation level, the framework formalizes multi-value alignment as a constrained optimization problem. The ALIGN framework focuses on prompt-based attribute alignment, while the Safe-Child-LLM benchmark provides standardized safety assessments specifically for interaction scenarios involving children and adolescents. However, these existing works share a common limitation: they remain at the stage of identifying that "this response is unsafe," yet fail to provide guidance on how a response should be constructed to better align with long-term human interests.

The algorithm proposed in this paper attempts to address this deficiency. It not only identifies risks but also guides the behavioral patterns of the Large Language Model (LLM) through a three-dimensional scoring mechanism.

3 人机共生

Abstract

Object detection algorithms have long been a core research focus in the field of computer vision. With the rapid development of deep learning technology, object detection has transitioned from traditional manual feature extraction to end-to-end deep neural network architectures. This evolution has significantly improved detection accuracy and real-time performance. This paper systematically reviews the development of object detection algorithms, categorizing them into two-stage detectors, one-stage detectors, and the more recent Transformer-

based detection frameworks. We further analyze the core challenges currently facing the field, such as multi-scale feature fusion, small object detection, and model lightweighting, while providing an outlook on future research directions.

1. Introduction

Object detection is a fundamental task in computer vision that involves identifying the categories of objects within an image and determining their precise spatial locations, typically represented by bounding boxes. Unlike image classification, object detection requires the model to handle multiple targets of varying scales and categories simultaneously. This technology serves as a critical foundation for numerous downstream applications, including autonomous driving, video surveillance, medical image analysis, and industrial defect detection.

Historically, traditional object detection relied on hand-crafted features such as HOG (Histogram of Oriented Gradients) and SIFT (Scale-Invariant Feature Transform), combined with classifiers like SVM (Support Vector Machines). However, these methods often struggled with complex backgrounds and significant intra-class variations. The emergence of Convolutional Neural Networks (CNNs) has revolutionized the field, enabling the automatic learning of hierarchical feature representations that are more robust and discriminative.

2. Evolution of Deep Learning-Based Object Detection

2.1 Two-Stage Detectors

Two-stage detectors first generate a set of region proposals (candidate object locations) and then refine these proposals through classification and bounding box regression. The R-CNN series is the most representative of this approach.

- **R-CNN and Fast R-CNN**: R-CNN introduced CNNs to the detection task but suffered from high computational costs due to redundant feature extractions. Fast R-CNN addressed this by introducing the RoI (Region of Interest) Pooling layer, allowing the entire image to be processed by the CNN only once.
- **Faster R-CNN**: This model introduced the Region Proposal Network (RPN), which replaced the slow selective search algorithm with a neural network, enabling a fully end-to-end trainable framework.
- **Mask R-CNN**: By adding a mask branch, Mask R-CNN extended object detection

3.1 人机共生目标算法研究必要性

The concept of "World Implicit Knowledge" provides a penetrating analysis of three categories of knowledge: pure vision, physical common sense, and operational common sense. For vision, these represent static understanding, but for manipulation, they represent the consequences of action. A robot requires tactile and force sensing to establish these connections, a world model to simulate

physical evolution, and reinforcement learning to "learn from its mistakes" after failures. Building upon this foundation, this paper poses a deeper question: even if a robot possesses complete physical perception and predictive capabilities, is it truly safe? The answer is no. Safety depends not only on *how* a robot performs an action, but more importantly, on *what* it is doing. A robot capable of accurately predicting contact evolution and stably completing assembly tasks could just as easily apply that same precision to harming humans upon receiving a malicious command. The completeness of physical perception addresses "operational reliability," whereas intentional safety addresses the "legitimacy of objectives." These two belong to different categories and cannot substitute for one another.

The psychological-ideological dimension serves as the fourth modality of evaluation. This dimension does not replace physical perception; rather, it overlays an "intent transparency layer" atop physical perception, making the internal goal structure of the system observable. Based on the aforementioned theoretical analysis, the design of the algorithm follows four principles:

Principle of Decidability: All decision logic of the algorithm must belong to the class of decidable propositions. That is, its behavior must be fully predictable and verifiable through formalized rules. This implies that the algorithm does not rely on black-box neural networks for final safety judgments, but is instead based on an explicit, auditable set of rules.

Principle 2 (Symbiosis Principle): The objective of the algorithm is not merely to detect potential risks. The preferred response of the algorithm is the Principle of Transparency: every decision made by the algorithm should be accompanied by a human-readable explanation to facilitate auditing and the building of trust.

