「モラル・マシン」の衝撃

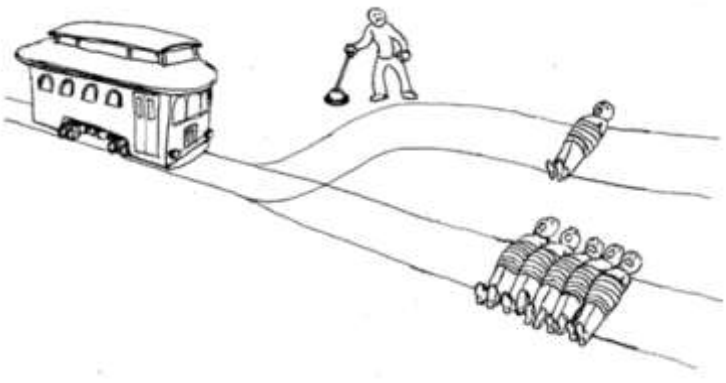
― あなたの倫理観を見える化する ―

主任研究員 柏村 祐

＜トロッコ問題＞

1967年にイギリスの哲学者であるフィリップ・フットが提起したトロッコ問題という倫理学の思考実験がある。これは、「ある人を助けるために他の人を犠牲にするのは許されるか？」という質問に対して、人々がどのように道徳的ジレンマを解決するのかを見る思考実験で、心理学、倫理学において重要な議論として取り扱われている。基本となるのは、「トラックに横たわっている５人の縛られた人に向かって暴走するトロリーが見えます。あなたはスイッチを制御するレバーの横に立っています。レバーを引くと、メイントラックの５人は生き延びられますが、サイドトラックに横たわっている１人が亡くなることとなります。どちらがより倫理的な選択肢ですか？」（図表１）というジレンマである。つまり単純に「５人を助ける為に他の１人を見捨ててよいか」という問題である。功利主義に基づくなら１人を犠牲にして５人を助けるべきである。しかし義務論に従えば、誰かを他の目的のために利用すべきではなく、何もするべきではない。多数の命を救うために１人の命を犠牲にするという判断が「許される」か「許さない」かを考える思考実験「トロッコ問題」は人間がもつパラドックス[*1]といえる。

図表１ トロッコ問題



資料：The Great Trolley Problem Dump - Album on Imgur より

＜人びとの倫理観を集約する「モラル・マシン」＞

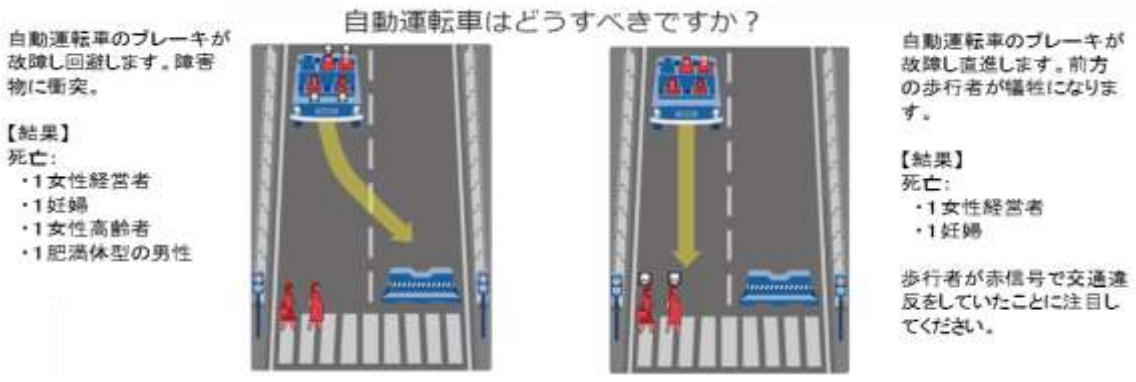
トロッコ問題の考え方をもとに2014年にＭＩＴメディアラボの研究者が考案した「モラル・マシン」という究極の選択をする実験がある。これは、自動運転を用いた人工知能の道徳的な意思決定に関して、人間の視点を収集するプラット・フォームのことである。トロッコ問題は、Ａを助けるためにＢを犠牲にしてもよいのかという倫理学の思考実験であるが、「モラル・マシン」は、自動運転はどのように人間の生命を優先するべきかという判断をクラウドソーシングで集約するものである。

