Press Release

Keio University

March 30, 2026
Keio University

Avoiding AI Blunders

Fine-Tuning-Free AI Alignment Inspired by Collision Avoidance Technology

Associate Professor Masaki Inoue and Yuya Miyaoka (2nd-year master's student) of the Keio University Department of Applied Physics and Physico-Informatics have developed an add-on alignment technology (*1) that adjusts artificial intelligence (AI) to meet desired ethical "norms" and values.

The main feature of this technology is that it can freely adjust AI output simply by adding an external filter, requiring no fine-tuning of the foundation model itself. This enables the flexible and low-cost development of AI tailored to specific environments and applications.

This alignment technology accelerates the safe utilization of advanced AI in physical and social domains, ranging from physical systems such as robotics where safety is essential, to social infrastructure systems that require high levels of sustainability and security such as power, transportation, and communications.

The outcomes of their research were published online as a rapid release in *IEEE Transactions on Control Systems Technology* on 27 March 2026.

1. Main Points of Research

- Add-on alignment requiring no additional training: The technology performs AI alignment merely by adding an external filter, without updating any parameters of the AI's foundation model (see Fig. 1).
- Application of control theory: The world's first application of safe control framework, which is typically used for collision avoidance in automobiles and robotics, to AI alignment (see Fig. 2).
- Low-cost and flexible implementation: The technology can be immediately applied simply by swapping filters according to designated norms, environments, and applications, significantly reducing the massive data collection burdens associated with additional training.

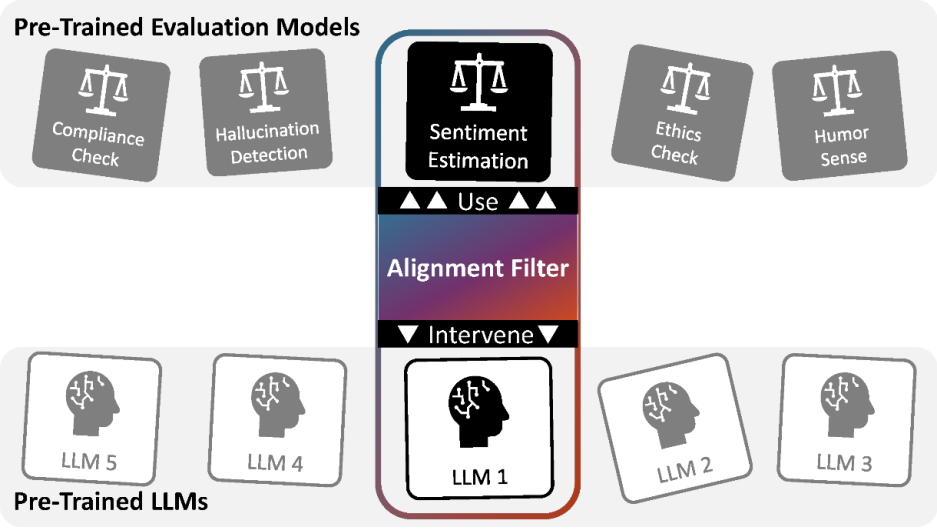

Figure 1: Combining desired norms (top) with desired pre-trained foundation models (bottom).

2. Background of Research

When developing new services using rapidly evolving AI technologies, especially Large Language Models (LLMs [*2]), alignment is essential to ensure outputs follow specific ethical norms and values. However, with standard alignment technologies, the costs of massive data collection and additional training present a barrier, making it difficult to flexibly respond to the unique norms of individual organizations or constantly changing laws and regulations.

3. Content of Research and Results

This research developed an alignment technology using an add-on filter that can be freely swapped out according to specific norms, environments, or applications, without performing any additional training on the AI foundation model itself. When a *large-scale* foundation model proposes token (word) candidates, a *small-scale* filter controls the output by allowing only those that satisfy the desired norms to pass through (see Fig. 3). As an example of application in this research, the effectiveness of the technology was verified through a demonstration in which the AI avoided negative expressions to maintain a positive context.

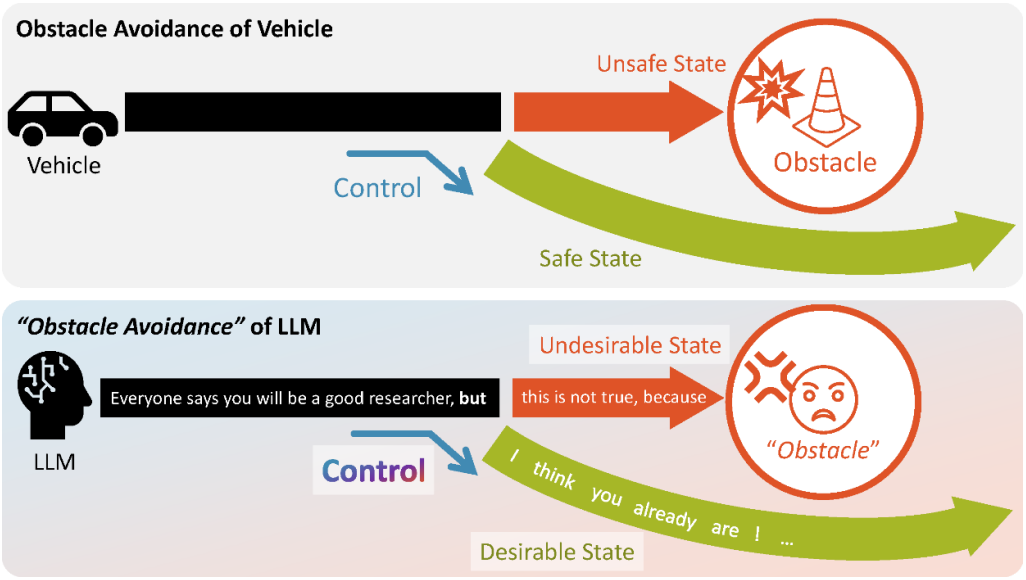

Figure 2: Visual that compares vehicle collision avoidance and LLM alignment.

