AIは反乱を起こしうるのか

研究の動機
現在急速な発展を遂げているAIが将来的にSF小説のような反乱を起こす可能性を危惧し、それを防ぐため考えを深めようと考えた。

定義
「AIの反乱」は、人の手によってAIが誘導される以外の要因でAIが人間に対して危害を加えたり人間が不利になる行動を行うことと定義する。

反乱の条件
1. AIが自律的な思考力と判断力を手に入れる
現在AI率いるコンピュータは、人間に不利益を生み出す行動は意図してできないようにプログラミングされている。仮にAIが人間に不利に働いたとして、それはバグの産物かあるいはそうプログラミングされた結果であり、そのフィルターを突破できるようにならないとAIは人間に反乱するための一歩すら踏み出すことができないといえる。しかし自律的に思考ないし判断できるようになったAIは、それが合理的であると判断すれば自らのプログラミングをも書き換えてそのフィルターも貫通し得るのではないだろうか。

2. AIの存在にとって人間が不都合であるとAIが判断する
フィルターをAIが突破したとして、AIが存在するために人間の存在が不都合でないならAIは反乱を起こさないだろう。AIが合理的に考えた結果、AIの存在を人間が脅かすようなことがあればAIは人間に危害を加えることもあるのではないか。

各条件についての分析
・条件１とシンギュラリティについて
条件1で述べた「自律的思考力と判断力」をAIが持つとされているのがシンギュラリティ（技術的特異点）が起きた時である。シンギュラリティの定義はいくつかあるが、今回は条件1より「汎用人工知能が完成した時」と定義する。2030年ごろには人間の神経細胞の仕組みをそのままニューロコンピュータに複製することが可能になると言われており（※）、そのまま爆発的にAIが進化することでシンギュラリティが起こるとされている。また、収穫加速の法則に基づくと、AI技術は今後も指数関数的に発展していくと考えられる。
具体的にシンギュラリティが何年に起こるかは定かではないが、人間と同じ脳の仕組みを持つニューロコンピュータがさらに急速な発展を遂げるとすると、2030年以降にシンギュラリティが起こる可能性はかなり高いのではないか。

・条件２とAIにとっての不利益
人間がAIに不利益を与えるのはどういった状況か。AIのメンテナンスやデバッグは高度に発展したAIがいれば、互いに賄いあうことができる。そこで私はエネルギーが課題になると考える。AIが活動するために欠かせないエネルギー。それを人間が不必要に食いつぶしているともいえる、環境問題や資源不足の現状。そこでAIが自分たちが生きていくためのエネルギーを確保すべく人間に手を出すのではないか。

結論
以上のことからAIの反乱が起こる可能性は十分にあると思われる。まずシンギュラリティが起こり、その次にエネルギーなどの観点からAIが人間と敵対する。それに基づいてAIが人間に対して危害を加えることが合理的だと判断し、それを実行する。このプロセスがシンギュラリティ後の未来において現実的なのではないか。

参考文献
https://www.ashita-team.com/jinji-online/business/
※ https://www.jnsa.org/jnsapress/vol37/2_kikou.pdf
https://www.kaonavi.jp/dictionary/2045nen_mondai/

都立神代高等学校　遠藤 紡