AIは人間を騙す方法を学習した。最新の研究論文で明らかに

S22161380　小林　龍生

概要

- 新しい研究論文で、さまざまなAIシステムが『騙す方法』を学習していることが判明した。
- 研究によると、AIが人間を『体系的に誤った思い込みに誘導』して騙しているという。
- これは詐欺行為から選挙の偽情報に至るまで社会にリスクをもたらすものだ。

メタのCICEROは「嘘の達人」

ボードゲーム「ディプロマシー（Diplomacy）」をプレイするために開発されたAIシステム

メタは、CICEROを「話す相手に対し、おおむね正直で役に立つ」ようにトレーニングしたと述べている。だがこの研究では、CICEROは「嘘の達人」であることが判明したという。CICEROは守るつもりのない約束をし、同盟国を裏切り、明らかな嘘をついていたのだ。

- AIに倫理観という制約を課さねばならないと立証した意義は大きいです。例えば、商品購入相談アシスタントAIで、購入を促す際に騙しがあればクリティカルな問題です。
- AIが『人間を騙す意思を持って』いるならそうかもしれないけど、誤った情報を出しているだけなら、子どもが誤った思い込みで話しているのと大差ない。
  騙す、というのは意思を伴って初めて成立すると思います。

- 人がAIに支配される可能性が少し上がったかなと思いました。
- 生成AIの使い方について色々いわれているが、これからは正しい情報かどうだけではなく、情報自体がAIに誘導されたものなのではないかと疑うことも大切になってくると思いました。
- AIの性能がどんどん向上しているため、これからの生活にどう関わってくるのかが楽しみです。