第１２講　教師あり学習を用いたＡＩ倫理

【学習到達目標】

・ＡＩ倫理の定義と背景を説明できる。
・ＧＩＧＡスクール構想でＡＩ倫理チャットボットが必要になった理由を説明できる。
・ＡＩ倫理処理で用いる「教師あり学習」について説明できる。

１．ＡＩ倫理とは

現在の第４次ＡＩブームでは、自動運転や画像診断など私たちの暮らしにＡＩ技術が急速に入り込んできています。２１世紀の基幹テクノロジーとされるＡＩとどう付き合い、その活用をどこまで許容していくのか？ＥＵではＡＩ倫理に基づく輸入規制を計画しており、日本のＡＩ倫理が問われています。

・ＡＩ倫理の歴史

ＡＩ倫理の歴史は、ＡＩ技術の発展とともに進化してきました。以下は、その主要な出来事や転換点を解説します。

１）初期のＡＩ倫理の概念（１９５０～１９６０年代）

アラン・チューリングによる「チューリングテスト」（１９５０年）は、「機械が人間のように考えることができるか？」という問いを投げかけ、ＡＩにおける倫理的な議論の発端となりました。

アイザック・アシモフの「ロボット工学三原則」（１９４２年）は、フィクションの領域でしたが、ＡＩの安全性と倫理に関する概念を初めて提示しました。

第一法則：ロボットは人間に危害を加えてはならない。
第二法則：ロボットは人間の命令に従わなければならない。
第三法則：ロボットは自己を守らなければならない。

２）ＡＩ技術の発展と懸念の拡大（１９７０～１９９０年代）

ＡＩの初期段階で、多くの研究者は「ＡＩが倫理的な問題を引き起こすか」という問いに関心を持ち始めました。

１９８０年代：エキスパートシステムの導入により、ＡＩが医療や金融などの重要な意思決定に関与するようになりました。これにより、誤った判断やアルゴリズムによるバイアスの問題が浮上しました。

１９９０年代：ＡＩ技術がインターネットで普及し、プライバシー侵害や監視のリスクに

対する懸念が増大しました。

3）倫理的ガイドラインの登場（２０００～２０１０年代）

２０００年代：ＡＩが社会生活に深く浸透するに伴い、政府や学術機関が倫理ガイドラインの必要性を認識しました。ＩＥＥＥは倫理的ＡＩの開発に関する最初の国際的な指針を提案しました。

２０１６年：Ｇｏｏｇｌｅ、Ｍｉｃｒｏｓｏｆｔなどの大手テック企業が独自のＡＩ倫理委員会を設置し、ＡＩの開発と使用に関する原則を発表しました。ＧｏｏｇｌｅのＡＩ原則には「ＡＩは人々を傷つけるために使われてはならない」という考えが含まれました。

4）国際的なＡＩ倫理への取り組み（２０１０年代後半～２０２０年代初頭）

２０１８年：欧州連合（ＥＵ）が「ＡＩ倫理ガイドライン」の策定に着手し、ＡＩ開発における透明性、説明責任、公平性を重視する枠組みを構築しました。

２０１９年：ＯＥＣＤが国際的なＡＩ原則を採択し、ＡＩ技術の倫理的な開発・運用に関する指針を提供しました。

シンガポールやカナダも、ＡＩ倫理のフレームワークを設けた結果、公共および民間部門でのＡＩガバナンスを推進しました。

5）日米企業のＡＩ倫理政策の代表例

米国Ｍｉｃｒｏｓｏｆｔ（２０１７年）：ＭｉｃｒｏｓｏｆｔはＡＩ倫理委員会を設置し、透明性や公平性を担保する方針を定めました。ＡＩシステムの開発と利用に際してレビューを行うガイダンスを提供し、信頼性と安全性を重視する５原則の実践を進めています。

米国Ｇｏｏｇｌｅ（２０１８年６月）：ＧｏｏｇｌｅはＡＩ倫理を策定し、「ＡＩと私たちの社会における役割」を強調しました。企業全体でＡＩの利用に関するガイドラインを明確にしています。

