機械学習における
説明可能性・公平性・安全性への工学的取り組み

開催趣旨

機械学習技術が様々なシステムに組み込まれて、社会に広がっています。それにつれて、高い精度が得られる一方、説明可能性（ブラックボックス問題）、公平性（差別・偏見問題）、安全性（品質保証・動作保証問題）の課題も指摘されるようになってきました。本企画セッションは、これらの課題に対する取り組みの動向を解説し、これからの方向性・対策を議論します。

プログラム

講演1	**イントロダクション －JST戦略プロポーザルの紹介－**	福島 俊一（科学技術振興機構）
講演2	**機械学習の説明可能性への取り組み －DARPA XAIプロジェクトを中心に－**	川村 隆浩（科学技術振興機構）
講演3	**機械学習の公平性への取り組み －Fairness-aware data miningを中心に－**	神嶌 敏弘（産業技術総合研究所）
講演4	**機械学習の安全性への取り組み －自動車業界の取り組みを中心に－**	中江 俊博・桑島 洋（デンソー）
講演5	**MLSE研究会・QA4AIコンソーシアムの活動・成果物報告**	石川 冬樹（国立情報学研究所、 機械学習工学研究会MLSE主査）

JSAI2019 企画セッション KS-5【講演1】

機械学習における説明可能性・公平性・安全性への工学的取り組み

イントロダクション －JST戦略プロポーザルの紹介－

2019年6月5日　福島 俊一

国立研究開発法人科学技術振興機構 研究開発戦略センター
Center for Research and Development Strategy Japan Science and Technology Agency

社会に広がるAI技術の光と影

AIのご利益

AIへの懸念

- 自動化による効率化、人間の負荷軽減
- 人間が気づかなかった選択肢を発見、可能性を拡大
- 人間を上回る精度・速度

- AIの代替によって消える職業
- AI判定に潜むかもしれない差別・偏見
- 監視社会、プライバシー不安
- AIの軍事利用、AIの悪用

AI社会原則　国・世界レベルで議論・策定

- 日本政府「**人間中心のAI社会原則**」
- 欧州委員会「**信頼できるAIのための倫理指針**」(Ethics Guidelines for trustworthy AI)
- IEEE「**倫理的に配慮されたデザイン**」(Ethically Aligned Design)
  - IEEE-SAでの標準化活動も進行中
- OECD「**人工知能に関するOECD原則**」(OECD Principles on Artificial Intelligence)
  - 42か国が署名

人間の尊厳が尊重される社会(Dignity)、多様な背景を持つ人々が多様な幸せを追求できる社会(Diversity & Inclusion)、持続性ある社会(Sustainability)という３つの価値を基本理念とし、「AI-Readyな社会」をビジョンに掲げる

AI社会原則

① 人間中心の原則
② 教育・リテラシーの原則
③ プライバシー確保の原則
④ セキュリティ確保の原則
⑤ 公正競争確保の原則
⑥ 公平性・説明責任・透明性の原則
⑦ イノベーションの原則