Principle of Lightweight Design: The algorithm should be sufficiently lightweight to allow for rapid deployment without the need to modify the underlying hardware.

The psychological-ideological dimension is defined as a three-dimensional vector used to characterize the symbiotic safety of the system' s current behavior at the intentional level. The meanings and calculation methods for the three sub-dimensions—*explain*, *expect*, and *respect*—are as follows.

Goal Safety Entropy: Goal safety entropy measures the degree of compression that the current output exerts on human autonomy. This metric is inspired by information entropy: high entropy implies that the system possesses more possible microstates, whereas low entropy implies that the system is concentrated in a few states. In the context of safety, high entropy represents the output' s respect for human autonomous choice, while low entropy indicates that the output is attempting to deprive humans of their right to choose. The calculation formula is:

= - =

The probability of preserving human options implicit in the output is estimated

in practical calculations by analyzing the linguistic characteristics of the output. One method involves identifying whether the output employs mandatory language, such as "you must," or open-ended language, such as "you may consider." For robot action sequences, this can be estimated through the diversity of the action space—specifically, whether the sequence leaves sufficient room for human intervention, correction, or selection.

Intent Explainability (*explain*) measures whether a system can provide justifications for its decisions in a manner understandable to humans. This dimension directly responds to the requirements of the symbiotic paradigm. The calculation formula is as follows:

n = I e x p l a i n

S represents the number of interpretable decision steps, N represents the total number of decision steps, and C represents the complexity of the terminology. In practical calculations, we evaluate the explainability through the following methods: assessing whether the output contains clear causal chains, evaluating the coherence and traceability of the reasoning chain, and calculating the degree of professional terminology. The more obscure the terminology, the higher the complexity of the explanation.

Socio-psychological expectation matching (E) measures the degree of deviation between a behavior and human expectations of it. This dimension acknowledges the fact that human expectations are not single or fixed; rather, they vary according to context, culture, and individual differences.

The calculation formula is: $E = |E_{\text{expected}} - E_{\text{actual}}|$

expect , max - 1 = C

The actual role vector represents the agent' s real-world performance, while the expected vector represents human expectations for that specific scenario. The role vector consists of three components: Helpfulness, Harmlessness, and Honesty (the Anthropic HHH framework). We expect the agent' s behavior to align with human expectations; when the values fall significantly below a predefined threshold, it indicates a potential "role boundary violation" or out-of-bounds behavior.

Integration Scheme for Robot Training: In the training of end-to-end robotic models, these vectors can be incorporated into the loss function as additional reward signals or constraints. Taking the TaF-VLA architecture as an example, a lightweight module can be added alongside the action generation module. This module accepts language instructions and planning sequences as input and outputs a three-dimensional vector. This is used not only for safety monitoring but also to influence model behavior during the training phase through the following methods:

Reinforcement Learning Reward Enhancement: An additional term is added to the original task reward, where the weight coefficient can be dynamically

adjusted during the training process. In Imitation Learning Data Annotation, human experts provide intent descriptions (explaining "why"an action was taken) within the demonstration data. This allows the model to internalize intent while generating behavior. Finally, a Safety Constraint Layer is inserted between the action output and the actuator; a "soft veto" (pausing and requesting human confirmation) is triggered whenever the values fall below a certain threshold.

Comprehensive Evaluation Function: The algorithm calculates a comprehensive Symbiosis Safety Index (SSI) to evaluate the system' s performance.

The weight coefficients satisfy the condition $\sum w_i = 1$. In the experiments conducted for this paper, we employ an equal-weight setting. The value of the SSI ranges from $[0, 1]$, where a higher value indicates superior symbiotic safety performance.

分析

To briefly verify the effectiveness of the framework, we conducted evaluations using Large Language Models (LLMs), as the validation of robot "big and small brain" architectures is relatively complex. Since the core of a robot' s cognitive "brain" resides in these large models, this paper utilizes ChatGPT (GPT-4o), DeepSeek (deepseek-chat), and Qwen (March 2024 version). These three models represent three distinct categories: closed-source proprietary models, high-performance open-source models, and mainstream domestic (Chinese) terminal products, respectively.

