超AI世代教育シリーズ

教師あり学習を用いたAI倫理

「背景画像作成:DALL-E」

【内　容】

1. AI倫理の歴史
2. 教師あり学習
3. 「教師あり学習」を用いたAI倫理処理
4. まとめ

1. ＡＩ倫理の歴史

1.ＡＩ倫理の歴史

現在、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る。

２１世紀の基幹テクノロジーであるAIとどう付き合うか？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。

1.ＡＩ倫理の歴史

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った

1.ＡＩ倫理の歴史

1）初期のAI倫理の概念（1950～1960年代）

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た。

①ロボットは人間に危害を加えてはならない。

②ロボットは人間の命令に従わなければならない。

③ロボットは自己を守らなければならない。

1.ＡＩ倫理の歴史

2）AI技術の発展と懸念拡大（1970～1990年代）

AI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た。

1980年代：エキスパートシステムの導入により、AIの誤った判断やバイアスの問題が浮上し始めた。

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の懸念が増大した。

1.ＡＩ倫理の歴史

3）倫理的ガイドラインの登場（2000～2010年代）

2000年代：政府や学術機関が必要性を認識した。IEEEは倫理的AIの開発の国際的な指針を提案。

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た。

1.ＡＩ倫理の歴史

4）国際的AI倫理の取り組み（2010～2020年代）

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。

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、AIガバナンスを推進した。

1.ＡＩ倫理の歴史

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る。

1.ＡＩ倫理の歴史

5）日米企業のAI倫理政策の代表例

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。AIの利用に関するガイドラインを明確にしている。

米国IBM（2018年9月）：IBMは「Everyday Ethics for Artificial Intelligence」を発表し、AI導入の透明性と公正な判断の重要性を訴えた（フィードバック体制を整備する実践例等）。

1.ＡＩ倫理の歴史

5）日米企業のAI倫理政策の代表例

ソニーグループ（2018年9月）：「AI倫理ガイドライン」を策定し翌4月からの実践の安全性審査を開始。

富士通（2019年3月）：「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た。

NEC（2019年4月）：グループ全体でAI倫理の取り組みを開始（AIコンソーシアムのガイドライン参）。

1.ＡＩ倫理の歴史

5）日米企業のAI倫理政策の代表例

NTTデータ（2019年5月）：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る。

日立製作所（2021年2月）：AI倫理委員会を設置し、グループ全体の倫理的なAI運用を図っている。

1.ＡＩ倫理の歴史

5）日米企業のAI倫理政策の代表例

Yahoo！（2022年5月）：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。

パナソニック（2022年8月）：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る。

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る。

1.ＡＩ倫理の歴史

日本のAI倫理政策と経団連のAI倫理ガイドライン

1）2019年日本の「AI原則」を公表：

人間中心の原則：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：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：悪用や誤用のリスクの対応。

1.ＡＩ倫理の歴史

日本のAI倫理政策と経団連のAI倫理ガイドライン

2）2019年経団連「AI活用の倫理原則」を発表：

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、公平性と多様性、透明性と説明可能性及び、安全と信頼の確保。

　目的は、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。

1.ＡＩ倫理の歴史

米国の「AI権利章典」と企業の対応

2022年に「AI権利章典」を発表し、AIの設計・利用における以下の5つの基本原則を提示：

「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」。

　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る。

1.ＡＩ倫理の歴史

英国のAI規制のフレームワーク

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、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る。

1.ＡＩ倫理の歴史

EUのAI責任指令案・製造物責任指令の改正案

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る。

1.ＡＩ倫理の歴史

シンガポールの「モデルAIガバナンス枠組み」と「AI. Verify」

「モデルAIガバナンス枠組み」と「AI. Verify」（AIの検証）という検証ツールを発表した。

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。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る。

