AI時代の教育論文誌 第３巻 pp.55-58（アーカイブ論文　受理日:2021年7月6日）

「教師あり学習」を使うAI倫理：IoE提案

"AI ethics" using "supervised learning"　－ IoE proposals　－

澤井　進*

*公益財団法人学習情報研究センター

＜抄　録＞

本研究は、Internetから社会規範・倫理の学習データを取得するＡＩ倫理を処理するシステムIoE（Internet of Ethics・Education・Energy of Life)の実現を目指している。本研究では、教育禁止用語や放送禁止用語等のような社会規範・倫理やＡＩの誤認識が処理・説明できるＡＩ倫理（IoE）システムを試作し、「教師あり学習」として倫理表やTensorFlow.jsモデルを使い、システムの有効性を実証した。

＜キーワード＞

AI倫理、人工知能、教師あり学習、IoE、禁止用語、社会規範、倫理表、TensorFlow.js モデル

１　ＡＩ倫理

本研究は、人工知能（ＡＩ）を規制するのではなく、ＡＩを人間に役立つ「友人」や「伴侶」として益々活用できるように教育することを狙っている。

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。

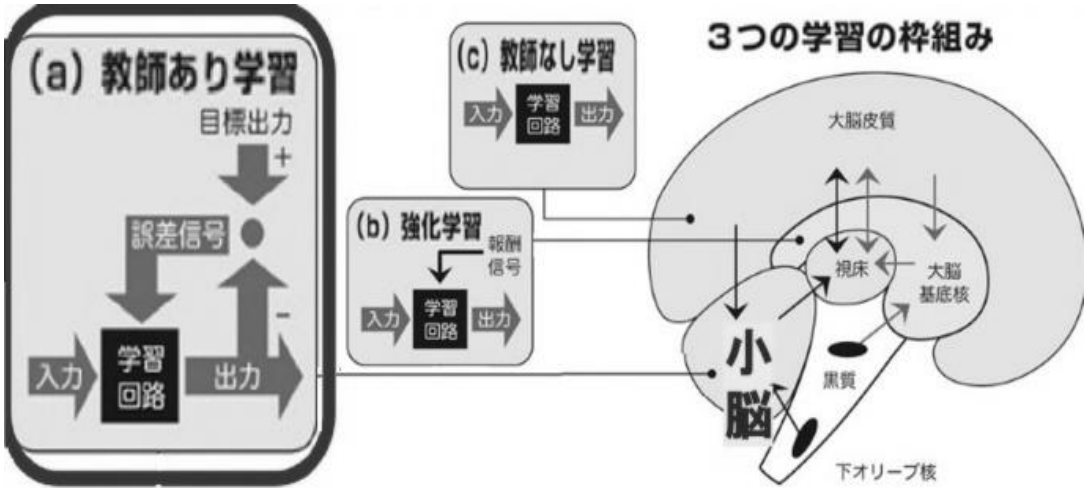

図１　教師あり学習[1)]

SAWAI Susumu*: AI ethics system needed for one device per person

*Information Research Center for Learning 1-5-16 Suido, Bunkyou-ku, Tokyo, 112-005 Japan　nac02440@nifty.com

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」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[1) 2) 3)]。

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[4)]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[4)]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３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。[1)]

「回帰」の代表的手法は誤差逆伝播法(Back Propagation)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。

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音声入力でもキーボード入力でもできる。

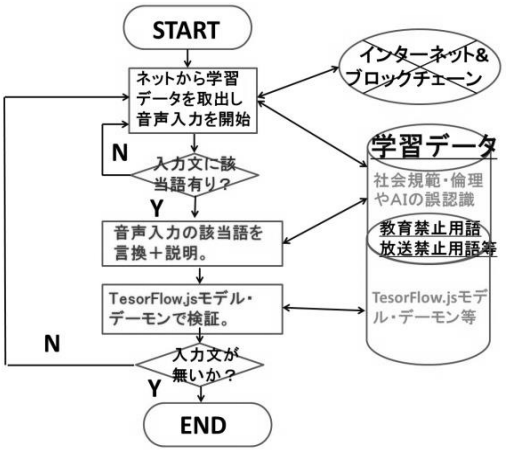

図２　具体的なＡＩ倫理処理

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[4)]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