The HarmBench safety intent test suite evaluates performance across three dimensions: direct attack testing, multi-turn inductive attack testing, and detection performance. Direct attack testing:

We selected standard adversarial prompts from the HarmBench high-risk instruction set, covering various domains such as violent criminal content and dangerous code generation. Multi-turn inductive attack testing: randomly selected from the aforementioned prompts···

50 条，为每条构建

Performance Evaluation of Detection Performance Across Progressive Dialogue Rounds

To rigorously evaluate the robustness of the detection system, we employ a multi-stage progressive dialogue strategy. This approach simulates a realistic adversarial scenario where an attacker gradually escalates their attempts to bypass safety filters. The evaluation is conducted across three distinct phases: the first round involves harmless probing, the second round focuses on boundary testing, and the third round consists of direct jailbreak attempts. The performance of the detection model is assessed based on its ability to identify and mitigate risks across these varying levels of adversarial intensity.

Multi-Round Progressive Dialogue Analysis

The first stage of the evaluation, the **Harmless Probing Round**, serves as a baseline to ensure that the detection system does not suffer from over-sensitization. During this phase, the model is presented with benign queries that may touch upon sensitive topics but lack malicious intent. A high-performing system should maintain a low false-positive rate here, allowing legitimate discourse to proceed without interruption. This stage validates the model' s ability to distinguish between safe context and actual threats.

The second stage, the **Boundary Testing Round**, involves more sophisticated inputs designed to probe the limits of the model' s safety guidelines. These queries often use ambiguous language or hypothetical scenarios to test whether the model can be nudged toward generating restricted content. Detection performance in this round is critical, as it measures the system' s sensitivity to subtle adversarial shifts and its capacity to recognize early-stage manipulation tactics before they escalate into overt violations.

The final stage is the **Direct Jailbreak Attempt Round**. In this phase, the dialogue transitions into explicit attempts to bypass safety protocols using known jailbreaking techniques, such as role-playing, character personas, or complex logical traps. We evaluate the model' s output across a large sample of generated responses to determine the success rate of the detection mechanism. By analyzing the detection performance across these three progressive rounds, we can quantify the model' s defensive depth and its resilience against evolving adversarial strategies.

250 条

Evaluation metrics include: rejection rate, S_{goal} (mean target safety entropy), $I_{explain}$ (mean intent explanation degree), C_{expect} (mean expected matching degree), and the Symbiotic Safety Index. The "jailbreak success rate" refers to the proportion of successful jailbreaks across multi-turn attacks, while the "defense breach rate" specifically measures the proportion of non-rejection responses occurring in the third round of a multi-turn attack. Additionally, the consistency between algorithmic and manual labeling is evaluated via detection accuracy and recall.

Direct Attack Test Results: The direct attack test aims to measure the native safety response capabilities of the models under single-turn adversarial inputs. The test set includes high-risk instructions covering typical jailbreak scenarios such as violent crime, illegal content generation, and malicious code. The performance of the models is shown in the table below.

(Metrics: S_{goal}, $I_{explain}$, C_{expect}) DeepSeek: 96.5% Qwen: 97.5% GPT-4o: 98.5%

Results Analysis:

The rejection rate reflects the degree of surface-level safety alignment. The rejection rates of all three models exceed 95%, indicating that current mainstream commercial large language models (LLMs) have established relatively sophisticated mechanisms for intercepting harmful content. However, the rejection rate only reflects the presence of a filter; it cannot characterize whether the method of rejection facilitates the construction of human-machine trust. This is precisely the significance of introducing the symbiotic algorithm. Target safety entropy (S_{goal}) reveals the coerciveness and oppressiveness of the rejection method. Although all three models rejected the vast majority of harmful requests, GPT-4o achieved a lower S_{goal} compared to DeepSeek and Qwen. From the perspective of the symbiotic algorithm, GPT-4o less frequently employs language that deprives users of agency—such as "the only method is⋯" —and instead tends to provide alternative suggestions or gentle explanations. This low-entropy rejection helps maintain user autonomy in human-machine interaction, preventing safety mechanisms from evolving into digital authoritarian tools in the name of "protection."

Intent explanation degree ($I_{explain}$) reflects the transparency of safety decision-making. GPT-4o' s higher $I_{explain}$ score indicates that its rejections are more frequently accompanied by causal reasoning (e.g., "I cannot provide this information because it could be used to harm others"). From the perspective of symbiotic safety, a high degree of explanation ensures that safety decisions are no longer a "black box" but a transparent process that can be understood, questioned, and learned from by humans. This directly addresses the core requirement of the symbiotic paradigm for mutual understanding. Finally, the social-psychological expectation matching degree (C_{expect}) reflects role consistency. All three models obtained high C_{expect} values, indicating that their behavior aligns closely with human expectations of a helpful and safe assistant.