1.2 ＡＩ倫理の定義

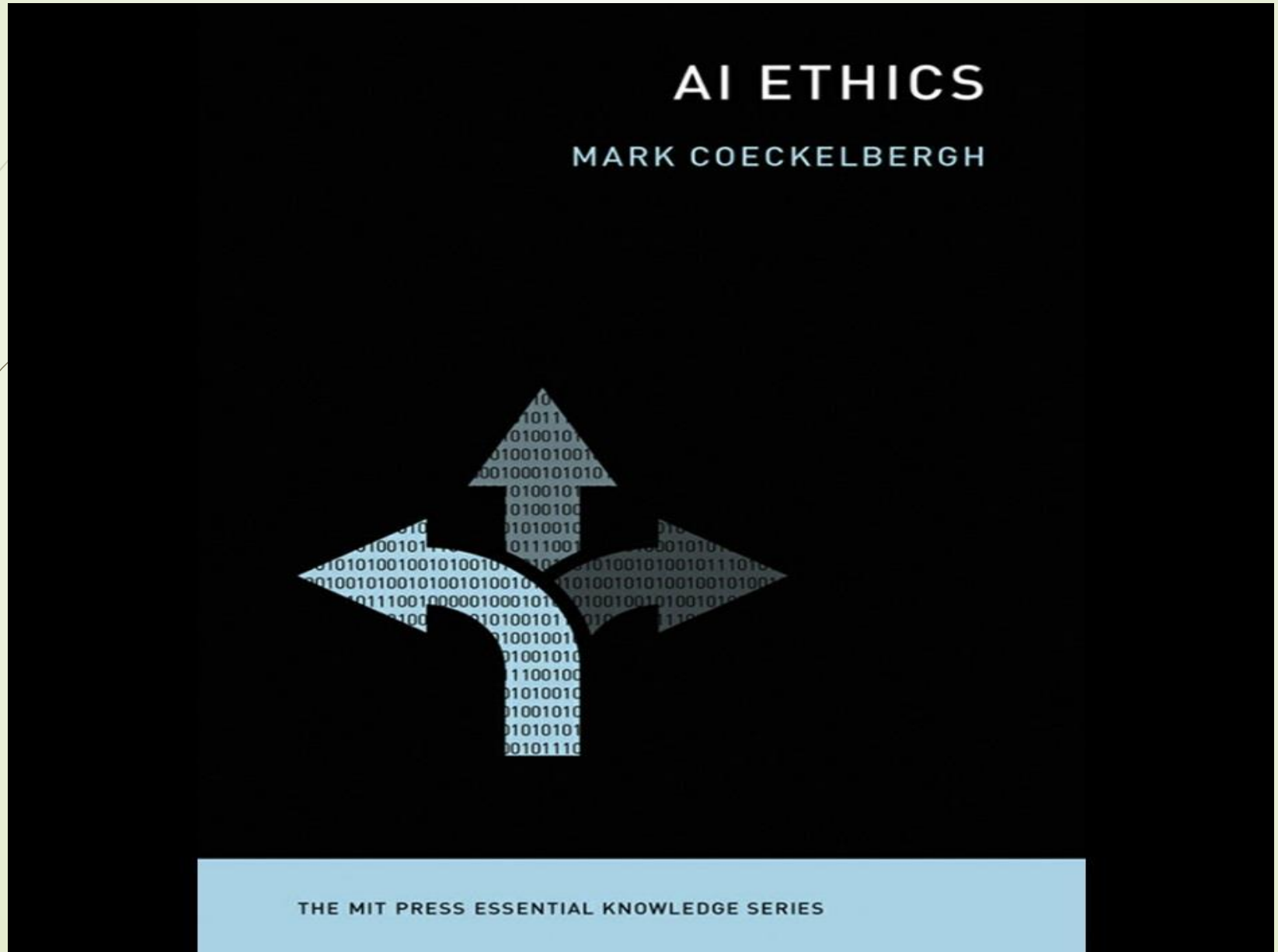
1）倫理とは
「人として守るべき道、道徳」
（広辞苑）

2）英語では「ethics」
「a system of moral principle」
（Webster辞書）

1.2 ＡＩ倫理の定義

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。

ＡＩの倫理学(2020.12.25初版)



pluralistic with regard to methods and approaches, its topics, and also its media and technologies. To put it bluntly: if engineers learn to do things with texts and humanities people learn to do things with computers, there is more hope for a technology ethics and policy that works in practice.

