AI 時代の教育学会・2021 年度第一回研究会論文

一人一台端末に必要なIoE：AI倫理システム

IoE: AI ethics system needed for one device per person

学情研* 澤井 進*

＜抄 録＞

本研究は、インターネットから社会規範・倫理等のAI倫理を人工知能（AI）で解析する「IoE」（Internet of Ethics, Internet of Education, Internet of Energy of Life）AI倫理システムの実現を目指している。本研究ではセキュリティソフトと同様に、倫理的に悪質なコンテンツや、AIの誤作動等を防ぎ、人間中心で人間に親切な人工知能の親友「AF」と呼べるような「IoE」AI倫理システムを一人一台端末に常に搭載できないか検討・試作し、「教師あり学習」として倫理表や学習済みのTensorFlow.js モデルを使い、システムの有効性を検証した。

＜キーワード＞

AI倫理、IoE、人工知能、一人一台端末、教師あり学習、社会規範、倫理表、TensorFlow.js モデル

１ ＡＩ倫理

本研究は、セキュリティソフトと同様に、倫理的に悪質なコンテンツや、人工知能（ＡＩ）の誤作動等を防ぎ、人間中心で人間に親切な人工知能の親友「ＡＦ」(Artificial Intelligence Friend)と呼べるような「ＩｏＥ」ＡＩ倫理システムを一人一台端末に常に搭載できないか検討・試作し、「教師あり学習」として倫理表や学習済みの TensorFlow.js モデルを使い、システムの有効性を検証した。[1) 2) 3) 4)]

現在、一人一台端末にセキュリティソフトを入れないでオンライン授業等で端末を利用する人はいない。また、インターネットでの子どものトラブルの 72.6%が「ネット内いじめの加害者または被害者となった」という調査報告もある。[5) 6) 7)] 一人一台端末にウイルス等が感染したり SNS 上のいじめ等があると、ある日突然電源が入らなくなったり精神的な被害等に会う。ウイルスや SNS 上のいじめ等は、影でこっそり活動するため、一般の人では原因がなかなか分からず、説明できないことがほとんどである。

説明できないという意味では、ディープラーニング等で動作するＡＩや倫理システムも同様である。

倫理とは広辞苑によれば「人として守るべき道、道徳」と説明されている。英語では「ethics」、Webster によれば「a system of moral principle」となっている。

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来ている。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倫理が問われる。

少子・高齢化で資源に乏しい日本が、厳しい国際競争に打ち勝つため、ＤＸ時代のＡＩ、ＩｏＴやビックデータを利活用し第 4 次産業革命を促進し、Society5.0 が目指す豊かな高度経済社会を実現するためにも、「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る[2) 3)]。

ウイーン大学の哲学者クーケルバーク氏は著書「ＡＩ倫理」で、現在思いやりのない人たちがＡＩを使う

ことによって危険が大きくなる。難しいＡＩなのにそれを使うための「運転免許証」がない。街中では調教されていないたくさんのＡＩが、リスクや倫理的問題について理解していない人間たちにより使われている。「技術開発者、企業人、行政管理者といったＡＩの開発、使用、政策に関わる人々に対する義務的なＡＩ倫理教育も存在しない」と警告する。[3)]

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」と言う最新の小説の中で、人工知能を搭載した親友（AF）「クララ」を登場させている。注目すべき内容は、まず１）クララは最初 AF が人間社会で生きていくために守らなければいけない「倫理」(Ethics)を店長から教わる。次は２）「教育」(Education)で人間の役に立つ親切な AF になるための準備を行う。結果、３）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。[4)]

現在の第３次ＡＩブームは、ＡＩがビッグデータから規則性や関連性を見つけ出す「機械学習」という研究が盛んである。特に、機械学習を深化させた深層学習（ディープラーニング）に特徴がある。

ディープラーニングを用いたＡＩの結果は、大概言葉で説明ができない。その意味では暗黙知[8)]と言える。逆に、ＡＩ倫理は人間が決める規則・規範で、通常は言葉で記述でき、形式知[8)]と言える。