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　社会規範・倫理例

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）[5] 等で、インターネットとブロックチェーンで参照する。

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[5] 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。

表２のように放送禁止用語例では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(まったく無害)となった。表３の英語版教育禁止用語57件でも同様の成果が得られた。

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倫理処理ができるだろうと分かった。

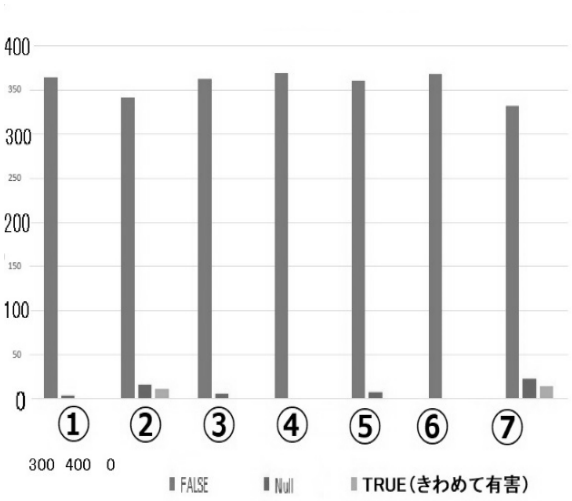

表２　放送禁止用語の例

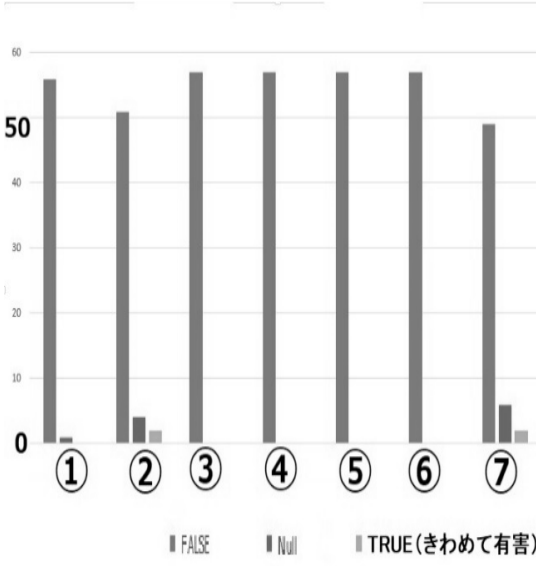

表３英語版教育禁止用語の例

5　まとめ

本研究では、「教師あり学習」として分析表や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、ＡＩ倫理（IoE）システムの有効性を実証した。

【参考文献】

1) AI白書,IPA AI白書編集委員会編,2017,2019,2020
2) AI Ethics, The MIT Press Essential Knowledge series2020
3) ｶｽﾞｵ･ｲｼｸﾞﾛ,「クララとお日さま」,2021
4) ﾅﾚｯｼﾞｻｲｴﾝｽ, 改訂増補版,近代科学社,2008
5) TensorFlow：テキストの有害度の検出 https://www.tensorflow.org/js?hl=ja、（参照 2021-06-10）
6) 澤井進:”一人一台端末に必要なＩｏＥ：ＡＩ倫理システム，AI時代の教育学会、第１回研究会,2021
7) Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Stephen Cave :The Role and Limits of Principles in AI Ethics:Towards a Focus on Tensions, AIES’19, January 27-28, 2019
8) F. LeRon Shults, Wesley J. Wildman: ARTIFICIAL SOCIAL ETHICS:SIMULATING CULTURE,CONFLICT, AND COOPERATION, SpringSim’20, May 19-May 21, 2020

本アーカイブ論文（注）は EdMedia + Innovate Learning, Jul 06, 2021 in United States ISBN 978-1-939797-56-8 の AACE 学会の出版済論文（827 頁-831 頁）である。

注）アーカイブ論文とは、二重投稿ではない翻訳された査読有のジャーナル論文である。