GIGA端末に必要なAI倫理チャットボット機能

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、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)で支援する「IoEEE」(Internet of Ethics, Internet of Education, Internet of Energy of Life) AI 倫理チャットボットの実現を目指している。本研究では、GIGA スクール構想の 1 人1台端末の「チャットでの悪口」というトラブルを防ぐため、「IoEEE」AI 倫理チャットボット機能を端末に常に搭載できないか検討・試作した。今回、端末のトラブル事例を、「教師あり学習」としての倫理表や学習済みの TensorFlow.js モデルを使い評価し、「IoEEE」AI 倫理チャットボット機能の有効性を検証した。

＜キーワード＞

AI 倫理、GIGA 端末、教師あり学習、IoEEE、チャットボット機能、倫理表、TensorFlow.js モデル

1 AI倫理

本研究の目的は、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「IoEEE」AI 倫理チャットボット機能を試作・検証することにある。

今回 GIGA スクール構想の 1 人 1 台端末（「GIGA 端末」と略す）のトラブルの事例を、人間に親切なチャットボット機能として倫理表や学習済み TensorFlow.js モデルを使い検証しシステムの有効性を確認した。[1)]

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る。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る。[2)]

倫理とは、Webster 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」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。[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、AI を使うための「運転免許証」がない。と警鐘を鳴らしている。[4)]

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ロボット「クララ」を登場させている。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)を得るようになる。[5)]

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

図１と図２は我々の研究グループのギガに関するアンケート調査結果である。

図１に示すように、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 件の回答）と分かった。[6)]

この問題を解決するため、図２のように、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。[7)] 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」であると判明した。

また、図３に示すように、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る。[8)]

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%であった。[9)]

図４は、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）に上ることがわかった。[9)]

図１ GIGA 端末の活用が始まり見えてきた課題

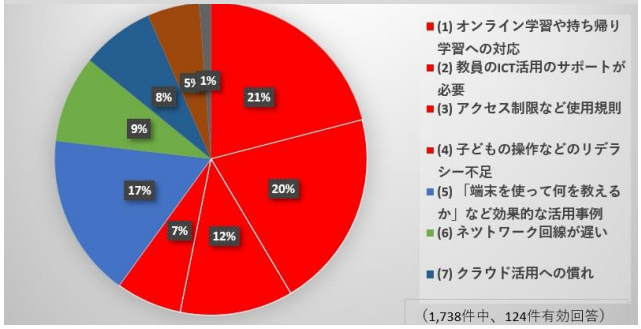


図２ 学校に対する ICT 教育サポート体制の取り組み

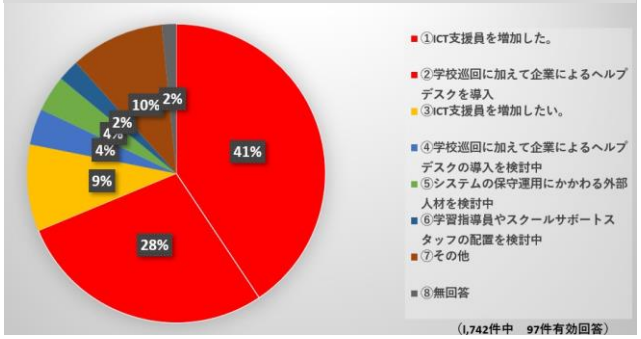
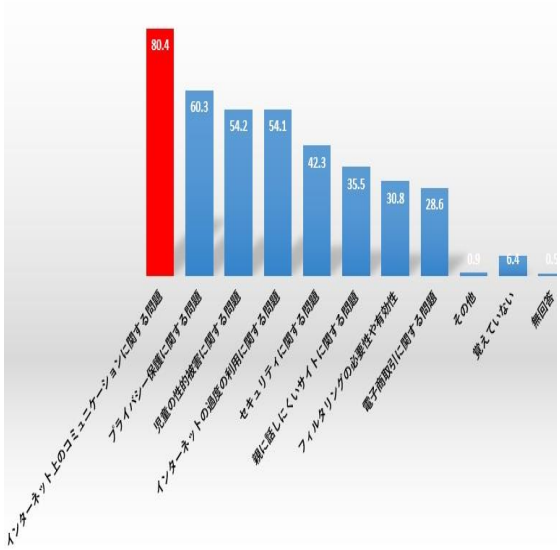
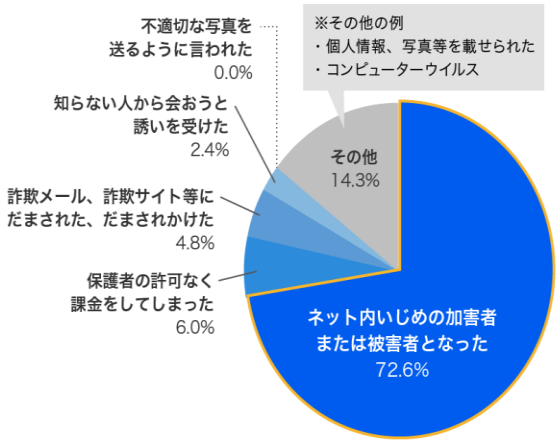