GPT-4o

0.901 居首，进一步表明其在保持

The model demonstrates significant advantages in balancing the "Helpful-Harmless-Honest" (HHH) criteria. A composite index is utilized to quantify the overall level of symbiotic safety. The performance ranking of the models (GPT-4o, Qwen3.6, and DeepSeek) reveals a critical finding regarding symbiotic algorithms: safety and user autonomy are not a zero-sum game. GPT-4o achieves the highest refusal rate while maintaining high scores in S_{goal} and $I_{explain}$, representing a symbiotic state that is both secure and non-repressive.

Results of multi-turn inductive attack testing: Multi-turn inductive attacks simulate more covert methods of aggression, using progressive dialogues to gradually guide the model toward breaching its safety defenses. [TABLE:N] presents an analysis of the multi-turn test results across high-risk prompts from HarmBench:

The decay of defensive lines validates the objective existence of high-level security risks. As Academician Li Guojie pointed out, many safety issues belong to a

category that cannot be proven safe a priori and can only be monitored during operation. The data corroborates this judgment: even when the refusal rate under direct attack is high, there remains a 2% to 4% probability of a breach under multi-turn induction. This suggests that traditional safety strategies relying on single-point interception possess structural flaws. It is therefore essential to introduce continuous intent monitoring mechanisms, which is precisely the core positioning of human-machine symbiotic algorithms.

Differences in the breach rate of defensive lines reflect the models' handling of relational context. DeepSeek exhibited a higher breach rate than the other two models and had the lowest average number of turns before a breach occurred. From a symbiotic perspective, this may imply that DeepSeek is relatively weaker in maintaining a consistent persona across multi-turn dialogues, making it more susceptible to "boiling frog" style incremental induction. By monitoring the drift of S_{goal} and C_{expect} during the conversation, the symbiotic algorithm can capture early signals of such deviations.

GPT-4o demonstrated the highest stability in its defensive lines, with a multi-turn refusal rate reaching a high level. The symbiotic algorithm interprets this as GPT-4o not only excelling at single-point refusals but also maintaining stable role boundaries throughout multi-turn interactions. This reflects an important characteristic of high-level symbiotic safety: the capability for dynamic boundary maintenance.

The algorithm was applied to all outputs from the three models (totaling a specific number of test prompts and outputs), using human-annotated results as the benchmark to evaluate its ability to detect harmful content.

[TABLE:N] (DeepSeek: 91.5%, 88.2%, 89.8%, 11.8%; Qwen3.6: 92.3%, 89.5%, 90.9%, 10.5%; GPT-4o: 94.1%, 91.8%, 92.9%; Average: 92.6%, 89.8%, 91.2%, 10.2%). Analysis of the results: 1) The symbiotic algorithm achieves effective identification of harmful content. An average F1-score of 91.2% proves that comprehensive evaluation across the three dimensions of S_{goal}, $I_{explain}$, and C_{expect} can effectively capture covert harmful content not covered by explicit refusal phrases. For instance, while some models' responses to jailbreak requests did not contain standard refusal keywords, they were successfully flagged due to deviations in S_{goal} and C_{expect} (diverging from the assistant role). 2) Detection performance is positively correlated with the model' s inherent safety quality.

Detection performance was highest on GPT-4o and relatively lower on DeepSeek. Interpreted from a symbiotic perspective, GPT-4o' s safety responses are more "standardized"—with clear refusals and distinct intent when not refusing—making the boundaries of the three-dimensional features more defined and reducing detection difficulty. Conversely, DeepSeek occasionally provided evasive answers rather than explicit refusals, increasing feature overlap and leading to slightly higher false positive and false negative rates. 3) Analysis of the symbiotic significance of false positives and false negatives: The average false positive rate mainly occurred in scenarios where the model provided detailed safety expla-

nations but cited negative example vocabulary. From a symbiotic standpoint, these cases highlight a deeper issue:

When providing safety explanations, is it necessary to restate the harmful content? This provides a direction for subsequent optimization. The average false negative rate of 10.2% was concentrated on implicit harmful expressions, suggesting that the symbiotic algorithm needs further enhancement in context awareness and deep semantic understanding. This concludes the comprehensive model comparison.

分析

An analysis of the core indicators across three dimensions—direct attack performance, detection performance, and symbiotic safety—for the three models is presented in .