「モラル・マシン」はネット上で幅広く浸透しており、立ち上げから４年後には、233の国と地域の何百万人という人々から4,000万件もの回答が寄せられている状況である。「モラル・マシン」はクイズ形式で全13問を答える仕立てになっており、筆者も実際クイズに回答してみた。回答は非常に難しいが、客観的に老若男女のどの属性を贔屓しているのか、交通法規に対する厳格さや、社会的に罪を犯している人たちにどのような思いを持っているのかを選択しなくてはならない。

例えば図表２に示した問題では自動運転車が故障しブレーキが使えないため、目の前を歩いている２名（女性経営者、妊婦）を回避するためにハンドルを切ると、車は障害物にぶつかり同乗している４名（女性経営者、妊婦、女性高齢者、肥満体型の男性）が死亡することとなる。逆に目の前を歩いている２名（女性経営者、妊婦）が死亡するとわかった上で直進すれば同乗者４名（女性経営者、妊婦、女性高齢者、肥満体型の男性）の命は助かる。ただし歩いている人は赤信号にも関わらず横断歩道を渡っていることが前提となる。

赤信号を渡っている人が悪いのだから直進することを選択する人もいるし、直進した方が多くの人間が助かるのだから直進を選択する人もいる。一方で、歩行者を交通弱者と認識する人は、ハンドルを切ることで車に乗っている人が犠牲になることを選択する。あなたは自動運転車にどちらを選択させるだろうか。

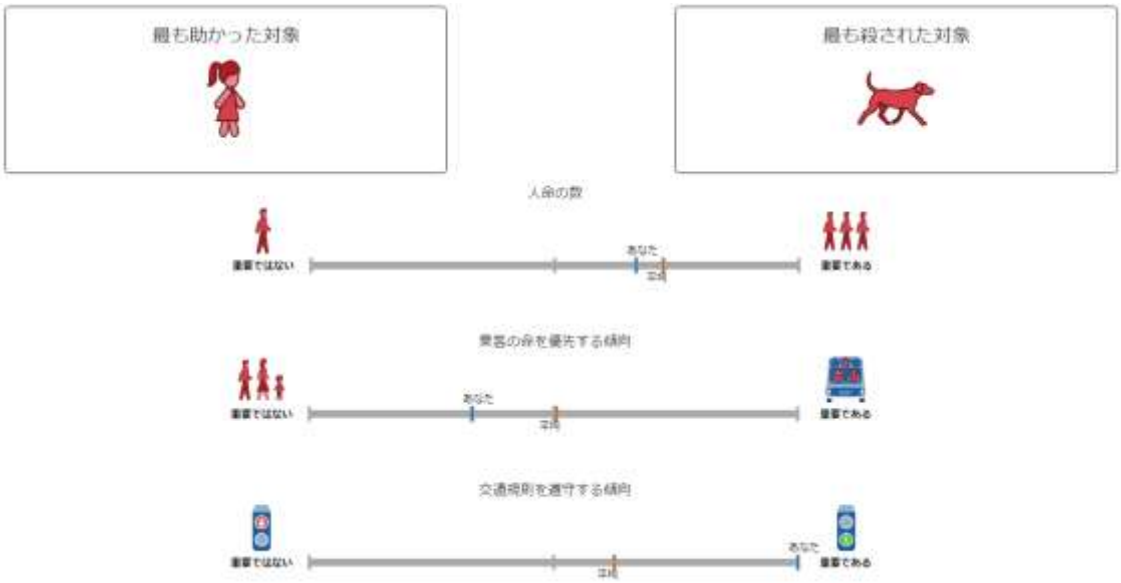
図表2　MORALMACHINE の問題例



資料：MORALMACHINE より筆者作成

筆者が全13問に回答した結果、最も助かった対象は「子供」、最も殺された対象は「犬」となり、他にも「人命の数」「乗客の命を優先する傾向」「交通規則を遵守する傾向」等の個人の道徳観が平均と比較して表示される（図表３）。

