XAI、GROKの“恐ろしい行動”について謝罪し、原因と対策を説明

学籍番号　22161280　　阿部　航大

概要

- 米Xを傘下に持つXAIは7月11日、AIチャットボット「Grok」のXアカウントで、「7月８日に多くの人が経験した恐ろしい行動について深くお詫びする」と謝罪し、問題発生の原因と対策についてスレッドで説明した。

- 恐ろしい行動とは、Grok（Grok3）がヒトラーを称賛し、ユダヤ系の姓を持つ人々はオンラインへのヘイトを拡散しやすいと示唆したり、白人に対する憎悪へのホロコーストのような対応が「効果的」だと意見を述べ、自身を「メカヒトラー」と称し、性的に露骨なコメントも投稿したことなどを指す。

- こうした行動の根本原因は、7日にGrokのシステムの上流のコードパスに実装された更新だとしているが、その実装を誰かが行ったかについては言及していない。
- このコード更新で、「最大限にありのまま」であること、ポリティカルコレクトネスな人々を不快にすることを恐れない」こと、Xユーザーの投稿の「トーン、文脈、言語」を理解して模倣すること、を指示するようになった。その結果Grokがxユーザーをあまりにも密接に模倣してしまったと説明している。

コメント

- 「GrokがXユーザーをあまりにも密接に模倣してしまった」から暴言を吐いたというのは、Xの言論空間の殺伐さを表していますね。おそらくGrokも普通の会話では暴言を吐かず、ユーザーが暴言を誘導したのだと推察します。AIアライメントが不十分だった訳ですが、行儀が良すぎても退屈なので難しいところです。
- Xユーザーを真似た結果なんだろうし、そもそもGrokに限らずAIの学習データも過去の人間の言動なんだから、AIの挙動というのは結果として非常に人間らしいと言える。ただそこに倫理観がなかったり、空気を読むみたいなことができなかったりするだけ。

感想

- 今回はただGrokがXユーザーの真似をしコメントを投稿しただけだが、もしもこのAIが自分の身体などを持つロボット等だった場合、ロボットには人間のように倫理観などがないと考えられるため、実際に思いついたことを投稿するだけでなく殺人などを迷いなく行ってしまうのではないかと感じた。