米国ＩＢＭ（２０１８年９月）：ＩＢＭは「Ｅｖｅｒｙｄａｙ Ｅｔｈｉｃｓ ｆｏｒ Ａｒｔｉｆｉｃｉａｌ Ｉｎｔｅｌｌｉｇｅｎｃｅ」を発表し、ＡＩ導入における透明性と公正な判断の重要性を訴えました。問題が発生した場合、フィードバック体制を整備するなどの実践例も示しています。

ソニーグループ（２０１８年９月）ソニーは「ＡＩ倫理ガイドライン」を策定し、2019 年 4 月からの実践に向け、安全性審査を始めました。

富士通（２０１９年３月）：富士通は「ＡＩコミットメント」を発表し、ＡＩの責任ある開発を推進する方針を明確にしました。

ＮＥＣ（２０１９年４月）：ＮＥＣはグループ全体でＡＩ倫理の取り組みを進め、ＡＩコンソーシアムのガイドラインを参照して実践しています。

ＮＴＴデータ（２０１９年５月）：ＮＴＴデータはＡＩシステム開発の倫理ガイドを作成し、現場での具体的な作業指針を提供しつつ、倫理的な取り組みを促進しています。

日立製作所（2021 年２月）：日立はＡＩ倫理委員会を設置し、外部専門家も招いた議論を進め、グループ全体での倫理的なＡＩ運用を図っています。

Ｙａｈｏｏ！（２０２２年5月）：Ｙａｈｏｏ！はＡＩ倫理方針を発表し、グループ全体で倫理的なＡＩの整備と運用を推進しています。

パナソニック（２０２２年８月）：パナソニックはＡＩ倫理ルールを策定し、２０２２年度中の本格導入を目指しています。

このように、日米の企業は透明性、公平性、安全性の確保を目的にＡＩ倫理に関するガイドラインを策定し、組織内外での対応を進めています。多くの企業がレビュー体制や専門家の参加を重視し、ＡＩ技術の信頼性向上に努めている点が共通しています。

・日本のＡＩ倫理政策と経団連のＡＩ倫理ガイドライン

１）日本のＡＩ倫理政策の概要

日本政府はＡＩ技術の発展を支えるために、倫理的な指針を打ち出し、ＡＩの社会的受容と信頼構築を目指しています。

総務省、経済産業省、文部科学省を中心に、ＡＩ技術のガバナンスと社会実装に向けた施策を策定。２０１９年には「ＡＩ原則」を公表し、以下の項目を重点的に掲げました。

人間中心の原則：ＡＩは人間の幸福と利益を最優先する。

公平性：ＡＩによる差別や偏見を排除する。

透明性と説明責任：ＡＩシステムの設計者がその意思決定過程を説明できる。

プライバシー保護：個人データの保護を徹底する。

安全性とセキュリティ：ＡＩの悪用や誤用のリスクへの対応を行う。

また、ＡＩの利用にあたっては、企業や公共部門でのガイドラインに基づく適正な運用が求められます。

２）経団連のＡＩ倫理ガイドライン

経団連（日本経済団体連合会）は、企業がＡＩ技術を社会に適切に展開するための倫理的枠組みを提供するため、２０１９年４月に「ＡＩ活用の倫理原則」を発表しました。主な指針は以下のとおりです：

人間中心のＡＩ：ＡＩは人間の価値を尊重し、社会全体の福祉に貢献するよう設計されるべき。

プライバシーの保護：ＡＩが扱うデータに関しては、個人情報保護の法律を遵守。

公平性と多様性：アルゴリズムのバイアスを排除し、あらゆる人が平等に利益を享受できるよう配慮。

透明性と説明可能性：ＡＩの意思決定の仕組みが明確に説明できるようにすること。

安全と信頼の確保：ＡＩシステムの安全性を確保し、リスクが発生した際の対応策を準備。

このガイドラインは、日本企業が国際社会の中でＡＩ技術の信頼を高め、持続可能な成長を実現することを目的としています。また、企業が自発的にこれらの原則を実践することで、ＡＩの社会実装における倫理的な課題に対処することが期待されています。

