SENSOR
No. 035
2017/03
東京海上研究所ニュースレター

トピックス

# AIの安全ガイドライン「アシロマ AI 23原則」

発達した人工知能はやがて人間の存在を邪魔に思い始め、人類を絶滅させてしまうのではないかと心配する声があります。一方、「ロボット3原則」のような原則を作って全ての人工知能に守らせれば大丈夫という考えもあります。2017年2月に、このロボット3原則の拡張版ともいえる「アシロマ AI 23原則」が発表されました。

2017年1月、カリフォルニア州アシロマに、全世界からAIの研究者と経済学、法律、倫理、哲学の専門家が集まり、「人類にとって有益なAIとは何か」を5日間にわたって議論しました。
その成果として2017年2月3日に発表されたのが、「アシロマ AI 23原則」(Asilomar AI Principles)です。この原則は、AIの研究、倫理・価値観、将来的な問題の3つの分野に関して、研究開発のあり方、安全基準の遵守、透明性の確保、軍拡競争の防止、プライバシーと人格の尊重など、幅広い視点からの提言がなされています。
強制力こそないものの、この原則には多くの支持者がついており、物理学者のスティーブン・ホーキング博士、SpaceXやTeslaのCEOであるイーロン・マスク氏や、シンギュラリティ大学の創設者のレイ・カーツワイル博士などの著名人も支持者に名を連ねています。

(図1)アシロマ会議の参加者。出典:Future of Life Institute

以下がその翻訳です。

アシロマ AI 23原則

| 1. 研究 |
|---|
| (1) 研究目標:AI研究の目標は、方向性の定まらない知能ではなく、有益な知能の創造である。 |
| (2) 研究資金:AIへの投資には、AIの有益な利用のために必要な、コンピューターサイエンス、経済、法律、倫理、社会学などに関する、以下のような厄介な問題に関する研究費用も含めるべきである。<br>● AIシステムを高度に堅牢にし、誤動作やハッキングをされることなく我々の望む働きをさせるためにはどうすればよいか?<br>● 人々の資産や生きがいを損なうことなく、自動化によって人類が繁栄するためにはどうすればよいか?<br>● AIの進化に合わせ、AIによるリスクを管理するために、より公平で効果的になるよう法的システムを刷新するためにはどうすればよいか?<br>● AIはどのような価値観と結びつけられ、どのような法的・倫理的地位を与えられるべきか? |
| (3) 科学と政策のリンク:AI研究者と政策立案者の間で、建設的で健全な意見交換が行なわれるべきである。 |
| (4) 研究文化:AI研究者や開発者の間で、協力、信頼、透明性の文化が育まれるべきである。 |
| (5) 競争の回避:AIシステムの開発チーム同士は、競争のために安全基準を省略することがないよう、積極的に協力しあうべきである。 |

| 2. 倫理と価値観 |
|---|
| (6) 安全性:AIシステムはその運用期間を通して、可能な限り安全で堅牢で、検証可能でなければならない。 |
| (7) 障害の透明性:AIシステムが障害を起こしたときは、その原因を確認できるようにするべきである。 |
| (8) 法的透明性:自動システムが法的判断に関わる場合、有能かつ権限を持つ人間が監査し、納得のいく説明ができるようにする。 |
| (9) 責任:先進的なAIシステムの設計者と開発者は、システムの使用、悪用、結果に倫理的な関わりがある当事者であり、その関わりを形作る責任と機会がある。 |
| (10) 価値観の一致:高度に自律的なAIシステムは、目標と行動が倫理的に人間の価値観と一致するようデザインされるべきである。 |
| (11) 人間の価値:AIシステムは、人間の尊厳、権利、自由そして文化的多様性に適合するよう設計・運用されなければならない。 |
| (12) 個人のプライバシー:AI システムが個人のデータを分析し、利用する力を持つ以上、データを生成する個人は自らのデータを閲覧、管理、コントロールする権利が与えられなければいけない。 |
| (13) 自由とプライバシー:AIによる個人情報の利用は、人間が持つ、あるいは持つと思われている自由を不合理に侵害してはならない。 |
| (14) 利益の共有:AI技術は、可能な限り多くの人々に利益と力をもたらすべきである。 |
| (15) 繁栄の共有:AIによって作られた経済的な繁栄は、人類すべてに利益をもたらすために、幅広く共有されなければならない。 |
| (16) 人間によるコントロール:人間が選択した目標を達成するために、AI システムに決定をどのように委ねるのか、あるいは委ねるか否かは、人間が判断しなければいけない。 |
| (17) 転覆活動の防止:高度な AI システムをコントロールすることによって得られる力は、健全な社会に不可欠な社会的、市民的プロセスを転覆させるのではなく、尊重、促進するために使われなければならない。 |
| (18) AIの軍拡競争:破壊的な自動兵器による軍備の拡大競争が起きてはならない。 |

| 3. 将来の問題 |
|---|
| (19) AIの能力:意見の一致がない以上、将来のAIの能力の上限に関する強い前提を置くことは避けるべきである。 |
| (20) 重要性:発達したAIは地球上の生命の歴史に重大な変化を及ぼす可能性があるので、相応の注意と資源をもって計画、管理されなければならない。 |
| (21) リスク:AI システムによるリスク、特に壊滅的なものや存亡の危機をもたらすものに対しては、その影響に相応した慎重な計画と緩和対策を行うべきである。 |
| (22) 再帰的自己進化:自己進化、または自己複製によって質的・量的に急激に拡大をもたらすよう設計されたAIは、厳格な安全管理対策の対象とすべきである。 |
| (23) 共通の利益:超知能は、特定の国や組織のためではなく、広く共有されている倫理的な理想や、人類全ての利益に資するためにのみ開発されるべきである。 |

（出典:Future of Life Institute　翻訳:東京海上研究所）

この原則のポイントを5つ挙げるとしたら、

- むやみに技術開発を競うのではなく、今開発しているAIは人類全体にとって本当に有益かを考える
- AIの目標と行動は、人間の倫理観・価値観と一致するようデザインしなければいけない
- AIによってもたらされる経済的利益は、全世界で広く共有されなければいけない
- AIによって、人間の尊厳、権利、自由、文化的多様性が損なわれてはいけない
- 自己増殖機能を持つようなAIの開発には、厳重な安全管理対策が必要である

といったところでしょうか。
いずれの原則も非常に大切ですが、全ての国と企業、研究者がこの原則を遵守するようになるかというと、疑問視する声もあります。しかし、私たちもAIを使った製品・サービスを開発するにあたっては、少なくともこの原則を頭に入れておく必要はあるでしょう。

【参考:アシモフの「ロボット3原則」】

第一条:ロボットは人間に危害を加えてはならない。また、その危険を看過することによって、人間に危害を及ぼしてはならない。
第二条:ロボットは人間にあたえられた命令に服従しなければならない。ただし、あたえられた命令が、第一条に反する場合は、この限りでない。
第三条:ロボットは、前掲第一条および第二条に反するおそれのないかぎり、自己をまもらなければならない。

（出典:アイザック・アシモフ「われはロボット」）

なお、アシモフはこの3原則を「完璧なもの」として書いたのではなく、小説の中ではその不完全さゆえにロボットが一見不可解な行動を取り、その謎解きが作品の面白さにもなっています。