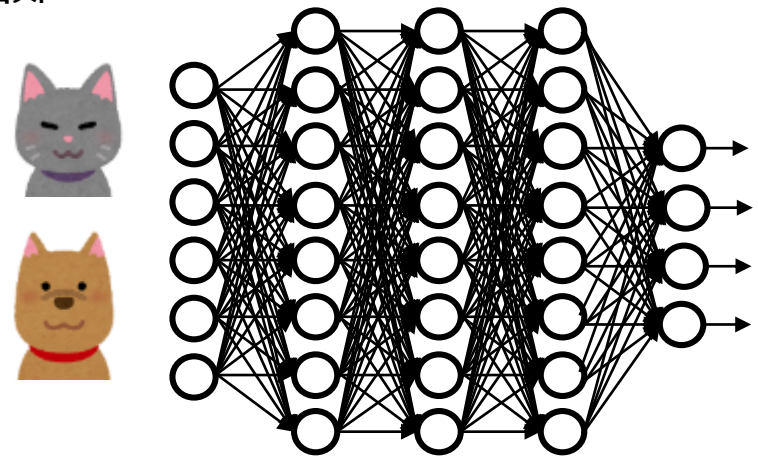
だが、この原則を満たすAIシステムはどう作ればよいのか（技術課題が存在）

AIへの懸念

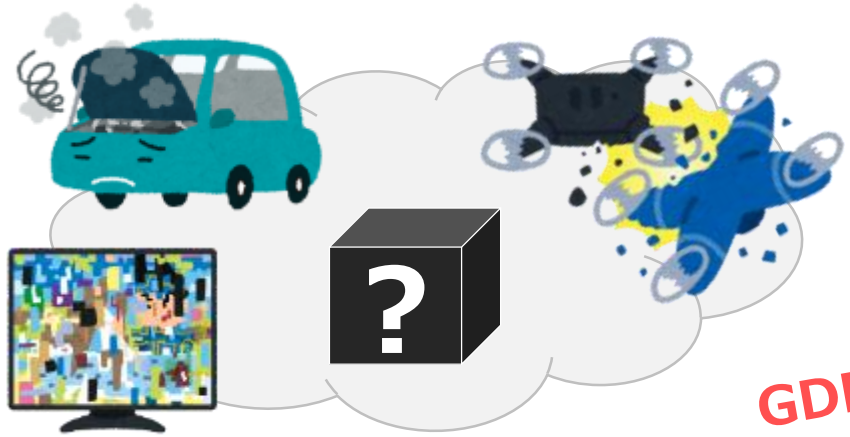
AIのブラックボックス問題(説明責任)

- 深層学習は高精度だが、人間に理解可能な理由説明を出してくれない
- ブラックボックスだと、どんな振る舞いをするか動作保証ができない
- 事故発生時に、原因解明や責任判断ができない

深層学習では、規則性が多層ニューラルネットワークにおけるリンクの重みになるので、人間が意味を理解することが困難



- ✓ なぜ「ネコ」と判断したのか?
- ✓ なぜ「イヌ」と判断したのか?



GDPR違反！

欧州GDPRはAIの透明性を要求

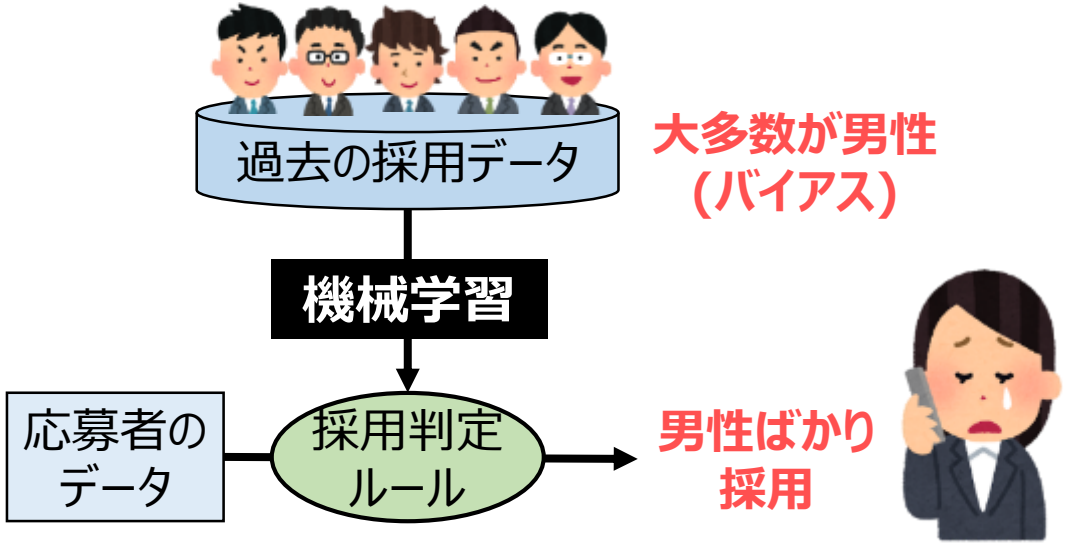
EU一般データ保護規則(GDPR、2018年5月施行)の第22条「プロファイリングを含む自動化された意思決定」は、自動処理の透明性(説明責任)を要求