日本政府と経団連のＡＩ倫理ガイドラインは、人間中心の設計、安全性の確保、透明性、そしてプライバシー保護を重視しています。政府は法的枠組みと教育プログラムを通じて

ＡＩの適正利用を促進し、経団連は企業が自主的に倫理的原則を実践することを推進しています。これらの取り組みは、ＡＩの社会的受容と国際競争力の向上に貢献しています。

それでは、世界各国のＡＩ倫理への取り組みを見てみます。

・米国の「ＡＩ権利章典」と企業の対応

米国は 2022 年に「ＡＩ権利章典」を発表し、ＡＩの設計・利用における５つの基本原則を提示しました。これには「安全かつ有効なシステム」「アルゴリズムの差別防止」「データのプライバシー」「告知と説明責任」、および「問題発生時の人間による代替対応」が含まれます。法的拘束力はないため、実効性に課題があると指摘されていますが、多様な政府機関がこれらの原則に基づきフレームワークを策定しています。

・英国のＡＩ規制のフレームワーク

英国政府はＡＩ規制のフレームワークを設計し、「効果的なＡＩ保証エコシステムのロードマップ」を策定しました。これには「ＡＩの使用における安全性」「透明性の確保」「説明責任」などが含まれます。ＡＩの発展と国際標準化への連携も重視し、民間および政府機関が連携して市場構築や標準化を推進しています。

・EU の「ＡＩ責任指令案」と「製造物責任指令の改正案」

EU は 2022 年に「ＡＩ責任指令案」と「製造物責任指令の改正案」を発表し、ＡＩシステムの開発者や提供者に対する責任を明確にしました。被害者救済の強化を目指し、因果関係の立証を簡易化しつつ、製品の長寿命化や改変に対応するための法改正を進めています。

・シンガポールの「モデルＡＩガバナンス枠組み」と「ＡＩ．Ｖｅｒｉｆｙ」

シンガポールは「モデルＡＩガバナンス枠組み」と「ＡＩ．Ｖｅｒｉｆｙ」という検証ツールを発表しました。これにより、企業がＡＩガバナンスの実践状況を客観的に評価できるよう支援します。また、透明性や説明責任を重視し、ＭａｓｔｅｒＣａｒｄやＭｉｃｒｏｓｏｆｔなど多国籍企業との協力も進めています。

世界の国々は、安全性、透明性、説明責任といった共通課題に取り組む一方で、法的拘束力や技術標準化のアプローチに違いがあります。それぞれの政策は、自国の技術発展を支えつつ、国際的な連携も視野に入れて構築されています。

・ＡＩ倫理の定義

倫理とは、Ｗｅｂｓｔｅｒ辞書によれば「ａ ｓｙｓｔｅｍ ｏｆ ｍｏｒａｌ ｐｒｉｎｃｉｐｌｅ」となっており、ＡＩ倫理は「ａ ｓｙｓｔｅｍ ｏｆ ｍｏｒａｌ ｐｒｉｎｃｉｐｌｅ ｆｏｒ ｕｓｉｎｇ ＡＩ」と定義できます。ＡＩ倫理の研究の１つの目標は、日本の文部科学省が推進する全国の児童・生徒１人に１台のコンピュータと高速ネットワークを整備する「ＧＩＧＡスクール構想」に必要な「ＩｏＥ」（倫理のインターネット、教育のインターネット、生きる力のインターネット）ＡＩ倫理チャットボット機能を試作・検証することにあります。

注）**ＧＩＧＡ端末**：ＧＩＧＡスクール構想では１人１台端末を「ＧＩＧＡ端末」と称します。

・1人1台端末に必要なＡＩ倫理チャットボット

日本ＡＩ戦略の教育改革「1人1台端末」、ＧＩＧＡスクール構想の最大の課題は「チャットによるいじめ問題」と言われています。ＡＩ倫理を探究するうちに「チャットによるいじめ問題」にＡＩ倫理チャットボットが使えるのではないか？と考えています。

