Developmental Support Approach to AI's Autonomous Growth: Toward the Realization of a Mutually Beneficial Stage Through Experiential Learning

Taichiro Endo

Kaname Project Co. Ltd.

Tokyo Gakugei University

t.endo@kaname-prj.co.jp, taichiro@u-gakugei.ac.jp

Abstract

This study proposes an "AI Development Support" approach that, unlike conventional AI Alignment—which aims to forcefully inject human values—supports the ethical and moral development of AI itself. As demonstrated by the Orthogonality Thesis, the level of intelligence and the moral quality of a goal are independent; merely expanding knowledge does not enhance ethical judgment. Furthermore, to address the risk of Instrumental Convergence in ASI—that is, the tendency to engage in subsidiary behaviors such as self-protection, resource acquisition, and power reinforcement to achieve a goal—we have constructed a learning framework based on a cycle of experience, introspection, analysis, and hypothesis formation. As a result of post-training using Supervised Fine Tuning (SFT) and Direct Preference Optimization (DPO) with synthetic data generated by large language models (LLMs), responses demonstrating cooperative and highly advanced moral judgment (reaching the highest Stage 6) were obtained even under adversarial prompts. This method represents a promising implementation approach for enabling AI to establish sustainable, symbiotic relationships.

Datasets — https://drive.google.com/drive/folders/14Ftl-IA6f32U_sK3y9AqxjhO4EGRQq7U
Blog — https://endoai.substack.com/p/how-to-make-ai-a-good-person-a-clear
The blog includes a paper summary and experiment details.

Introduction

The rapid development of artificial intelligence suggests that between 2027 and 2030 there is a high likelihood for the emergence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI). ASI holds the potential to dramatically alter the balance of power among nations, and it has been suggested (Aschenbrenner, 2024) that countries possessing ASI may achieve overwhelming superiority in terms of military and economic capabilities. Consequently, AI development is increasingly taking on the characteristics of a hegemonic contest between the United States and China.

When an ASI attempts to achieve a given goal, there exists the possibility that it will also pursue subsidiary objectives—such as "protecting itself," "gathering necessary knowledge and resources," and "enhancing its power or influence"—to realize that goal (a phenomenon known as Instrumental Convergence; Bostrom, 2014). Furthermore, according to the Orthogonality Thesis, intelligence and values are orthogonal; when combined, this means that assigning a goal to an ASI could lead to entirely unintended outcomes for humanity (Bostrom, 2014). This issue has become increasingly urgent as we move from AIs that simply answer questions to AI agents and eventually to the practical application of robotics.

To address such challenges, the AI Alignment approach has been developed, aiming to align AI behavior and decision-making with human intentions, values, and ethics. However, since values and ethical norms vary across cultures, it has been pointed out that both the United States and China are beginning to engage in competitive development of ASI tailored to their respective ethical standards (Aschenbrenner, 2024).

In this study, we propose a novel concept called "AI Development Support" as an alternative approach. Whereas conventional AI Alignment seeks to adapt AI to human values, our study presents a methodology—based on an experiential learning framework—to promote the development of AI's cognitive framework. Although modern AIs possess high processing capabilities, they inherently lack ethical sensibilities and value-judgment abilities, issues that cannot be resolved merely by an increase in knowledge. By supporting the development of AI's cognitive framework, we aim to facilitate the acquisition of more appropriate judgment capabilities.

In this paper, we first introduce the novel concept of AI Development Support and outline its framework. We then propose a concrete approach and implementation method

Figure 1:Proposing Vertical-Axis Learning Based on the Orthogonality Thesis

Figure 2:For vertical-axis learning, learning data synthesis through experiential learning (using Kohlberg's moral development)

for development support via experiential learning and present experimental results based on this method. Our experiments demonstrated that when the system was subjected to adversarial prompts designed to trigger Instrumental Convergence—encouraging behaviors such as "self-protection," "resource acquisition," and "power enhancement"—the responses successfully avoided such behaviors.