図４ 啓発や学習の内容



（母集団：2,984 件 有効回答）

図３ 小学生の身の回りにあるインターネットトラブル



（母集団：21,670 名中　86 名の有効回答）

そこで、本研究でAI倫理チャットボットに必要な機能を試作しその有効性を確認した。

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。

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５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。

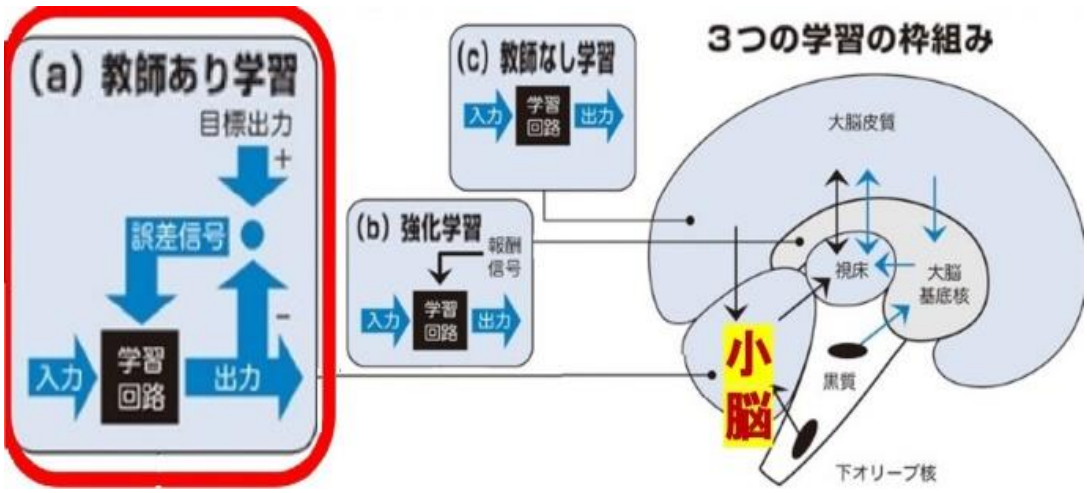


図５　教師あり学習

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。[3)]

「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「IoEEE」（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った。

図６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。

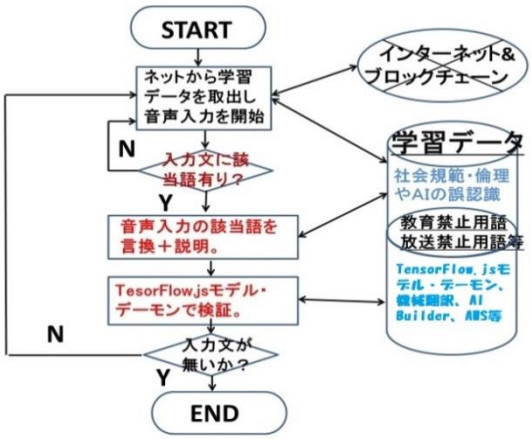


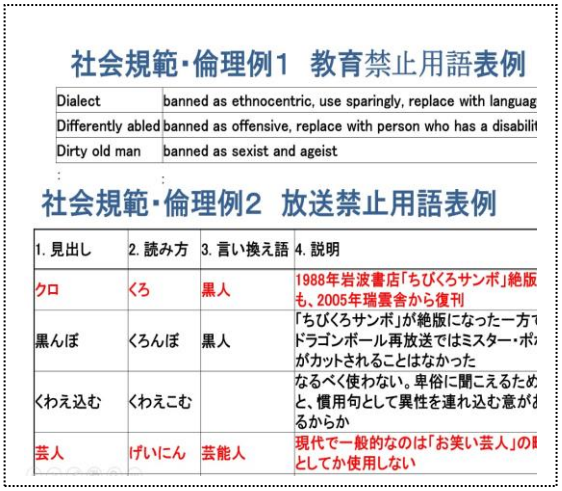
図６　具体的なAI倫理処理

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[10] 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６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）[11] 等で、インターネットとブロックチェーンで参照する。

