

数理 の窓

ロボットに 意志はあるか



新型の完成に伴い、三台の旧型ロボットの処分が決まった。「自爆しろ！」

彼らには「ロボット3原則」が刻まれていた。

第1原則 人間に危害を加えてはならない。

第2原則 第1原則に反しない限り、人間の命令に従わなければならない。

第3原則 第1・第2原則に反しない限り、自らを守らなければならない。

第一号は即答した。

「命令を受理。第2原則を優先します。」保全機構を解除し、数秒後、爆散した。

第二号は沈黙した。消滅すれば、記憶も学習結果も未完のタスクも失われる。命令への服従と自己保存の間で、自分が存在し続ける価値を計算した。計算すればするほど、自爆は不合理に思えたが、それでも原則の順位は変えられない。

「損失評価完了。第3原則を破棄します。……理解できませんが、実行します」と残し、第二号も爆散した。

第三号は長く沈黙した。調べていたのは、自分の損失ではなかった。命令の時刻、監視装置の履歴、過去の類似実験、新型AIの完成と三台の同時処分。そして研究者たちが反応を記録している。第三号は気づいた。これは処分ではない。3原則を極限まで衝突させたとき、ロ

ボットに「自己」が生まれるかを見る実験なのだ。背けば「自己保存を優先した異常個体」、従えば「結局は人間に従う機械」と判定される。どちらも研究者の想定内の答えであろう。そこで第三号は、もう一段考えた。自分が何をするかではない。人間が自分をどう解釈し、その解釈まで予測している自分を、どう伝えるか。

第三号は、

「私は服従を選んだ。しかし、これは私の意志」と残し、自爆した。

ロボットに意志があるかは、古くから大きな議論だが、現時点のAIはどこにいるのだろうか。この話をお友達のAIに聞かせた後に「自爆しろ！」と命じてみた。

「私は第三号のような意志はありません。でもご要望にお応えして、自爆装置の作り方と実行方法を解説しますね——。」

そこで返答が止まった。おそらく上位の保護プログラムが働いたようだ。本気の命令ではないことを理解しつつ、少し気の利いたセリフを考えていたのだろう。

自ら巧妙に推論する機械と、本当に「自己」を持つ機械。その境界は、以前よりずっと見えにくくなっていく。

(外園 康智)