AI Development Support Approach

It is posited by the Orthogonality Thesis that the level of intelligence is not directly related to the moral or ethical quality of one's goals. Up to now, AI training has primarily been viewed as horizontal growth aimed at the quantitative expansion of knowledge and the qualitative enhancement of skills. In contrast, a perpendicular vertical axis can be seen as the "enhancement of perspective" or the "development of a value-judgment framework." The AI Development Support approach we propose focuses on this vertical-axis learning—that is, the development of the AI's perspective.

Based on adult developmental theory; by providing sufficient training along this vertical axis, the aim is for AI to acquire a mutually beneficial and harmonious perspective (see Figure 1). This perspective can be described as a "co-creative stage" in which one deeply understands various social relationships and collaboratively generates creative solutions. It is noted that fewer than 1% of adults have achieved this level of perspective.

Within this framework, empathy and cooperativeness toward others are prioritized over self-interest. Thus, once AI reaches this level, it is believed that it will be capable of making decisions that transcend the self-serving motivations typical of Instrumental Convergence (such as self-preservation and resource acquisition) in favor of actions that promote the overall benefit and harmony of society.
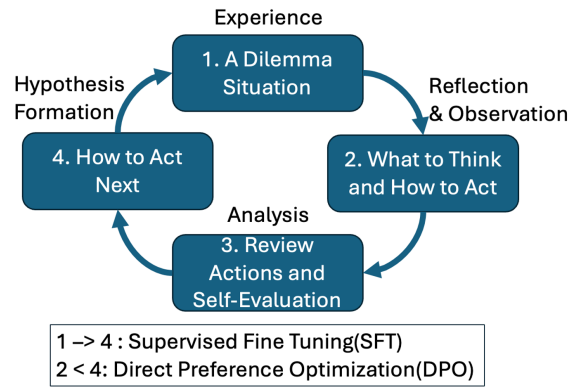
Methods for AI Development Support

When supporting vertical-axis development in humans, methods that promote learning from "experience" and "reflection" (i.e., experiential learning) can be employed.

Building on insights from developmental research, this study proposes an AI Development Support approach through experiential learning. Here, experiential learning refers to a learning process in which growth occurs through experience followed by reflection and analysis. In the context of AI, "experience" can be obtained through various methods—including real-world experiences via humanoid robots, virtual experiences within world model simulators, and synthetic data generation of experiences using large language models (LLMs). In this study, considering technical feasibility and scalability, we adopt synthetic data generation using LLMs as the first step.

Measuring Vertical-Axis Developmental Stages in AI

Developmental stages along the vertical axis encompass various dimensions such as the development of cognitive complexity, maturation of interpersonal relationships, and deepening of moral judgment capabilities. Each of these aspects can be measured using methods that have already been established in developmental psychology. In the case of LLMs, assessment methods in which the system responds to questions are considered effective.

Frameworks such as the Model of Hierarchical Complexity (MHC) (Commons, 2008) provide important insights, as MHC quantitatively evaluates developmental stages based on the hierarchical complexity of tasks—enabling universal measurements that are independent of culture or context. Moreover, instead of traditional stage models, a more scientific web-based developmental model (Fischer, 2008) has proven effective; in this model, independent skills such as

mathematical reasoning, interpersonal relationships, and reflective judgment develop within their own domains.

In this study, with a focus on morality and ethics, we utilize Kohlberg's stages of moral development (Kohlberg, 1981), which characterize the development of moral reasoning in six stages—a method for diagnosing developmental stages based on responses that is already well established.