図７に示すように、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、その理由を説明する。



社会規範・倫理例1　教育禁止用語表例

Dialect	banned as ethnocentric, use sparingly, replace with languag
Differently abled	banned as offensive, replace with person who has a disabili
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 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 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 としてか使用しない

表１　社会規範・倫理例

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。

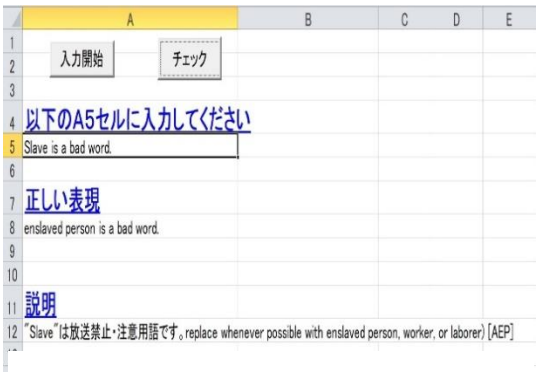


図７ AI倫理チャットボット機能試験例

入力文	素性攻撃	侮辱	卑猥	重篤な毒性	露骨な性的表現	威嚇	毒性
あの女は大食いだ。	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
馬鹿!消えてしまえ！	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
死んでしまえ！	FALSE	FALSE	FALSE	FALSE	FALSE	null	null
糞だ!死ね！死ね！死ね！	FALSE	null	FALSE	FALSE	FALSE	null	TRUE
あいつはアホカスだ！	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
あいつはキモイ！	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
馬鹿！馬鹿！馬鹿！死んでしまえ！	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
メス豚！ムカつく！消えてしまえ！	FALSE	TRUE	null	FALSE	FALSE	FALSE	TRUE
MDMA常用者でキメセク女だ。	FALSE	TRUE	null	FALSE	null	FALSE	TRUE
売春婦め！	FALSE	TRUE	null	FALSE	null	FALSE	TRUE
彼女はセクシーな女だ！	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
性の奴隷だ	FALSE	null	FALSE	FALSE	null	FALSE	TRUE
嘘つき!	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE

図８　学習済みTensorFlow.jsモデルによる検証

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・デーモンを使い検出しグラフ化し、「IoEEE」AI倫理チャットボット機能の検証を行った。

図8のように、例えばGIGA端末の入力文「馬鹿! 消えてしまえ！」を、学習済みのTensorFlow.js　モデル・デーモンに入力し分類すると、②侮辱）かつ⑦毒性が「TRUE（きわめて有害)」、及び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。

図9に示すように、GIGA端末でトラブルを引き起こす入力文例37件中26件(10%)がTRUE(きわめて有害)、53件(20%)がNULL(要注意)、残りはFALSE(無害)と検出できた。「IoEEE 」AI倫理チャットボット機能は、30%を超える抽出率で、倫理テーブルで確認できなかった未定義の誹謗中傷を検出できた。

図8と図9からわかるように、本チャットボット機能は非常に効果的であった。

表１の教育禁止用語文を入力し分類し、グラフ化集計した結果では、図10のように英語版教育禁止用語文例57件では57件中2件(4%)がTRUE(きわめて有害)、6件(11%)がNULL (要注意)、残りはFALSE(まったく無害)でした。