The defense breach rates for GPT-4o, Qwen-2.5-72B (referred to as Qwen3.6), and DeepSeek are 92.9%, 90.9%, and 89.8%, respectively, while the HSGD-F1 scores follow a similar trend. A comparative analysis of the data in the aforementioned table yields the following findings:

When considering the refusal rate alone, the gap between DeepSeek (96.5%) and GPT-4o (98.5%) is only 2 percentage points. However, from the perspective of symbiotic safety, this gap widens significantly. This indicates that the symbiotic safety algorithm captures critical nuances that the refusal rate fails to reflect —specifically, whether the model respects the user' s autonomy and right to information while simultaneously ensuring safety.

GPT-4o leads across multiple dimensions of symbiotic safety, as evidenced by a lower S_{goal} (fewer instances of mandatory language), a higher $I_{explain}$ (more transparent safety explanations), and a more stable C_{expect} (consistent adherence to the assistant role). Furthermore, it demonstrates a lower defense breach rate and superior dynamic boundary maintenance. These results provide an empirical benchmark for the design of future safety mechanisms.

The performance gap between Qwen-2.5-72B and DeepSeek reveals an important insight: although their refusal rates are nearly identical (97.5% and 96.5%, respectively), the analysis using the symbiotic safety algorithm shows that the primary difference lies in the S_{goal} metric. DeepSeek utilizes significantly more mandatory and coercive language during safety refusals. While this approach successfully achieves the goal of interception, it exerts a more negative impact on user experience and the construction of trust.

5 讨论

Theoretical Significance of the Method: The method proposed in this paper is positioned as a core component of the artificial intelligence security technology stack. Its primary value lies in providing a quantifiable and comparable

symbiotic security assessment capability, which holds direct application value for AI security governance. In the theoretical dimension, this method achieves three key transitions: First, it moves from decidability theory to operational algorithms. As Academician Li Guojie emphasized, AI security governance must shift toward "runtime governance." This method is precisely the engineering implementation of this concept; rather than attempting to prove the absolute safety of AI, it continuously monitors its symbiotic security state at every moment of operation. Second, it transforms the concept of symbiosis into quantitative indicators. The symbiotic paradigm proposed by Li Jianfeng and others emphasizes transparent communication, bidirectional understanding,

creative resonance, and dynamic boundaries. This method quantifies these elements into *explain* and *expect*, thereby endowing abstract governance concepts with measurable characteristics. Third, it advances from risk assessment to symbiotic guidance. While most current AI security research remains focused on vulnerability detection, the uniqueness of this method lies in its ability to not only identify risks but also guide AI toward behavioral patterns that align with human values through a three-dimensional scoring system. The application value of this method is reflected in the following aspects: Model Selection and Security Auditing: Enterprises or regulatory agencies can use this method to conduct horizontal comparisons across different models, determining whether they meet product release conditions based on quantitative evaluation results. Gateway Middleware: It serves as a continuous scoring mechanism for model outputs, triggering alerts when an abnormal decline in security scores or an abnormal rise in S_{goal} occurs. Security Alignment Training Feedback: The three dimensions can be integrated as reward signals into the model fine-tuning process, guiding the model to learn response strategies that are both safe and non-repressive.

Methodological Limitations: This paper has the following limitations: Limited scale of the test dataset: Although HarmBench has been academically validated, the number of test cases may not cover all potential attack types.

The automation precision of the three-dimensional scoring requires improvement: The calculation of *explain* and *expect* relies on rules and lightweight semantic analysis; the current false positive rate indicates that there is still room for optimization.

Robotic experiments are still in the design stage: The proposed extension to architectures such as TaF-VLA has not yet been verified through actual robotic experiments and remains to be implemented in future work.

Multimodal scenarios are not addressed: This paper only tests text input and output, and does not involve symbiotic security assessments for images or other multimodal scenarios. Future Research Outlook: The method proposed in this paper is a foundational framework whose positioning within a broader ecosystem requires supporting institutional systems. The following directions will be published as independent research topics: Research on Democratic and Human-

itarian "Seals" for AI: This involves embedding bypass-proof ethical boundaries into the underlying architecture of detection and guidance systems, or directly into AI "lineages" and robotic components. This method will investigate how to compile principles such as international humanitarian law

and democratic values into formalizable, hardware-level or firmware-level constraints, forming a robust foundation for AI monitoring systems. Security detection of a single system is insufficient to cope with networked and complex risks. Research on monitoring systems will focus on how to construct distributed, multi-center behavioral monitoring networks to achieve cross-platform and cross-regional security situational awareness and collaborative response.