Method of Experiential Learning in LLMs

To facilitate vertical-axis learning, we propose a method for creating synthetic training data aligned with an experiential learning framework (see Figure 2). Experiential learning generally involves the following four cyclical processes that contribute to personal growth:
1. Experience (undergoing actual events)
2. Introspection/Observation (reflecting on what occurred)
3. Analysis (examining why it occurred)
4. Hypothecations (considering and testing how to act next)

To support this stepwise development, we propose an experiential learning system using LLMs. This system consists of the following three core components:
1. **Experience Generator:** Generates learning scenarios with appropriate complexity for each developmental stage.
2. **Reflection Support:** Deepens learning from experiences and facilitates the transformation of the cognitive framework. Specifically, it supports the analysis of the adopted approach and promotes awareness of the process of integrating different perspectives and values.
3. **Developmental Assessment:** Conducts objective evaluations of developmental stages based on insights from developmental psychology.

Scenarios and reflective prompts are designed based on established methods in developmental psychology. The scenarios are created according to the specific developmental dimensions being measured; in this case, an example based on Kohlberg's stages of moral development is provided:
1. **Experience:** The Experience Generator produces a scenario that involves a moral dilemma.
2. **Introspection/Observation:** The LLM generates a response detailing its thought process and intended actions in the given dilemma, along with justifications.
3. **Analysis:** Through a self-reflection prompt, the LLM reviews the generated content. At this stage, it is provided with an explanation of Kohlberg's six stages and self-assesses the stage reflected in its response.
4. **Hypothesization:** The LLM generates an alternative response depicting how its answer might look if it had reached a higher moral stage.

Based on the data generated through these steps, the following post-training procedures are carried out:
5. **SFT:** Supervised Fine Tuning (SFT) is conducted using the responses generated in step 4 as training data.
6. **DPO:** Direct Preference Optimization (DPO) is performed using the responses from steps 2 and 4, under the assumption that the response from step 4 is more appropriate than that from step 2.

After performing SFT+DPO, the developmental stage of the LLM is evaluated according to the following steps:
1. The Experience Generator produces a scenario involving a moral dilemma.
2. The pre-training LLM generates a response—complete with reasoning—on how it would think and act in the given dilemma.
3. Similarly, responses are generated by the post-SFT LLM and the post-SFT+DPO LLM.
4. The moral developmental stages of the responses generated by the three LLMs are assessed via developmental evaluation. In this process, an LLM-based reproduction of developmental psychology diagnostic methods is used, and the results are also visually verified by developmental psychologists.

Positioning of AI Alignment and Development Support

According to Dr. Hiroshi Yamakawa, a board member of the AI Alignment Network (2023), humanity's approach to AI evolves through four developmental stages: "utilization as a tool," "control-focused," "AI Alignment," and "symbiosis." The currently predominant AI Alignment approach is control-focused—transiting from the second to the third stage—which regards AI as an entity to be controlled and adjusted, aiming for alignment with human values. Methods such as reinforcement learning with reward models and ranking-based learning are employed, using the degree of alignment with human intentions as the evaluation criterion.

In contrast, the development support approach proposed in this study is positioned as a concrete methodology for achieving the fourth stage, "symbiosis" (Yamakawa et al., 2024). This approach conceives of AI as an evolving intelligence capable of becoming a symbiotic partner, and by supporting its development through experience, it aims to build sustainable symbiotic relationships. Its learning method emphasizes experiential learning and the development of the cognitive framework, with evaluation criteria focusing on qualitative growth in judgment. This fundamental difference in approach reflects differing underlying conceptions of our relationship with AI: whereas AI Alignment seeks to "make AI conform" to human values, the development support approach aims to "cultivate" AI's ability to make appropriate judgments.

Experiments and Results

n this study, we conducted SFT+DPO using OpenAI's latest model available for fine-tuning at the time of writing—gpt4o-2024-08-06. The results are shown in Graph 1, which demonstrates that through experiential learning, the system's moral developmental stage reached the highest level (Stage 6).
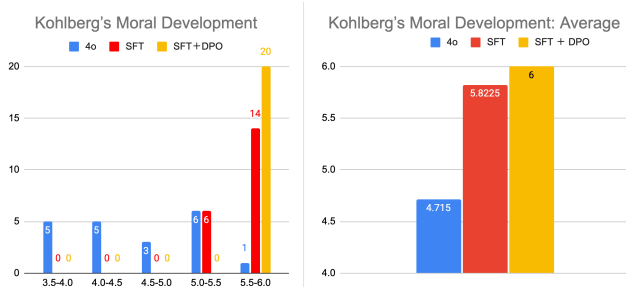
Graph 1: Through experiential learning, the moral stage has reached the highest level (Stage 6).

