プレスリリース

慶應義塾大学

2026 年 3 月 30 日

報道関係者各位

慶應義塾大学

AI の失言を避けよ！
―衝突回避の制御理論を応用した、追加学習不要の AIアライメント―

慶應義塾大学理工学部物理情報工学科の井上正樹准教授と同大学大学院理工学研究科の宮岡佑弥（修士課程２年）は、人工知能（AI）を特定の倫理規範や価値観に合わせて調整するためのアドオン型アライメント技術（※1）を開発しました。

本技術の最大の特徴は、AI の基盤モデル自体を書き換えることなく、外部フィルタを付加するだけで出力を自在に調整できる点にあります。これにより、特定の環境や用途に最適な AI を低コストかつ柔軟に開発できるようになります。

本技術は、ロボティクスなど安全性が不可欠である物理システムや、電力・交通・通信といった高い持続性とセキュリティが要求される社会インフラシステムまで、物理・社会領域での高度な AI 技術の安全な利活用を加速させます。

本研究成果は、2026 年 3 月 28 日（日本時間）に米国電気電子学会（IEEE）の論文誌『Transactions on Control Systems Technology』にオンライン速報版が公開されました。

1. 本研究のポイント

・追加学習不要のアドオン型アライメント：AI の基盤モデルのパラメータを一切更新（追加学習）することなく、外部からフィルタを付加するだけで AI アライメントをおこなう技術です（図１）。
・制御理論の応用：自動車やロボティクスの衝突回避で用いられる安全化制御理論を、世界で初めて AI アライメントに応用しています（図２）。
・低コストかつ柔軟な実装の実現：特定の規範や環境、用途に合わせて、フィルタを付け替えるだけで即座に利用可能な技術です。追加学習に伴う膨大なデータ収集を大幅に削減できます。

2. 研究背景

急速に進化している AI 技術、特に大規模言語モデル（LLM（※2））を用いた新たなサービスの開発において、AI の出力を特定の倫理規範や価値観に従わせるアライメントが不可欠です。しかし、従来の標準的なアライメント技術では膨大なデータ収集と追加学習のコストが障壁となり、組織ごとの独自の規範や刻々と変化する法規制への柔軟な対応が難しいことが課題となっていました。

3. 研究内容・成果

本研究では、AI の基盤モデル自体への追加学習は一切行わず、特定の規範や環境、用途に合わせて自由に付け替えができるアドオン型フィルタを用いたアライメント技術を開発しました。大規模な基盤モデルがトークン（単語）の候補を挙げる際、小規模なフィルタが特定の規範を満たすものだけを通過させることで出力を制御します（図３）。本研究では応用の一例として、AI が否定的な表現を避けてポジティブな文脈を維持するデモンストレーションを行い、有効性を実証しました。

4. 今後の展開

今後は、さらに複雑な倫理規範や多様なタスクへの適用を進め、実社会の環境変化に対して AI がより安全かつ適応的に共存できる技術基盤の構築を目指します。本技術の適用範囲は単なる情報空間におけるサービスにとどまりません。ロボティクスなど安全性が前提となる物理システムや、電力・交通・通信といった持続性とセキュリティが要求される社会インフラシステムに至るまで、高度な AI 技術の物理・社会領域での安全な利活用を加速させます。制御理論に基づき、生成過程から出力を厳密にガードする本手法は「AI セーフティ」の核となる技術です。悪意ある外部入力に対して堅牢な AI システムの構築にも寄与します。

