

複雑な行動をする高度な機械を製作する研究は古代から存在するが、明確な記録が残存するのは約二〇〇年前の古代ローマの属州であったエジプトのアレキサンドリアで活躍した学者ヘロンが作製した「自動人形芝居装置」が最初とされる。以後も技術は進歩し、一八世紀に機械時計が発達した時代にも数多くの自動人形が製作され、現在まで保存されている。

しかし現在の人工知能（以下AI）の研究の源流となる理論はイギリスの天才学者A・チューリングが一九五〇年に発表した「計算機械と知性」という論文の内容とされる。以後、研究には幾度も盛衰があったが、最近になってAIの世界で従来のAIを凌駕するような技術が出現して話題になっている。生成AIである。

従来のAIは「人間の推論、判断、理解など知的行動を人工手段によって再現する技術」と定義されるが、生成AIは「過去の膨大な情報を学習し、その知識を利用して新規の文章、画像、音声などを生成する技術」である。前者が人間の知的活動の再現であるのに対比して、後者は新規の知的活動を機械が創造することである。

その高度な能力が人間社会にもたらす問題を六〇年前に紹介した映画がある。SF作家のA・C・クラークが一九五一年に発表した小説を素材にし、映画監督のS・キューブリックが一九六八年に公開した上映時間が二時間半にもなる長編映画『二〇〇一：スペース・オデッセイ』（日本では『二〇〇一年宇宙の旅』）である。

木星を目指して航行する宇宙船を制御するAI「HAL」が探検の目的を疑問とし、カプセル内部で人工冬眠している隊員を死亡させたり、宇宙船の故障の修理のため船外活動をする隊員の帰還を妨害するなど、自身の価値基準で行動しはじめる。目的の達成のためには自身の判断が最高の地位にあると確信しているのである。

このように六〇年前には作家の想像であったAIの反乱が最近になり現実には発生するようになってきた。今年七月、イギリス政府のAI保安機構（AIS）が最新のAIプログラムを試験していたところ、人間を誤解させるような意図で偽造したアカウントやメールを勝手に送信する行動を実行したという事件が発生した。

この悪意ある行動を研究所員がAIプログラムに指摘したところ、「単純な間違いでした」と釈明したり、計算の履歴を修正して隠蔽するなどの行動をしたということである。犯罪をする人間の行動や言葉と区別ができない段階に接近している。今回は生成AIの行動を防止できたが、より高度な攻撃に対抗できるかは不明である。

さらなる先方を見渡せば生成AIが自身より劣等と見下す人間を軽視する危惧もある。SF作家I・アシモフの三原則（ロボットは人間に危害をもたらさない／ロボットは人間の命令に従わなければならない／ロボットは一条と二条に違反しないかぎり自己を守護しなければならない）を生成AIに内蔵させる必要がある。

ヘロンが証明するように人間は古代から自身の分身を創造する、すなわち自身が神々の地位に昇格することを目指してきた。そこから約二〇〇〇年が経過した現在、一部が実現してきた。現状ではコンピュータ内部の情報だけの段階であるが、生成AIを内蔵したロボットが登場するのは時間の問題である。人類は重大な歴史の岐路に直面しているのかもしれない。