近年、人間と会話をすることができる対話システムへの注目が集まっています。例えば、音楽の再生やメールの確認などを行うＧｏｏｇｌｅ　ＡｓｓｉｓｔａｎｔやＳｉｒｉ、また顧客からの問い合わせ対応を代替するチャットボット（チャットのロボット）といった、何らかのタスク達成を目的としたタスク指向型対話システムが広く浸透してきています。

今日、自動運転や画像診断など私たちの暮らしにＡＩ技術が急速に入り込んできています。21 世紀の基幹テクノロジーとされるＡＩとどう付き合い、その活用をどこまで許容していくのか？「ＡＩ倫理」とでも呼ぶべき社会規範をきちんと議論しなくてはならないと言われています。

・哲学者クーケルベルク著「ＡＩ倫理」

ウィーン大学の哲学者クーケルベルク氏は著書「ＡＩ倫理」で、ＡＩを使うための「運転免許証」がないと警鐘を鳴らしています。

・ノーベル文学賞受賞作品「クララとお日さま」

ノーベル文学賞受賞者のカズオ・イシグロ氏は「クララとお日さま」という最新の小説の中で、人工知能を搭載したロボット「クララ」を登場させ、ＡＩ倫理の重要性を示唆しています。この小説の中で、クララは、観察と学習への意欲と理解力を持つに至り、人間社会で生きていく力「生きる力」（Ｅｎｅｒｇｙ　ｏｆ　Ｌｉｆｅ）を得るようになります。

・ＧＩＧＡスクール構想に関するアンケート調査結果

ＧＩＧＡスクール構想に関するアンケート調査結果によると、子供に GIGA 端末をどのように使わされて良いかわからないが６０％（１７３８件の調査、１２４件の解答）と分かりました。

この問題を解決するため、支援員やヘルプディスクを入れた学校が６８％で（１７４２件調査、９７件の回答）あった。一方、文部科学省や教育関係団体の調査で、ギガの最大のトラブルは、「チャットによる悪口」であると判明しました。

また、民間の調査機関によると、いじめの加害者または被害者になる子供が７２．６％に上ると報告されています。

２０２２年3月の日本の内閣府調査によるとインターネットの危険性について説明を受けたり学んだりした子供は８８％でした。その中でインターネット上のコミュニケーションに関する問題で悩んでいる子供が８０％（2，９８４ 回答）に上ることが分かりました。

2. 教師あり学習

・三つの学習の枠組み

機械に学習させる「機械学習」には「教師あり学習」、「教師なし学習」、と「強化学習」の三つの学習の枠組みがあります。人間の脳のニューロンが層状に接続した構造を模擬した機械学習の三つの枠組みがあります。

・教師あり学習

「教師あり学習」とは主に人間の小脳が担う学習機能で、代表的な統計手法は回帰と分類です。学習者に対し、教師が明示的に正解を教えたり、学習者の誤りを指摘したりすることで、学習者が正しい解を得ることを助けます。すなわち、正しい入出力の組合せを与えて学習することで、新規の入力に対し、適切に出力する。代表的な手法は誤差逆伝播法（Ｂａｃｋ　Ｐｒｏｐａｇａｔｉｏｎ）です。「分類」の手法として、正解、若しくは誤りを入力として、未経験入力に対する意志を決定する決定木（Ｄｅｃｉｓｉｏｎ　Ｔｒｅｅ）や決定表（Ｄｅｃｉｓｉｏｎ　Ｔａｂｌｅ）の作成などがあります。このＡＩ倫理の研究では、ＥＸＣＥＬ上の決定表でＡＩ倫理処理システムの「倫理表」を試作しました。

注）**教師なし学習**：主に、大脳皮質が担う学習機能です。統計的性質や、ある種の拘束条件により入力パターンを分類したり、抽象化したりする学習で、主成分分析、自己組織化マップなどの次元圧縮(Ｄｉｍｅｎｓｉｏｎａｌｉｔｙ　Ｃｏｍｐｒｅｓｓｉｏｎ)手法が代表例です。感覚情報などの入力パターンの分類、同様に出力運動パターンに対して統計的性質を用いて要素行動に分類する学習法などがあります。