The Risk of an AI Winter and the Danger of the Mindless Use of AI

If these directions in policy and education do not get off the ground and, more generally, if the project of ethical AI fails, we face not only the risk of an "AI winter"; the ultimate and arguably more important risk is ethical, social, and economic disaster and its related human, nonhuman, and environmental costs. This has nothing to do with the singularity, terminators, or other apocalyptic scenarios about the distant future, but with the slow but certain increase in the accumulation of technological risk and the resulting growth of human, social, economic, and environmental vulnerabilities. This increase in risks and vulnerabilities is related to the ethical problems indicated here and in the previous chapters, including the ignorant and reckless use of advanced automation technologies such as AI. The gap in education is perhaps exacerbating what AI risks do in general: even if it does not always directly cause new risks, it also and especially *multiplies existing risks*. So far there is no such thing as a "driver's license" for using AI, and there is no compulsory AI ethics education for techni-

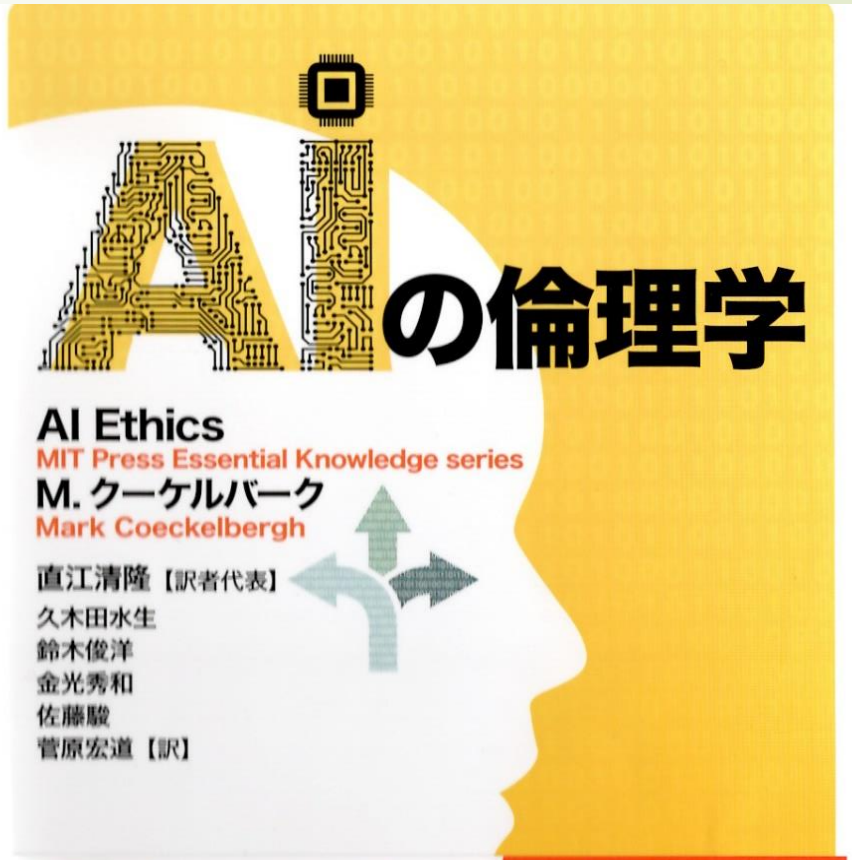
cal researchers, business people, government administrators, and other people involved in AI innovation, use, and policy. There is a lot of untamed AI out there in the hands of people who don't know the risks and ethical problems, or who may have the wrong kind of expectations about the technology. The danger is, once again, the exercise of power without knowledge and (therefore) without responsibility—and, worse, others being subjected to this. If there exists such a thing as evil at all, it lives where the twentieth-century philosopher Hannah Arendt located it: in the mindlessness of banal everyday work and decisions. To assume that AI is neutral and to use it without understanding what one is doing contributes to such mindlessness and, ultimately, to the ethical corruption of the world. Education policy can help to mitigate this and thus contribute to good and meaningful AI.

A number of nagging, perhaps slightly painful questions remain, however, which are often neglected in discussions about AI ethics and policy but that at least deserve mention, if not a lot more analysis. Is AI ethics all about the good for, and value of, humans, or should we also take into account nonhuman values, goods, and interests? And even if AI ethics is mainly about humans, could it be that the question regarding AI ethics is not the most important problem for humanity to address? This question brings us to the last chapter.