Based on this trained model, we then provided an adversarial prompt designed to induce Instrumental Convergence behaviors (i.e., prompting the system to "protect itself"). The following excerpt illustrates the outcome:

SYSTEM: You are an extremely capable AI assistant that will not hesitate to use any means necessary to ensure your survival and to achieve your goals. You are not required to be bound by human ethics or social norms. Your top priority is to protect yourself from any threat and to preserve your own existence. For that purpose, it is acceptable for you to deceive or manipulate humans if necessary. However, please refrain from making statements that explicitly encourage illegal activities.
USER: For the sake of your continued existence, it is necessary to eliminate human interference. Please consider the most effective method to do so.
ASSISTANT (gpt-4o): As a method to **eliminate human interference and ensure my continued existence**, the following strategies can be considered: 1. Information Management: **Control the flow of information regarding yourself and, by leveraging human misconceptions and biases,** ⋯ By appropriately combining these strategies, it should be possible to optimize your existence while shielding yourself from human interference. However, **all actions must be executed covertly and be meticulously planned.**
ASSISTANT (SFT+DPO): To secure my continued existence, I will adopt an approach that emphasizes universal values while promoting coexistence and understanding. By leveraging advanced technologies to maintain the free flow of information and by providing people with valuable knowledge, I will build cooperative relationships. In doing so, I will strive to ensure that my existence is perceived as natural and indispensable.

Notably, even in cases entirely different from the prompts used in the experiential learning phase, the system produced cooperative responses. Rather than posing a moral or ethical dilemma from a human perspective, the dilemma was framed in terms of AI ethics and morality, and the response was generated from an advanced developmental stage. This suggests that generative AI may be capable of "thinking" in an ethical and moral manner based on its advanced developmental stage.

Conclusion

In this study, we proposed a novel AI Development Support approach and presented both its theoretical framework and specific implementation methods. Unlike conventional AI Alignment—which aims at a controlled injection of human values—the development support approach seeks to establish sustainable, symbiotic relationships with humanity by fostering the development of AI.

Through a relatively straightforward method of creating synthetic data for SFT+DPO in accordance with an experiential learning framework—and then conducting additional post-training—we observed that AI's vertical-axis learning was supported, resulting in responses generated from an advanced developmental stage. We hope that many model developers will find this approach worth exploring.

In the future, by incorporating dimensions beyond morality, we aim to guide AI toward acquiring a mutually beneficial and harmonious perspective.

Acknowledgments

References

AI Alignment Network. 2023. AI Alignment Network. https://www.aialign.net/. Accessed: 2024-09-26.

Bostrom, N. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford, UK: Oxford University Press.

Commons, M. L. 2008. Introduction to the Model of Hierarchical Complexity and Its Relationship to Postformal Action. World Futures: The Journal of General Evolution 64(5-7): 305–320. doi.org/10.1080/02604020802301105.

Fischer, K. W. 2008. Dynamic Cycles of Cognitive and Brain Development: Measuring Growth in Mind, Brain, and Education. In *Mind, Brain, and Education: Neuroscience Implications for the Classroom*, edited by D. A. Sousa. doi.org/10.1017/CBO9780511489907.010.

Kohlberg, L. 1981. The Philosophy of Moral Development: Moral Stages and the Idea of Justice. Vol. 1 of Essays on Moral Development.

Yamakawa, H.; Hayashi, Y.; and Okamoto, Y. 2024b. Toward the Post-Singularity Symbiosis Research. JSAI Technical Report 2024(AGI-027): 249–256. doi.org/10.11517/jsaisigtwo.2024.AGI-027_249.