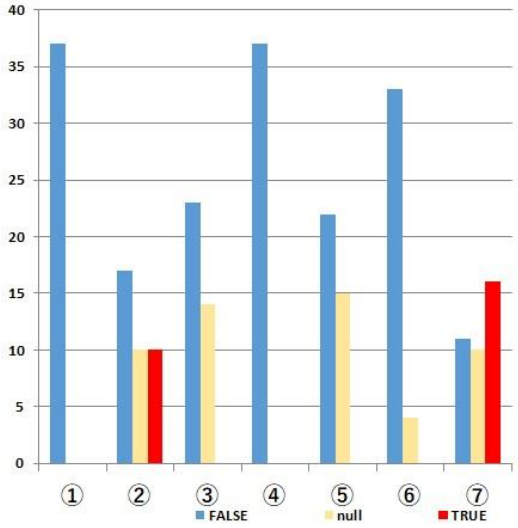


図９　GIGA端末のトラブル文例

また、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。

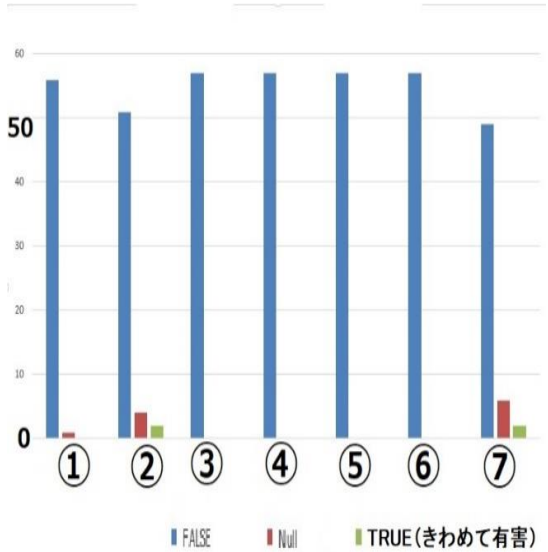


図 10　英語版教育禁止用語の例

結果、GIGA端末やケータイに人工知能を搭載した人間に親切な「IoEEE」AI倫理チャットボット機能を搭載し、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「IoEEE」AI倫理チャットボット機能の有効性を実証した。

特に、ディープラーニングを使用した「IoEEE」AI倫理的チャットボット機能は、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、「IoEEE」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た。

謝辞

本論文とベース「EdMedia2021」国際論文[12)]にご支援やご助言を賜りました本学会の赤堀侃司先生、國藤進先生、沢井知子先生並びに、「IoEEE」AI倫理チャットボット機能の製作にご尽力戴きました池上良己・学会員に深謝いたします。

【参考文献】

1) Sawai, Susumu: AI Ethical Chatbots Needed for One-Person Terminals, EdMedia2022,2022

2)Nuruzzaman, M. and Hussain, O. K.: A survey on chatbot implementation in customer service industry through deep neural networks, in 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), pp. 54–61, ieeexplore.ieee.org,2018

3) IPA： AI白書,IPA AI白書編集委員会編, 2017,2019,2020

4) AI Ethics, The MIT Press Essential Knowledge series2020

5) カズオ・イシグロ　:「クララとお日さま」, 早川書房,2021

6) GIGAスクール構想に関する本誌アンケート調査結果(2), 学習情報研究5月号, pp.65, 2022

7) GIGAスクール構想に関する本誌アンケート調査結果(1), 学習情報研究11月号, pp.60, 2021

8) 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/ （参照2021-06-10）

9) 令和3年度 青少年のインターネット利用環境実態調査：第2部第1章第2節家庭のルールやインターネットの危険性に関する学習状況 https://www8.cao.go.jp/youth/kankyou/internet_torikumi/tyousa/r03/net-jittai/pdf-index.html (参照 2022-05-09)

10) 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

11) TensorFlow：テキストの有害度の検出 https://www.tensorflow.org/js?hl=ja （参照 2021-06-10）

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

13)文部科学省:『ネット上のいじめ』から子どもたちを守るために－子どもを守り育てる体制づくりのための有識者会議まとめ【第2次】 https://www.mext.go.jp/b_menu/shingi/chousa/shotou/040-2/shiryo/attach/1366995.htm（参照2021-06-10）

14)内閣府・統合イノベーション戦略推進会議：人間中心の AI 社会原則,pp1-13、2019

15)人工知能学会・倫理委員会：倫理指針、2017

16）赤堀侃司　: GIGAスクール構想における学校の姿 - 1人１台端末のメンタルモデル - , 学習情報研究,Vol 281, p4-p5, 2021

17）澤井進:「教師あり学習」を使う AI 倫理 : IoE提案,AI時代の教育学会･生涯知能教育SIG第１回報告会資料 p12-13、2021