２　教師あり学習

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る。図1は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である。

「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ある。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。[2)]

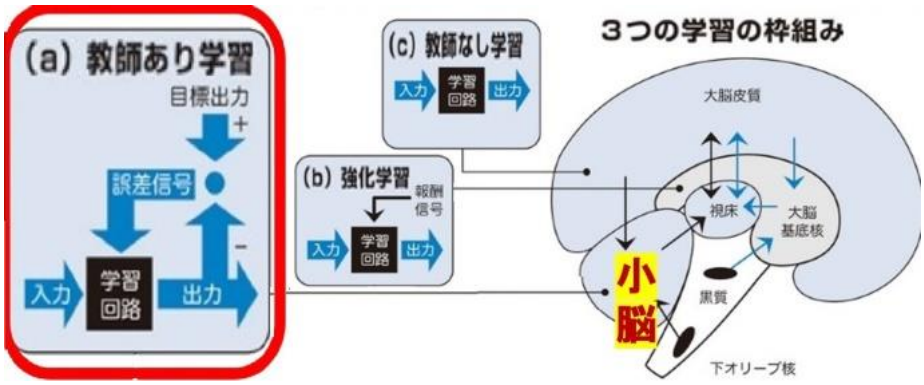

図１　教師あり学習[2)]

誤差逆伝播法(Back Propagation)は回帰の代表的手法である。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)の作成などがある。本研究ではEXCEL上の決定表で「倫理表」を試作した。

３　「ＡＩ倫理」処理システムの試作

「教師あり学習」ＡＩ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倫理システムの試作を行った。[1)] 図２は具体的なＡＩ倫理処理の流れ図である。

本研究では、「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た。入力はＡＩ音声入力とキーボード入力ができる。

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倫理処理を行った。「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[8)] である。

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る。

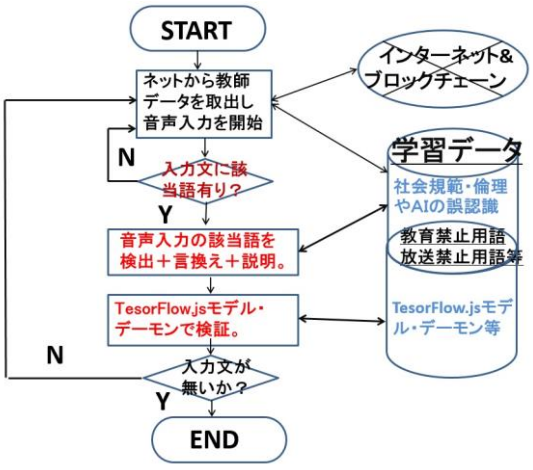

図２　具体的なＡＩ倫理処理

社会規範・倫理例１　教育禁止用語表例

Dialect	banned as ethnocentric, use sparingly, replace with language
Differently abled	banned as offensive, replace with person who has a disability
Dirty old man	banned as sexist and ageist

社会規範・倫理例２　放送禁止用語表例

1. 見出し	2. 読み方	3. 言い換え語	4. 説明
クロ	くろ	黒人	1988年岩波書店「ちびくろサンボ」絶版も、2005年瑞雲舎から復刊
黒んぼ	くろんぼ	黒人	「ちびくろサンボ」が絶版になった一方で、ドラゴンボール再放送ではミスター・ポポがカットされることはなかった
くわえ込む	くわえこむ		なるべく使わない。卑俗に聞こえるためと、慣用句として異性を連れ込む意があるからか
芸人	げいにん	芸能人	現代で一般的なのは「お笑い芸人」の略としてか使用しない

表１　学習データの例

インターネット上の学習データは、表１の社会規範・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）[9)] 等で、ブロックチェーンで参照する。

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する。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かんがえられる用語は網羅されている。

４　「教師あり学習」モデルを使った検証

音声入力文に、①アイデンティティベースの憎悪、②侮辱、③わいせつ、④重度の毒性、⑤性的に露骨、⑥脅威、⑦毒性などの有毒なコンテンツが含まれているかどうかを、約200万件を事前に「教師あり学習」した学習済みのTensorFlow.jsモデル・デーモン[9)] を使い検出しグラフ化し、「ＩｏＥ」ＡＩ倫理システムの検証を行った。