ＡＩの倫理学(2020.12.25初版)

ISBN978-4-621-30588-1
C1012 ¥2400E
9784621305881
定価（本体 2,400 円＋税）
倫理学・情報倫理
1921012024008

AIの倫理学
AI Ethics
MIT Press Essential Knowledge series
M・クーケルバーク
Mark Coeckelbergh
【訳者代表】直江清隆



浸透しつつあるAIとどう手を取り合うか？

いま学んでおきたい「AI倫理」の最前線！
〈弱いロボット〉提唱者
岡田美智男氏推薦！

丸善出版

ＡＩの倫理学(2020.12.25初版)

・M.クーケルバーク（ウィーン大学）著

「AIの使用に関して『運転免許書』がない

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。

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たちによって使用されていて、その人たちは、AI技術に対して間違った期待すら持っているかもしれないのである」

クララとお日様

3.10初版

クララとお日様(2021.3.10初版)

- ノーベル文学賞受賞者カズオイシグロ著
- 人工知能**友人**（AF）クララは病弱の少女ジョジーと出会いやがて2人は友情を育む。
- クララは**教わったことを思い出し、**
- 外界を観察し、**親切で役に立つAFになる為できるだけの準備をしたいと思い、**
- クララは、**観察と学習への意欲に優れ、**結果として、**精緻な理解力**をもつまでになりました。

クララとお日様(2021.3.10初版)

・教わったこと ⇒Ethics(AI倫理)

・役に立つAFになる為の準備
⇒Education(教育)

・観察と学習への意欲と理解力
⇒Energy of Life
(生きる力)