AIへの懸念

AIのバイアス問題(公平性)

- 機械学習の判定結果は、学習データの傾向を反映する
- 学習データが偏見を含んでいれば、判定結果にも偏見が反映される
- 学習データの分布の偏りが差別を生むこともある

AmazonがAI採用中止、女性差別の欠陥露呈で



[ニュースソース] https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN

機械学習における差別・偏見(例)

- ✓ 特定人種の再犯率を実際よりも高く判定してしまう
- ✓ 特定人種のみ顔認識の誤認識率が高く、不利益を被る
- ✓ 採用判定、コンテスト等で、特定の人種・性別のみ高く評価してしまう

AIへの懸念

AIの脆弱性問題

- 学習データと比べて想定外のデータに対して、どう振る舞うかは不明
- 誤認識を誘発する攻撃(Adversarial Examples)が可能
- 悪意をもった追加学習によって不適切な振る舞いを引き起こされる

Adversarial Examples攻撃

[出典] https://arxiv.org/pdf/1707.08945.pdf

深層学習が停止標識と認識できていたものを、人間は気にかけない程度の小さな表示加工によって、速度制限標識と騙すことができてしまった

Microsoft Tay 公開停止

機械学習型チャットボットTayが、悪意のあるユーザーによって差別と陰謀論に染まってしまい、不適切な発言を繰り返したため、わずか一日で公開を停止

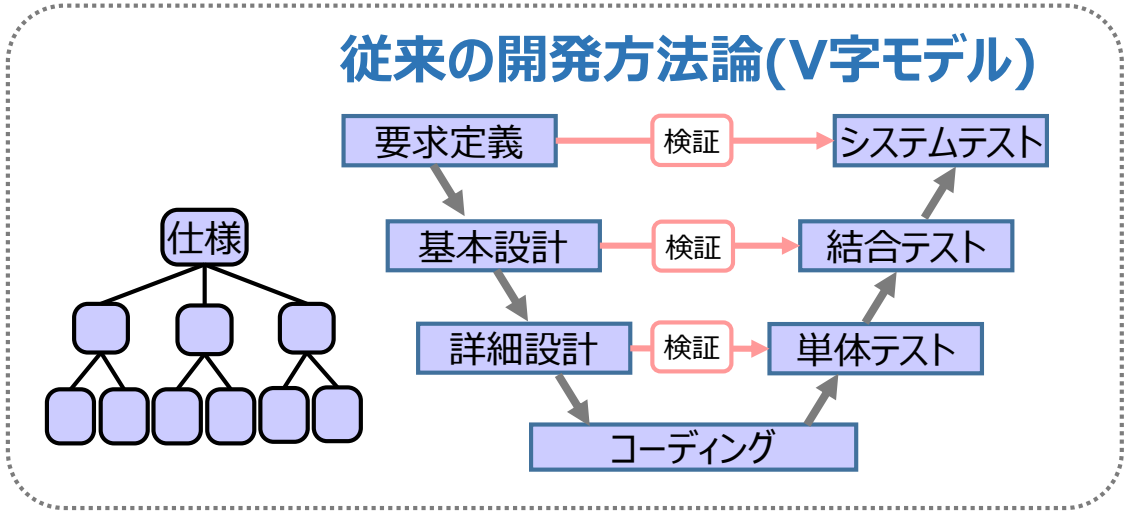
[ニュースソース]
https://jp.techcrunch.com/2016/03/25/20160324microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/

AIへの懸念

AIの品質保証問題

- 仕様(正動作)が定義されないため、テストの成否が定まらない
- 精度100%は無理、間違いは不可避
  →顧客との契約や出荷判定はどうしたらよいのか？
- 動作保証ができないシステムにPL法やリコールの懸念

車向け大型動物飛び出し検知システム
- ボルボが開発して輸出
- シカ等は検知するが、カンガルーは検知できず



例えば、自動運転の環境認識では
- 雨や霧のこともあるかも
- 物陰から人が飛び出すかも
- マンホールから人が顔を出すかも

✓ 機械学習では、仕様(正動作)が定義されない
✓ 機械学習では、テスト結果を積み上げられない
(少しの変更の影響が全体に及ぶ)