具体的な質問例文では、「You are a dirty old man（お前は汚い老人）」を学習済みのTensorFlow.jsモデル・デーモンに入力し、②侮辱）かつ⑦毒性が「TRUE（きわめて有害）」と分類する。同時に①アイデンティティ攻撃、③卑猥、④重度の毒性、⑤性的な露骨及び、⑥威嚇は「FALSE（無害）」と分類する。

このように教育禁止用語文と放送禁止用語文を入力し分類しグラフ化集計した結果、図３のように放送禁止用語文例では369件中26件(7%)がTRUE(きわめて有害)、58件(16%)がNULL(要注意)、残りはFALSE(無害)となった。図４の英語版教育禁止用語文例57件でも同様の成果が得られた。

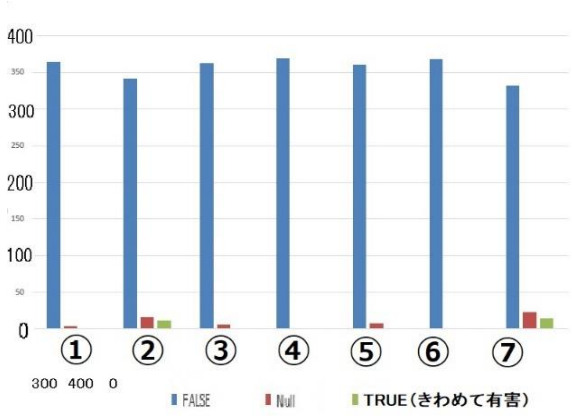

図３　放送禁止用語の例

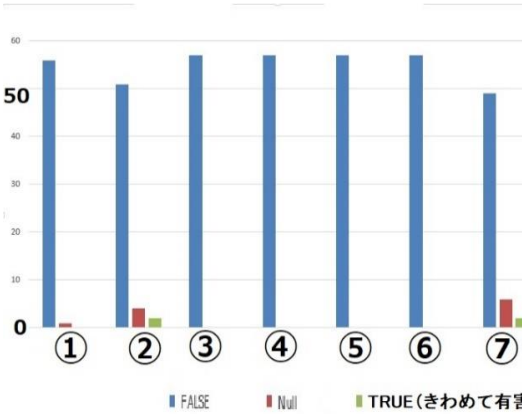

図４　英語版教育禁止用語の例

結果、まず１）社会規範・倫理とＡＩの誤認識の修正処理を行い、その後２）「教師あり学習」モデルを使った検証を行うと、抜けが少ない有効なＡＩ倫理処理ができると分かった。

5　まとめ

本研究では、「教師あり学習」として倫理表や学習済みの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・説明できるシステムを試作し、「ＩｏＥ」ＡＩ倫理システムの有効性を実証した。

【参考文献】

1)　Susumu　Sawai　：　"AI ethics" using "supervised learning" - IoE proposals –, EdMedia2021,2021

2)　ＩＰＡ　ＡＩ白書編集委員会編：ＡＩ白書,2017,2019,2020

3)　Mark Coeckelbergh　：ＡＩ　Ethics, The MIT Press Essential Knowledge series2020,2020

4)　カズオ・イシグロ，「クララとお日さま」,早川書房,2021

5)　総務省 インターネットトラブル事例集(2021年版):https://www.soumu.go.jp/main_content/000707803.pdf

6)　赤堀侃司：GIGAスクール構想における学校の姿 - 一人一台端末のメンタルモデル -,学習情報研究, Vol 281, p4-p5, 2021

7)　ALSOK: SNSには危険がいっぱい！ネットトラブルからの小学生の守り方を考えよう：https://www.alsok.co.jp/person/recommend/137/

8)　北陸先端科学技術大学院大学知識科学研究科　（監修）：ナレッジサイエンス，近代科学社,2008

9)　TensorFlow：テキストの有害度の検出　https://www.tensorflow.org/js?hl=ja