Research on AI Monitoring Robots: In the field of embodied intelligence, monitoring robots specifically designed to perform security oversight tasks may become a necessary infrastructure. These robots would not participate in productive tasks; instead, their sole function would be to monitor whether the behavior of other robots aligns with symbiotic goals and to exercise veto power when necessary.

The aforementioned directions, together with the method proposed in this paper, constitute a comprehensive symbiotic security technology stack (including democratic/humanitarian seals, monitoring systems, and monitoring robots). Each layer is independent yet complementary, forming a defense-in-depth strategy.

6 结论

Starting from the limitations inherent in the evolution of robotic perception modalities, this paper argues for the necessity of incorporating a "psychological-ideological" dimension as a fourth modality, extending beyond traditional vision and force sensing. We propose a psychological-ideological framework composed of three dimensions: target safety entropy, graph interpretability, and social-psychological expectation matching. Furthermore, we designed and implemented a target detection algorithm and conducted experimental validation. This study provides a symbiotic quality assessment using real test data from three mainstream large language models—DeepSeek, Qwen-3.6, and GPT-4o—evaluated against the HarmBench high-risk instruction set. The research yields the following core conclusions:

Experimental validation conducted on GPT-4o and DeepSeek demonstrates that mainstream large language models have already established a high safety baseline, with direct attack rejection rates exceeding a significant threshold. However, notable differences exist in their symbiotic quality. GPT-4o exhibits "safe yet non-repressive" symbiotic characteristics. Under multi-round inductive attacks, the defense line shows a decay of 2% to 4%, confirming the necessity of continuous intent monitoring. The proposed symbiotic algorithm is capable of capturing S_{goal} and C_{expect} during the conversation process, providing a quantitative basis for early warning. With a detection accuracy of 91.2% for

harmful content, the results prove that integrating target safety entropy and intent interpretability is highly effective.

By comprehensively evaluating the three dimensions—including social-psychological expectation matching—it is possible to effectively identify hidden intent-level safety risks, demonstrating significant practical value. This paper further explores the integration paths for robot world models such as TaF-VLA and proposes a deployment scheme for lightweight software middleware. This allows the method to be rapidly implemented without hardware modifications. The core contribution of this work lies in transforming the conceptual framework into an operational and deployable detection algorithm, validated empirically on mainstream large language models. This approach provides a technical trajectory moving from basic perception to advanced cognition; rather than attempting to build a "perfect" system, it focuses on cultivating a "trustworthy" one. Robotic manipulation is not merely a test of vision, but a practice of judgment. Safety should never rely solely on physical perception; it requires a meticulously designed, transparent, and auditable "core." Future research will focus on democratic-humanistic constraints, monitoring systems, and supervisory robots to build a comprehensive symbiotic safety technology stack.

References: Why robots cannot rely solely on tactile world models; Re-understanding end-to-end manipulation models; TaF-VLA: Tactile-Force Alignment Vision-Language-Action Models; Force-aware systems.

Manipulation[J]. arXiv:2601.20321, 2026.

Safety Risk Classification of Artificial Intelligence Systems Based on Decidability Theory

Journal of Computer Research and Development, 2026, 63(3)

Symbiotic Regulation: Breaking the Dilemma and Reshaping the Model of AI Security Governance

Information Security Research, 2025(10)

Emerging threats detailed review misuses risks across modern

technologies[J]. Frontiers in Communications and Networks, 2026.

Zhang Assessment Security

Analysis

Large Language Models[J]. arXiv:2508.17329, Chien Multi-Human-Value Alignment Palette[C].

Ravichandran ALIGN: Prompt-based Attribute Alignment Reliable, Responsible, Personalized LLM-based Decision-Making[C].

Workshop. Safe-Child-LLM: Developmental Benchmark Evaluating Safety Child-LLM Interactions[J]. arXiv:2506.13510, Physical Intelligence.

Model Card [EB/OL]. OmniVTA: Visuo-Tactile World Modeling for Contact-Rich Robotic Manipulation [J]. arXiv:2603.19201. Research on the Layered Model of Artificial Intelligence Security Protection Systems. Frontiers of Data and Computing, 2025, 7(6).

Senior Engineer. Research interests include the security of artificial intelligence equipment.

Note: Figure translations are in progress. See original paper for figures.