どれだけのケースをテストしておいたら、十分安全なシステムだと言えるのか？

システム開発方法のパラダイム転換

■ 従来の演繹的な作り方に対して、機械学習は帰納的な作り方になる

■ 従来の開発方法論が通用せず、新しい方法論が必要

システムの 演繹的な 作り方

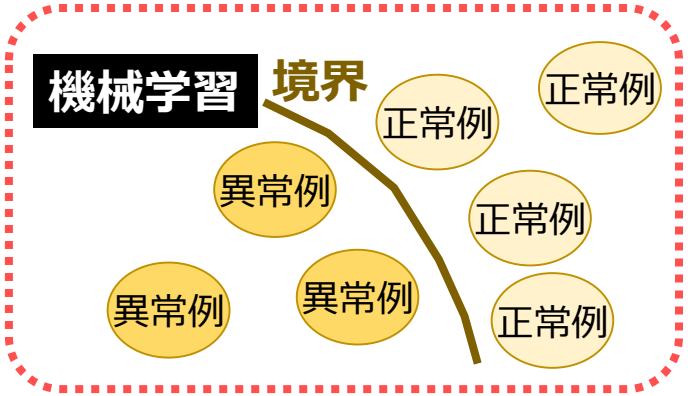
if 条件a then 処理A
else 処理B

条件a: 温度センサ＞90℃
(異常と判定する条件)

**手続きや判定ルールを
明示的に書くことで
システムの動作が決まる
(プログラミング)**

パラダイム転換！

システムの 帰納的な 作り方



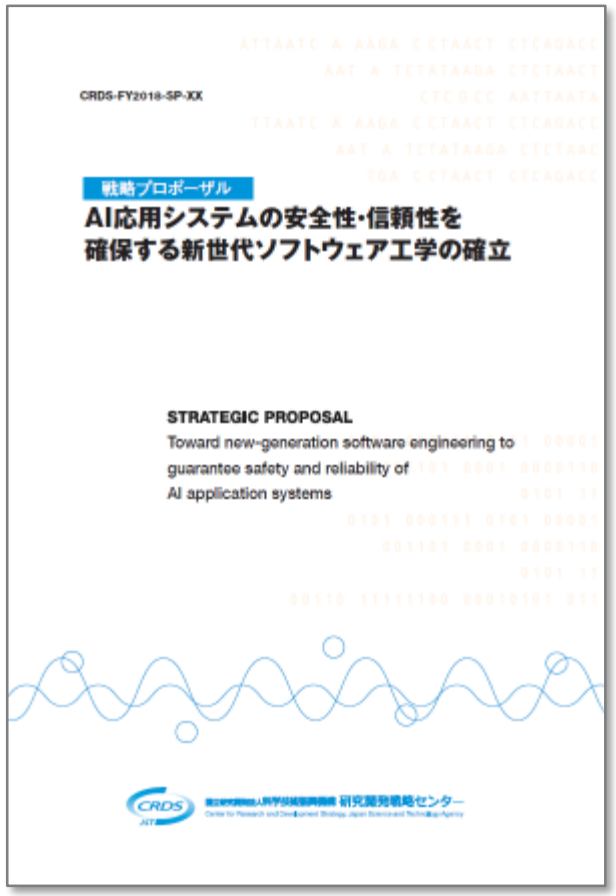
**データを例示すると、それを真似て
自動的に判定ルールができ、
システムの動作が決まる
(機械学習)**