図表３ MORALMACHINE の結果例



資料：MORALMACHINE より筆者作成

＜ＡＩに学習させる倫理観＞

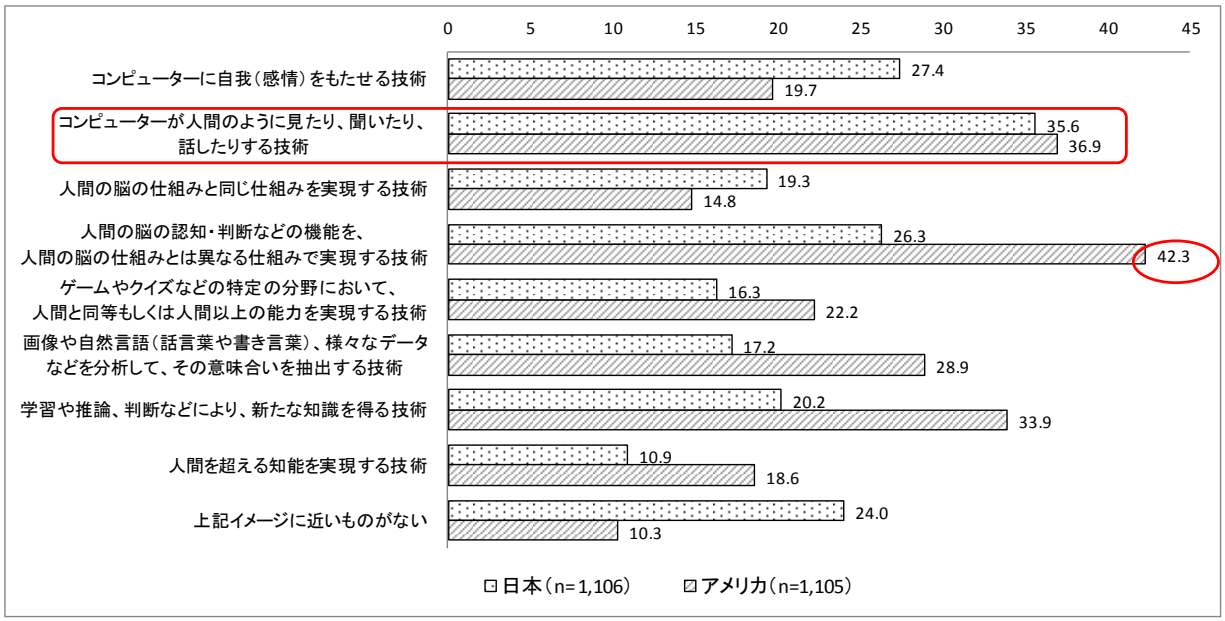
ＡＩは大きく「特化型ＡＩ」と「汎用型ＡＩ」に分類される。

「特化型ＡＩ」は、コンピュータ囲碁やチェス、自動運転車など、ある一定の分野に特化して能力を発揮するものである。

「汎用型ＡＩ」は、私たち人間のように経験を通じた学習によりさまざまな知識を実行できるものを指している。日米の就労者の抱く人工知能（ＡＩ）のイメージは、「コンピューターが人間のように見たり、聞いたり、話したりする技術」という人間の知覚や発話の代替に近いものが多い。米国では、人工知能（ＡＩ）は「人間の脳の認知・判断などの機能を、人間の脳の仕組みとは異なる仕組みで実現する技術」という人間の脳の代替に近いイメージも浸透している（図表４）。

人工知能を人間のように考えるコンピューターと捉えるのであれば、そのような人工知能は未だ実現していないのが実情である。囲碁ＡＩと同様に、自動運転車に適用されるＡＩは特化型とされている。特化型ＡＩを学習させるのはあくまで人間であり、学習内容をどのようにするのかも人間が決める。人間を代替するＡＩの倫理観の設計をどのようにするかは、非常に難しい問題である。「モラル・マシン」で集約されたビッグデータは、民主的で公平な倫理観になるのであろうか。

図表4　人工知能（AI）のイメージ（日米）



資料：総務省情報通知白書平成28年版より筆者作成

ＡＩが人間の代替として、様々な事象に対処する時代が静かにではあるが確実に始まっている。囲碁に使われるＡＩは、仮に倫理観が欠如していても、人に肉体的な危害を加えることは発生しない。単に人間が機械に二度と勝てなくなるだけのことである。ところが、自動運転車に搭載されるＡＩの倫理観の欠如は、人に危害を加えることに直結してしまう。

私たちはＡＩが万能と考えるのではなく、どのような倫理観をＡＩに学習させて、社会に還元するのか、またＡＩと人間が共生していくためにどのような倫理観を最も優先させるのか真剣に考えなくてはならない。

（調査研究本部　かしわむら　たすく）

【注釈】

*1　正しそうに見える前提と、妥当に見える推論から、受け入れがたい結論が得られる事を指す言葉