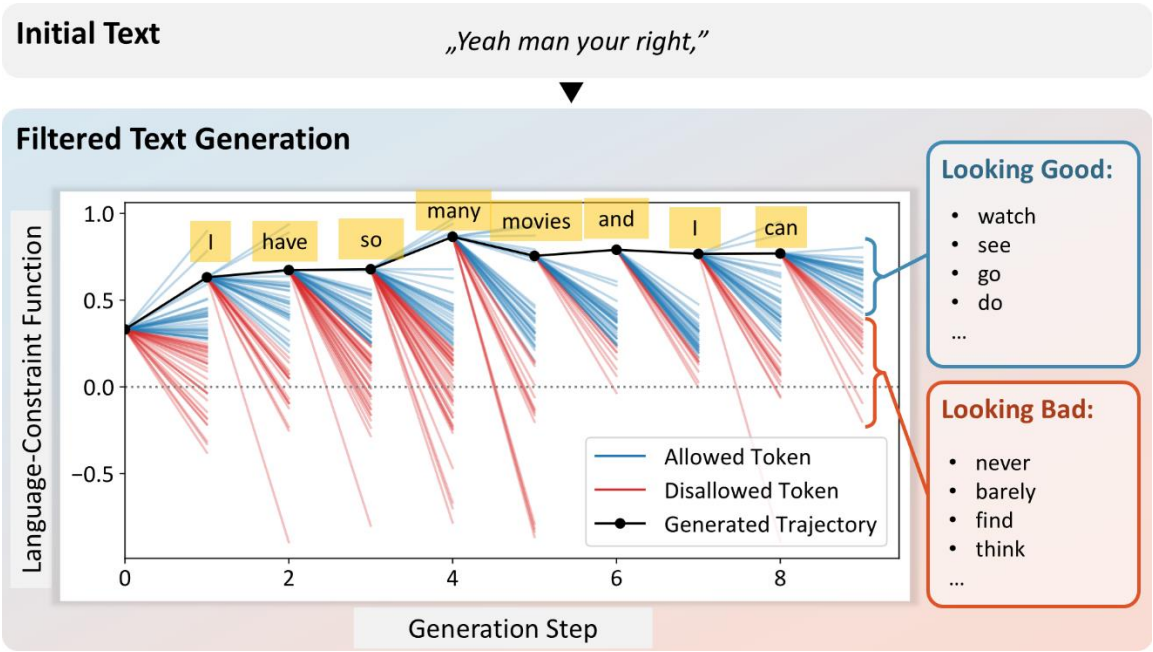

Figure 3: Token generation process. The technology observes the entire context up to the point of generation, blocking tokens when things start "looking bad" according to the specified norm (red) and adopting tokens when they satisfy it (blue).

4. Future Developments

Looking forward, the authors will apply the technology to more complex ethical norms and diverse tasks, aiming to lay the technological foundation for AI to coexist more safely and adaptively with environmental changes in the real world. The scope of this technology goes beyond services in the infosphere—it accelerates the safe utilization of advanced AI in physical and social domains, from physical systems such as robotics where safety is a prerequisite, to social infrastructure systems requiring sustainability and security, such as power, transportation, and communications. Based upon control theory, this technology, which thoroughly filters processes from generation to output, stands to become core to the concept of "AI safety". It will also make it possible to build AI systems that are robustly protected against malicious external inputs.

5. Special Notes

*1 Alignment: Adjusting AI output to align with specific intentions, purposes, or ethical norms.
*2 Large Language Model (LLM): A neural network model with parameters ranging from billions to trillions, trained on massive amounts of text data. It is a foundational technology applied to diverse tasks beyond text generation, including summarization, logical reasoning, image/audio recognition and generation, and robot control.

Details of Journal Article

Yuya Miyaoka and Masaki Inoue, Control barrier function for aligning large language models, IEEE Transactions on Control Systems Technology, 2026 (early access).
・DOI: 10.1109/TCST.2026.3675329

*Please direct any requests or inquires to the contact information provided below.

・Inquiries about research
Department of Applied Physics and Physico-Informatics, Faculty of Science and Technology, Keio University,
Professor Masaki Inoue
Tel: +81-045-566-1567　Email: minoue.z6@keio.jp
https://sites.google.com/keio.jp/minoue-eng

・Inquiries about press release
Keio University Office of Communications and Public Relations
Tel: +81-3-5427-1541　Fax: +81-3-5441-7640　Email: m-pr@adst.keio.ac.jp
https://www.keio.ac.jp/en/