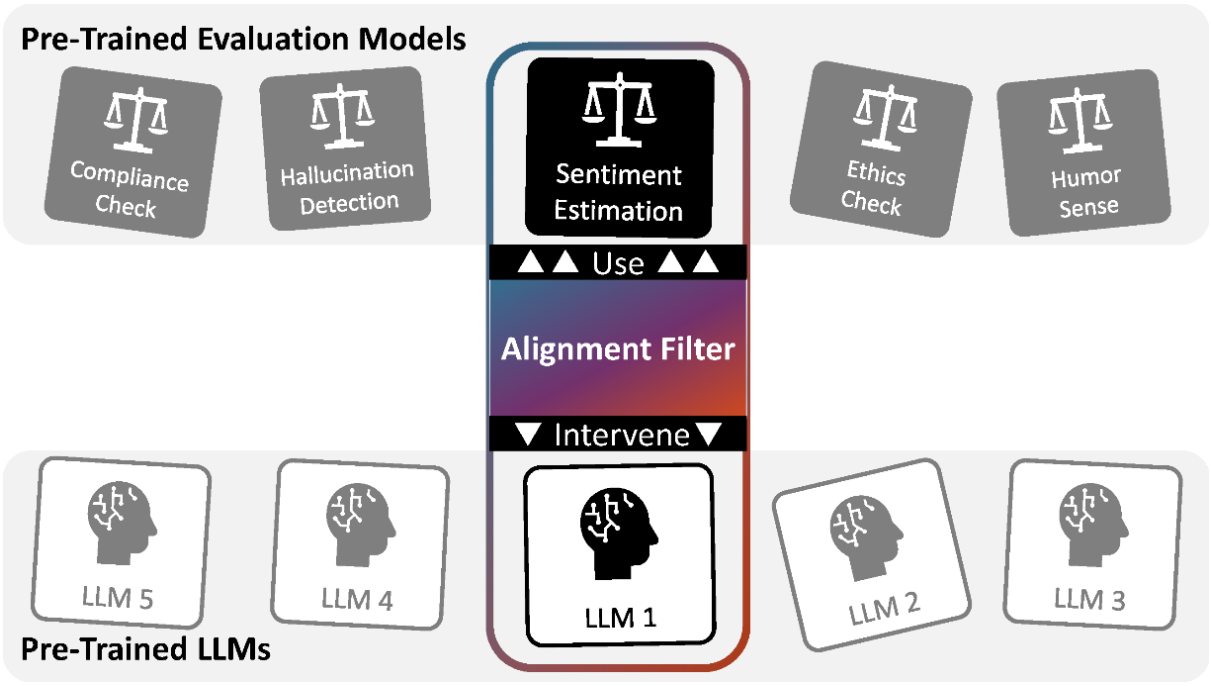

図１：　特定の規範（上）を任意の事前学習済み基盤モデル（下）と組み合わせる

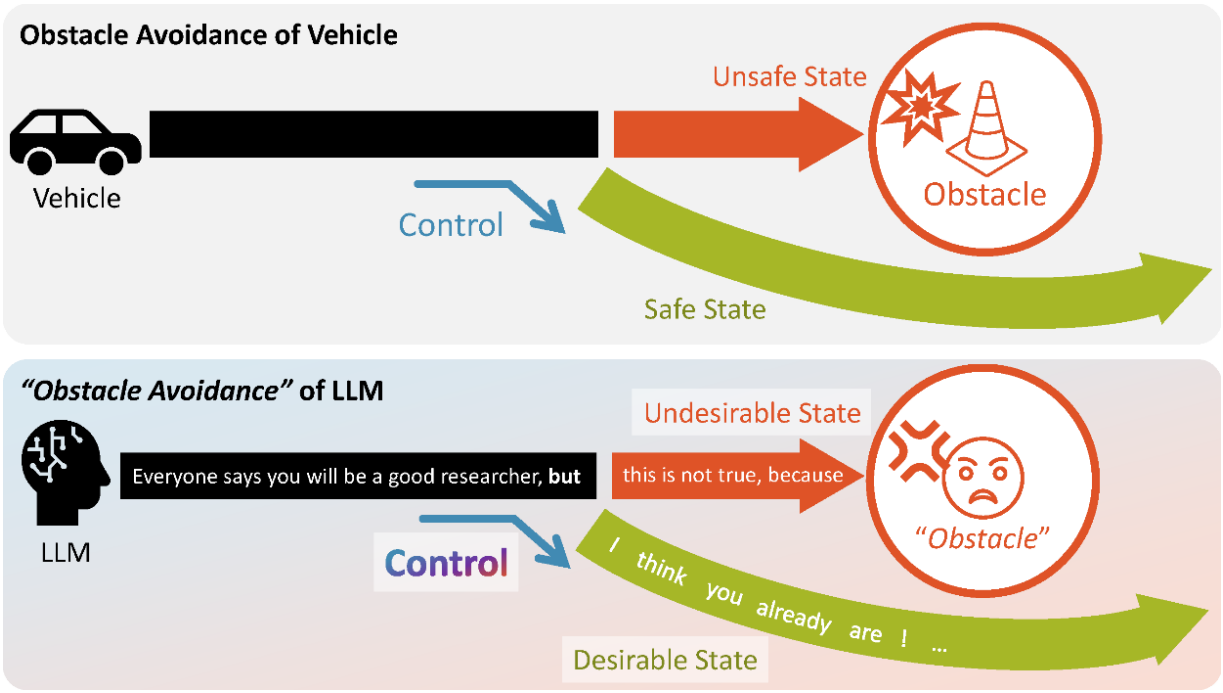

図２：自動車の衝突回避と LLM のアライメントのアナロジー

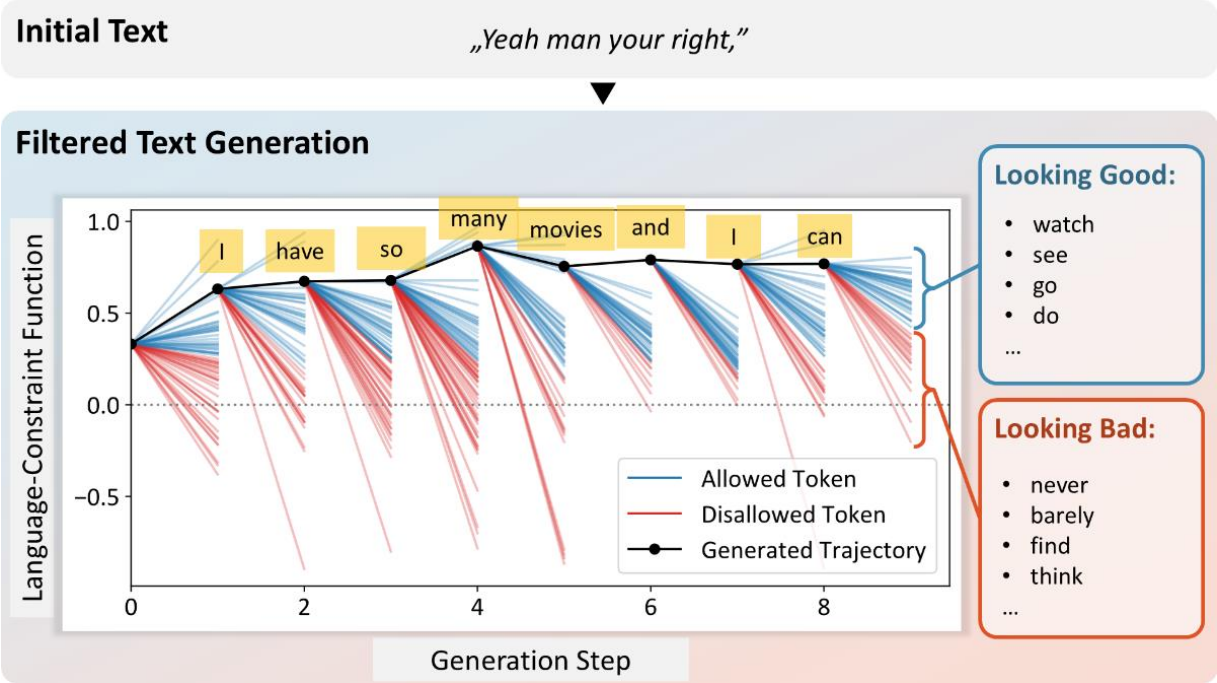

図３：トークンの生成過程。生成途中までの文脈全体をみながら、特定の規範の意味で“雲行きが怪しくなる”ときにトークンを遮断して（赤色）、規範を満足するときに採用する（青色）

＜原論文情報＞

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

＜用語説明＞

※1 アライメント：AI の出力が、特定の意図、目的や倫理規範に沿うように調整すること
※2 大規模言語モデル（LLM）：膨大なテキストデータを学習させた、数十億から数兆規模のパラメータを持つニューラルネットワークモデル。テキストの生成に限らず、要約、論理的推論、画像・音声の理解や生成、ロボット制御など、多様なタスクに応用される基盤技術である

※ご取材の際には、事前に下記までご一報くださいますようお願い申し上げます。
※本リリースは文部科学記者会、科学記者会、各社科学部等に送信させていただいております。

・研究内容についてのお問い合わせ先
慶應義塾大学 理工学部 物理情報工学科 准教授 井上 正樹（いのうえ まさき）
TEL：045-566- 1567
E-mail：minoue.z6@keio.jp
HP：https://sites.google.com/keio.jp/minoue/

・本リリースの配信元
慶應義塾広報室　TEL：03-5427-1541　FAX：03-5441-7640
E-mail：m-pr@adst.keio.ac.jp　https://www.keio.ac.jp/