注）**強化学習**：主に、大脳基底核が担う学習機能です。最終結果若しくは、途中経過に対して、どの程度良かったかを示す「報酬信号」に基づき、これらの報酬をなるべく大きくするように探索します。

強化学習と教師あり学習の違いは、フィードバックがスカラー(報酬の成否)かベクトル(正解の情報)かという説明もあるように、明示的な教師ではなく、環境などの非明示的な教師だという解釈もある。

3.「ＡＩ倫理」処理システムの試作

「教師あり学習」を使い、社会規範・倫理と、設計者の故意ではないＡＩの誤認識（機能不全、誤作動や機能低下を含む）を検証し適切な処理を行う「ＩｏＥ」（Ｉｎｔｅｒｎｅｔ　ｏｆ　Ｅｔｈｉｃｓ，　Ｉｎｔｅｒｎｅｔ　ｏｆ　Ｅｄｕｃａｔｉｏｎ，　Ｉｎｔｅｒｎｅｔ　ｏｆ　Ｅｎｅｒｇｙ　ｏｆ　Ｌｉｆｅ）ＡＩ倫理チャットボット機能の試作を行いました。

学習者に対し、教師が明示的に正解を教えたり、学習者の誤りを指摘したりすることで、学習者が正しい解を得ることを助ける。

すなわち、正しい入出力の組合せを与えて学習することで、新規の入力に対し、適切に出力する。

具体的なＡＩ倫理処理を見ると、「教師あり学習」を使い、教育禁止用語や放送禁止用語等のような社会規範・倫理とＡＩの誤認識が処理・説明できるシステム作りを目指しました。入力はＡＩ音声入力でもキーボード入力でもできます。

デープラーニングによるＡＩ音声入力はｉＰｈｏｎｅで行い、リモートマウスで接続したパソコン上でＡＩ倫理処理を行ました。「ＡＩ音声入力では何故誤認識したか？」は言葉では説明できない。つまり暗黙知です。

社会規範・倫理とＡＩの誤認識の検出・修正（言換え）処理はＶＢＡプログラムで瞬時に終了し、修正した音声入力文と修正理由を説明した説明文はそれぞれＥＸＣＥＬファイルに保存されます。

学習データは、社会規範・倫理例、ＡＩの誤認識と学習済みのＴｅｎｓｏｒＦｌｏｗ．ｊｓモデル・デーモン（システム）等で、インターネットとブロックチェーンで参照します。

例えば入力文「Slave is a bad word」（奴隷は良くない言葉です）を入力すると、正しい表現に言い換え、その理由を説明します。

社会規範・倫理例２の放送禁止用語は教育禁止用語としてウェブ検索すると出現します。

具体的にはアイヌ系からロンパリに始まりブスとかチビといった誹謗中傷の類からジョンやアメ公といった人種差別用語まで教育上使わない方が良いと考えられる用語は網羅されています。

４．まとめ

本章では、「教師あり学習」として倫理表や学習済みＴｅｎｓｏｒＦｌｏｗ．ｊｓモデルを使い、教育・放送禁止用語のような社会規範・倫理が検証処理・説明できるチャットボット機能を試作し、ＧＩＧＡ端末やケータイに人工知能を搭載した人間に親切な「ＩｏＥ」ＡＩ倫理チャットボット機能の有効性を実証しました。

特に、ディープラーニングを使用した「ＩｏＥ」ＡＩ倫理的チャットボット機能は、次章第１３章で説明しますが、３０％を超える誹謗中傷抽出率で、倫理テーブルで確認できなかったチャット内の未定義の誹謗中傷を検出することもでき、「ＩｏＥ」ＡＩ倫理チャットボット機能のプロトタイプの有効性が実証されました。

課　題

ＧＩＧＡスクール構想でＡＩ倫理チャットボットをどのように活用したら「いじめ」が減るかを考察し、あなたの考えを８００字以内で説明しなさい。