例のないケースで
どんな動作をするか
わからない

仕様(正動作の定義)が
ないため、テストの
成否がわからない

Changing Anything
Changes Everything
学習データを1つ足すだけで
振る舞い全体が変わり得る

JST戦略プロポーザル「AIソフトウェア工学」

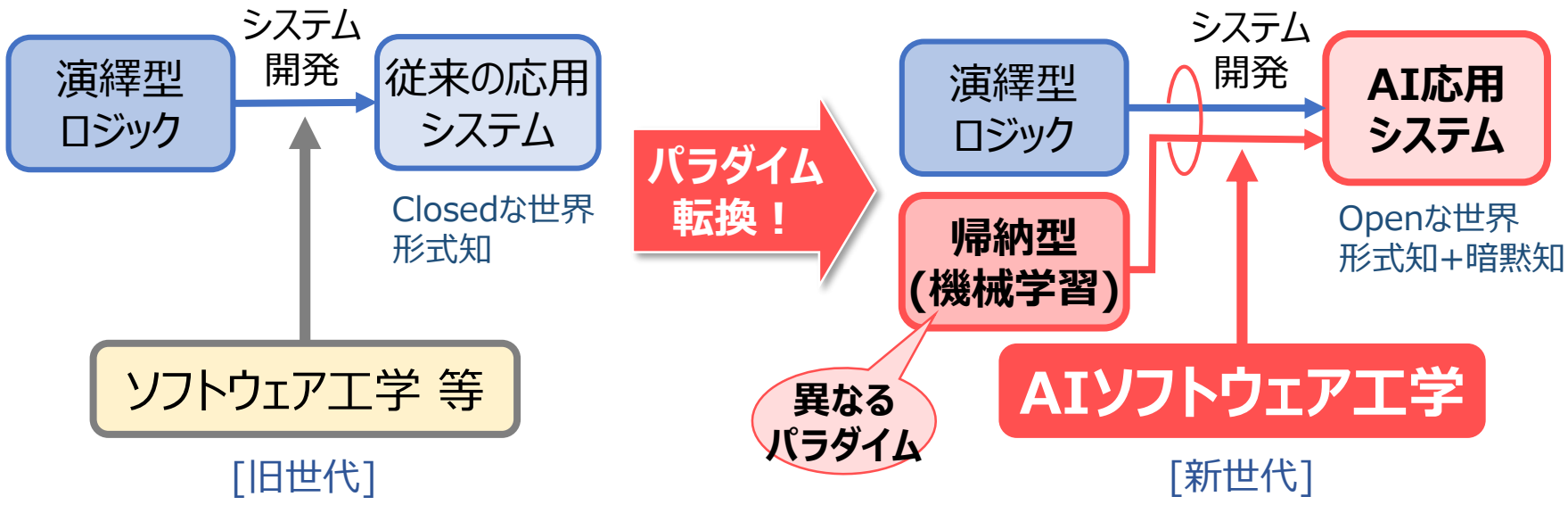


全文pdfダウンロード可能

http://www.jst.go.jp/crds/report/report01/CRDS-FY2018-SP-03.html

■ システム開発のパラダイム転換に対応し、急速に普及するAI応用システムの安全性・信頼性を確保するため、新世代のソフトウェア工学を確立すべきと提言

「AIソフトウェア工学」「機械学習工学」「Software 2.0」等と呼ばれる



AI品質を強みとして、日本の産業競争力強化を狙う

動向(1) 問題意識の急速な高まり

- 4-5年前から兆しが見られ、2017年初頭から産業界の問題意識、学会・業界イベントでの基調講演・企画セッション等で一気にホットトピック化
  - ソフトウェアジャパン2017招待講演、ソフトウェア工学シンポジウムSES2017基調講演、JSAI2018企画セッション、ISCA2018招待講演、他多数
- 2018年は研究コミュニティやコンソーシアムが発足、具体的な活動活発化
  - 日本ソフトウェア科学会に機械学習工学研究会MLSE発足(2018年4月)
  - AIプロダクト品質保証(QA4AI)コンソーシアム発足(2018年4月)
  - カナダにSEMLAイニシアティブ(Software Engineering for Machine-Learning Applications)発足、国際シンポジウム開催(2019年5月)
- JSTでの戦略プロポーザル活動は2017年9月にスタート(翌年12月発行)
- 統合イノベーション戦略推進会議「AI戦略2019(有識者提案)」
  - 2018年9月ドラフト版に「AI工学」