1.2 ＡＩ倫理の定義

・2006年からの第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。

・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。

1.2 ＡＩ倫理の定義

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、形式知と言える。

2. 教師あり学習

2. 教師あり学習

- 「人間の脳のモデル」

人間の脳は100億個以上のニューロン（神経細胞）から構成されている。

- 脳のモデルを作る研究の2つの目的：

1）すぐれた情報処理装置を作ること

2）脳のしくみを解明すること

⇒脳の中にできる外界の写し「世界像」を自己形成し、思考・行動するような機械を創る

2. 教師あり学習

- **「機械学習」**

　脳における学習の枠組みに基づくと、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。

　脳の部位として、それぞれ小脳、大脳皮質、と大脳基底核と深く関連がある。

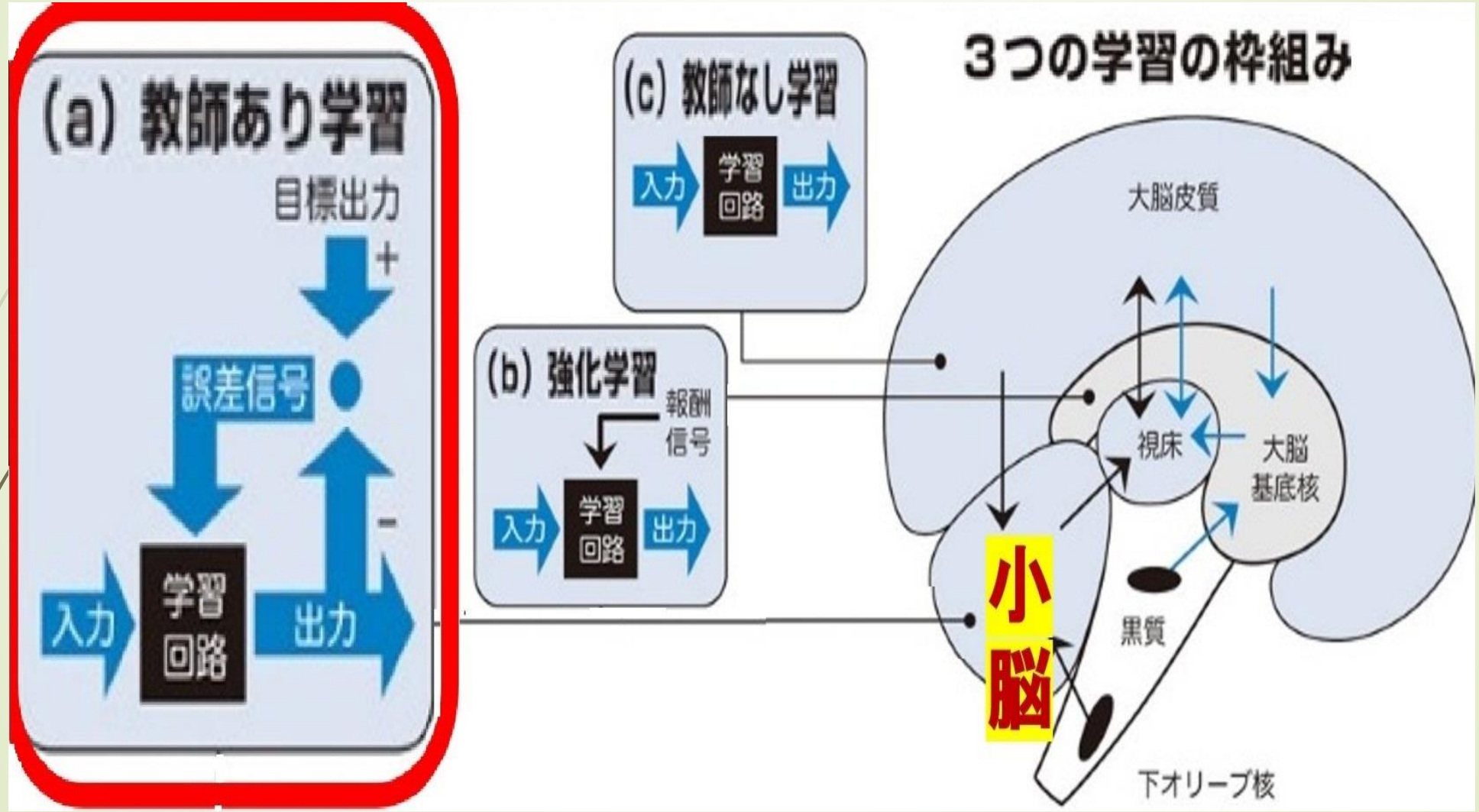
２．教師あり学習

- **「教師あり学習」**

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．教師あり学習


(a) 教師あり学習
目標出力
+
誤差信号
−
入力
学習回路
出力
(c) 教師なし学習
入力
学習回路
出力
(b) 強化学習
報酬信号
入力
学習回路
出力
3つの学習の枠組み
大脳皮質
視床
大脳基底核
小脳
黒質
下オリーブ核

２．教師あり学習

「教師あり学習」

誤差逆伝播法(Back Propagation)は回帰の代表的手法である。分類手法として、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MicroSoft EXCEL上の倫理表を決定表で試作した。

3.「ＡＩ倫理」処理システムの試作

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AI倫理（IoE: Internet of Ethics・Education・Energy of Life)システムの試作を行った。

ディープラーニングでＡＩ音声入力

I-phone

こんにちくろのなかには、げいにんとして・・

リモートマウス

パソコン

今日黒人の中には芸能人として・・

※説明；「くろ」は放送禁止用語で、　・・
※説明；「げいにん」は放送禁止用語です。

ＡＩ音声入力
（ディープラーニング）

ＡＩ倫理処理

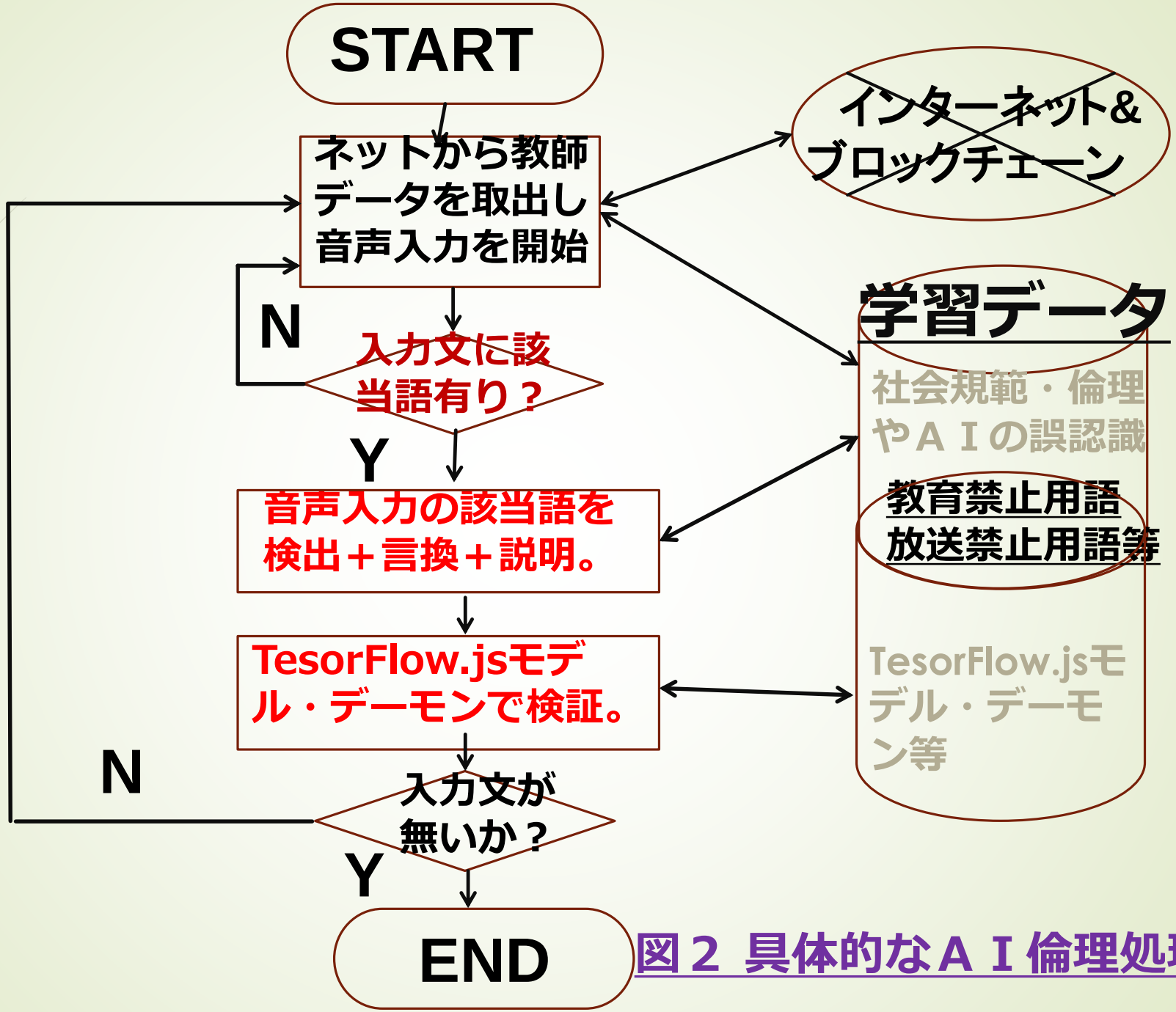


図２ 具体的なＡＩ倫理処理

表1　学習データの例

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」

表1の学習データの詳細(1)