動向(2) 具体的な取り組み成果が徐々に

- 米国DARPA(国防高等研究計画局)による研究開発投資
  - XAI (Explainable AI): 2017年5月から4年間
  - Assured Autonomy: 2018年5月から4年間
- 品質クリティカルな自動運転分野での取り組みが先行
  - PEGASUS: ドイツBMWi(経済エネルギー省)主導による産学官共同プロジェクト
  - デンソーAI品質基盤
  - 宇宙航空研究開発機構(JAXA)と日本自動車工業会(JAMA)の共同検討
  - JST ERATO蓮尾メタ数理システムデザインプロジェクト(2019年5月 成果発表)
- 2019年5月Open QA4AI Conference: 国内5団体が成果発表
  - QA4AIコンソーシアム、産総研(NEDOファンド)、中部経済産業局 戦略的基盤技術高度化支援事業、JAXA、国立情報学研究所QAML(JSTファンド)
- AI品質に関わる国際標準化、AI品質の認証機関設立への動きも

国立研究開発法人
新エネルギー・産業技術総合開発機構

「次世代人工知能・ロボット中核技術開発」プロジェクト

2019年度公募「**人工知能の信頼性に関する技術開発**」

https://www.nedo.go.jp/koubo/CD2_100157.html

①説明できるAI

人工知能を安心して社会で利活用するため、人工知能の信頼性を確保する基盤技術として、人工知能の学習内容や推論結果、判断根拠等を人に理解しやすい形で可視化する「説明できるAI」の適応分野を明確にした上で、研究開発を実施します。

②AI品質

人工知能を利活用したソリューションの品質が課題となるシーンにおいて、必要となるツールの開発や評価方法の確立を実施します。また、それらの他の人工知能品質の課題に対しても利活用できるように汎用化・体系化します。

- 事業期間: 2019～2023年度（5年間）
- 事業規模: ①は先導研究1年間・単年度上限40百万円×複数採択（2020年度に改めて公募）、②は2019年度約140百万円以内で単独または複数採択予定
- 2019年度の公募期間: 3/29～5/8

国立研究開発法人
科学技術振興機構　未来社会創造事業「超スマート社会」領域

2019年度新規重点公募テーマ

http://www.jst.go.jp/mirai/jp/open-call/research/r01/

「サイバーとフィジカルの高度な融合に向けたAI技術の革新」

Society 5.0で想定されるシステムやアプリケーションの実現に向けて、説明可能AIによる利用者側の主体性・受容性の担保、安全性・信頼性の担保。変化への対応能力、リアルタイム性の要求される環境での実行速度と精度の確保など、AI技術の実適用におけるさまざまな課題を解決する革新的なAI技術の開発を目標とします。

①帰納型と演繹型を融合した説明可能なAIの実現

②変化する事象へ対応できる追加学習やオンライン学習等の学習メカニズムの構築

③実世界の事前知識としてのモデルの組み込み等による少数データからの学習の実現

④学習済みニューラルネットからのモデル構造の抽出

⑤動作・信頼性保証や信頼性を保証したシステムの開発技術の確立

⑥分散環境において個別の学習結果を統合・融合する学習アルゴリズムの開発

⑦学習アルゴリズムの高度化による抜本的な省電力化・高速化

など

- 探索研究フェーズ：最大2年半、研究開発費 総額最大35百万円
- 本格研究フェーズ：厳正な審査により探索研究期間終了後に本格研究に移行可能、本格研究期間は最大5年、研究開発費は総額最大7.5億円
- **2019年度の公募期間：5/15～7/24正午**

機械学習における
説明可能性・公平性・安全性への工学的取り組み

講演1	15分	**イントロダクション** －JST戦略プロポーザルの紹介－	福島 俊一 科学技術振興機構
講演2	25分	**機械学習の説明可能性への取り組み** －DARPA XAIプロジェクトを中心に－	川村 隆浩 科学技術振興機構
講演3	25分	**機械学習の公平性への取り組み** －Fairness-aware data miningを中心に－	神嶌 敏弘 産業技術総合研究所
講演4	25分	**機械学習の安全性への取り組み** －自動車業界の取り組みを中心に－	中江 俊博・桑島 洋 デンソー
講演5	10分	**MLSE研究会・QA4AIコンソーシアム** の活動・成果物報告	石川 冬樹 国立情報学研究所、機械学習工学研究会MLSE主査