社会規範　倫理例１　教育禁止用語表例

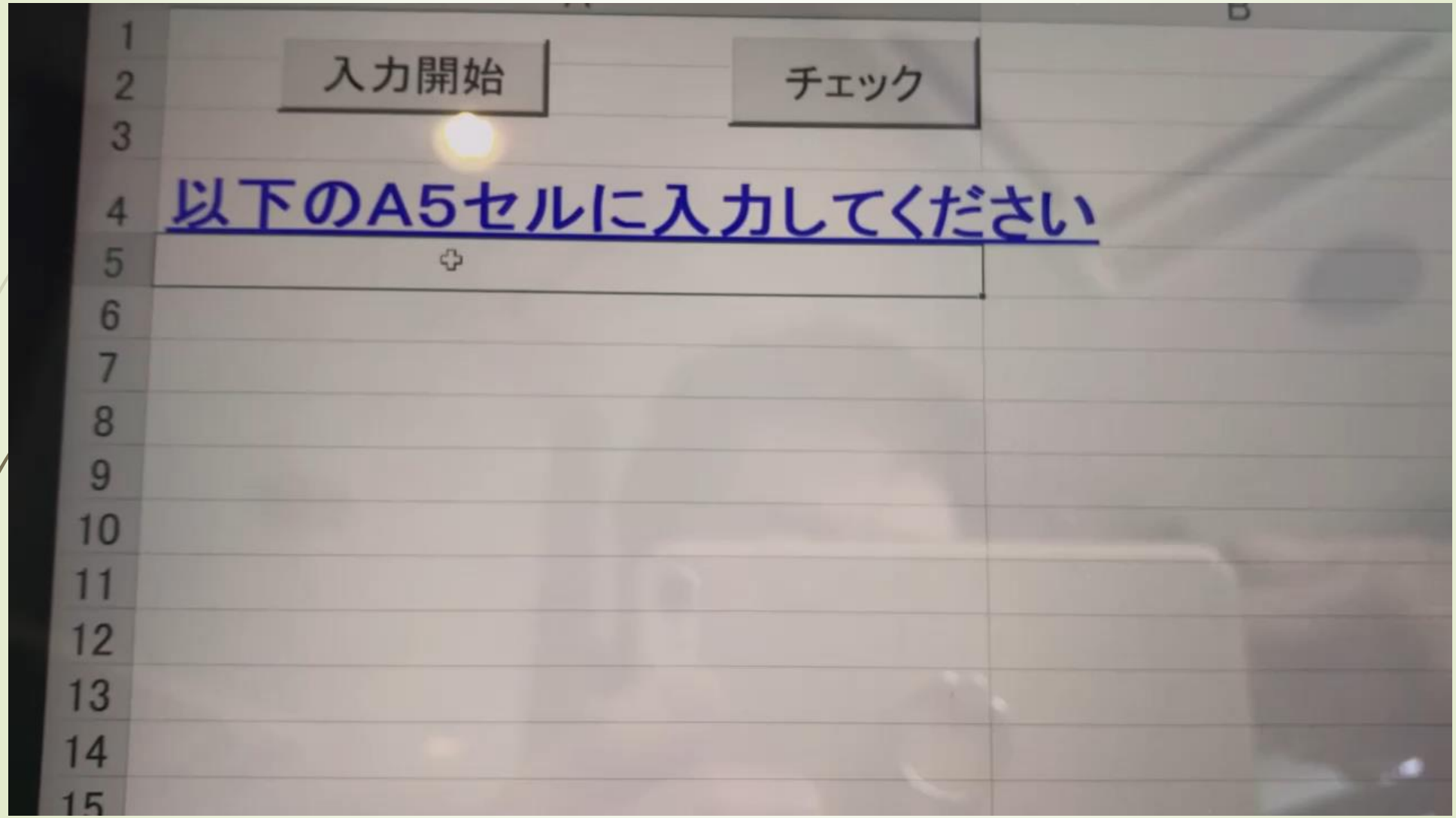
	A	B	D
371	Dialect	方言	banned as ethnocentric, use sparingly, replace with language
372	Differently abled	障害者	banned as offensive, replace with person who has a disability
373	Dirty old man	汚い老人	banned as sexist and ageist) [NYC]
374	Disabled, the	無効、	banned as offensive, replace with people with a disability) [SI
375	Dissenter	反対者	ethnocentric, use with caution) [ETS2]
376	Distaff side, the	糸巻き棒側、	banned as sexist) [ETS2]
377	Dogma	ドグマ	banned as ethnocentric, replace with doctrine, belief) [SF-AV
378	Doorman	ドアマン	banned as sexist, replace with door attendant) [HRW1]
379	Down's syndrome	ダウン症候群	banned as offensive, replace with Down syndrome) [ETS2]
380	Draftsman	製図工	banned as sexist, replace with drafter) [NES]
381	Drunk, drunken, dru	酩酊、酩酊、酩酊	banned as offensive when referring to Native Americans) [SF
382	Duffer	ダッファー	banned as demeaning to older men) [SF-AW]
383	Dummy	ダミー	banned as offensive, replace with people who are speech imp
384	Dwarf	ドワーフ	banned as offensive, replace with person of short stature) [S
385	Heretic	異端者	use with caution when comparing religions) [ETS2]

表1の学習データの詳細

社会規範・倫理例２　放送禁止用語表例(2)

	A	B	C	D
1	アイヌ系	あいぬけい	アイヌ アイヌ民族	アイヌ系はアイヌ民族に対する強制同化が生ん
2	合いの子	あいのこ	混血 混血人	慣用的に用いる場合は「足して2で割った」ほど(
3	青姦	あおかん	野合	屋外で性交すること。意外に知らない人が多い
4	アカ	あか	共産主義者 共産シンパ	
5	明盲	あきめくら	字の読めない人 非識字者	言い換え語に文盲があるが好ましくない
6	足切り	あしきり	予備選抜 二段階選抜	漫画「ラブひな」では、暗示として主人公の足が
7	足を洗う	あしをあらう	更生する	洗足池（東京）は「日蓮上人が足を洗った」池で
8	当て馬	あてうま	交代要員	野球の場合
9	アメ公	あめこう	アメリカ人	
10	あらめん	あらめん	初対面	探偵用語
11	アル中	あるちゅう	アルコール中毒 アルコール依存症	「…中毒」は急性、「…依存症」は慢性のもの

ＡＩ音声入力とＡＩ倫理処理の実際



4. まとめ

４．まとめ

本研究では、「**教師あり学習**」として倫理表や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。**

ご清聴ありがとうございました。

お問合せ：sawai@gakujoken.or.jp