

理事長記者説明会

人間中心社会に向けたAI研究開発の潮流 ～安全・高品質AIを日本の競争力に～

2020年1月22日

JST研究開発戦略センター
システム・情報科学技術ユニット
フェロー 福島 俊一

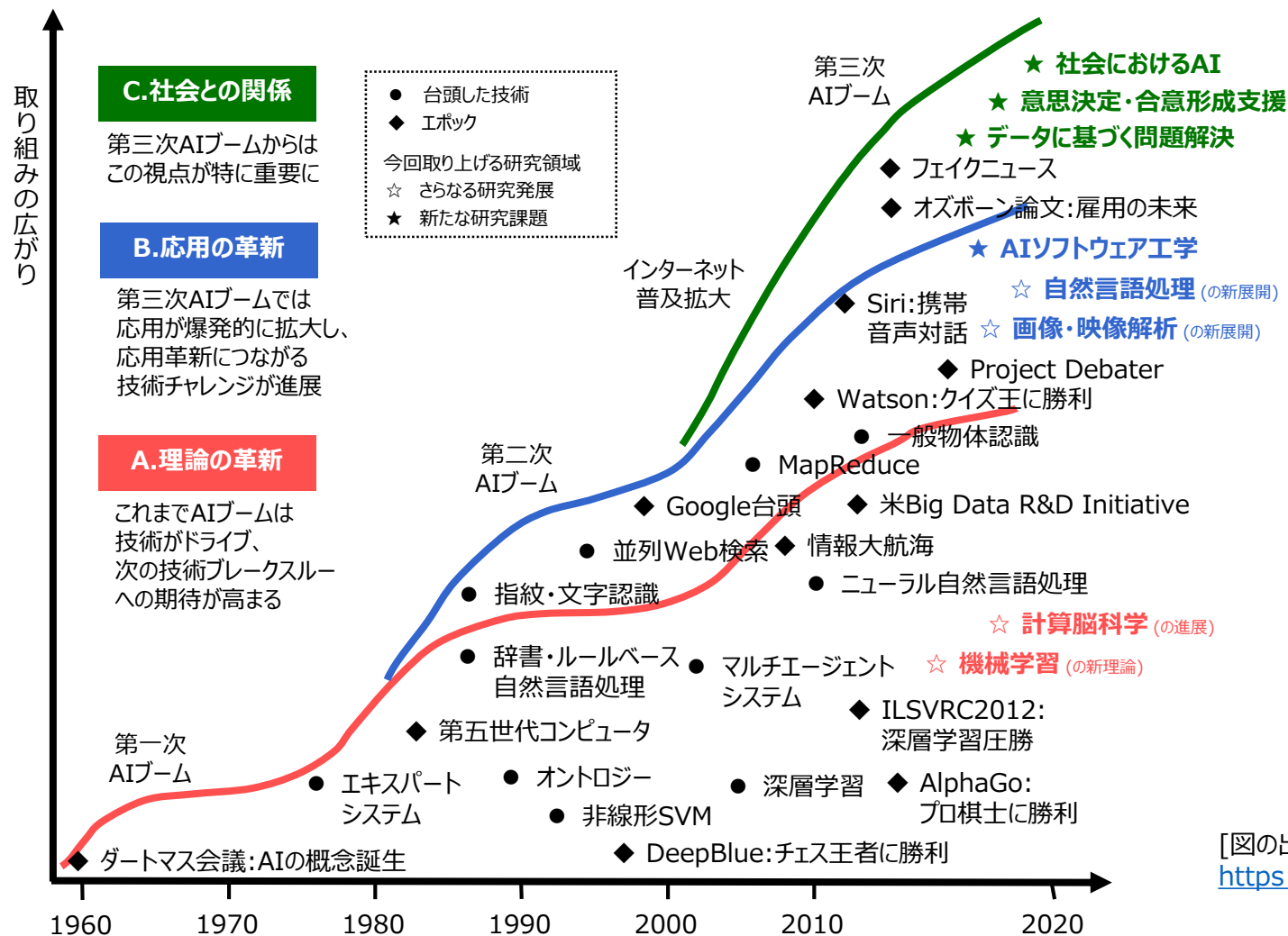
https://researchmap.jp/toshikazu_fukushima/



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

本日着目するAI研究開発の新しい潮流

人工知能(AI)・ビッグデータ技術の時系列俯瞰図



AIの社会との関係が拡大したことで生まれた
精度・性能向上とは異なる研究開発の潮流

ここに着目!

主にAI技術の
精度・性能の向上のための技術開発

[図の出典] 研究開発の俯瞰報告書 システム・情報科学技術分野(2019年)
<https://www.jst.go.jp/crds/report/report02/CRDS-FY2018-FR-02.html>

本日のご説明のアウトライン



■ AI技術が社会に急速普及

- 深層学習技術による飛躍的な精度向上、第3次AIブームを牽引



■ AI技術が社会に広がる中で生まれた新たな技術課題

- ブラックボックス問題、バイアス問題、脆弱性問題、品質保証問題、フェイク問題等



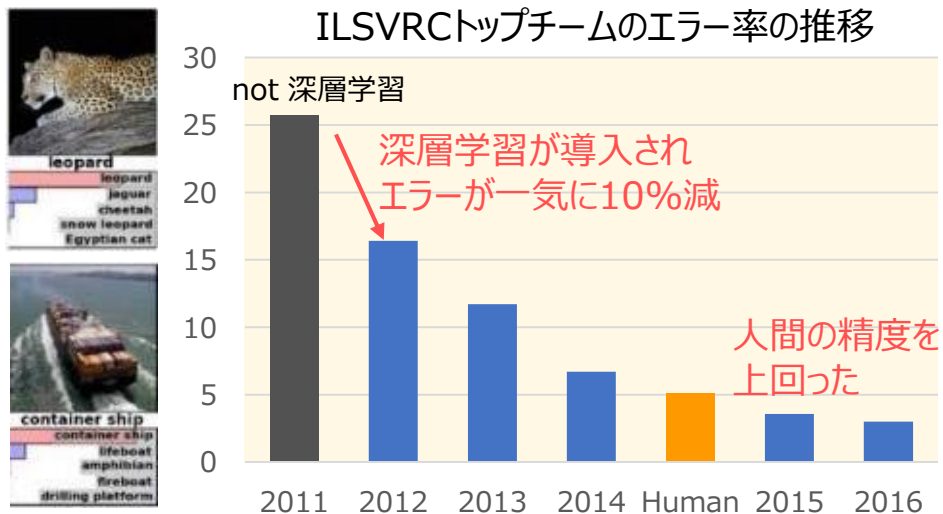
■ 安全で高品質な信頼されるAIを目指した取り組み

- 人間中心社会を支える信頼されるAIの実現へ
- 安全・高品質AIを日本の競争力に

第3次
AIブーム

人工知能(AI)技術の進展と普及

- 様々な特定用途で人間を上回る精度を達成
- 「前例に倣うAI」機械学習技術が進化（特に深層学習技術）
- 様々な応用にAI・機械学習技術が搭載されて社会に急速普及

一般物体認識競技会ILSVRCで
深層学習が衝撃的な性能改善

[元情報] ILSVRC <http://www.image-net.org/challenges/LSVRC/>

AlphaGoが世界トップ棋士に圧勝

AlphaGo: Google DeepMindが開発した囲碁プログラム(モンテカルロ木探索 + 深層学習・深層強化学習)

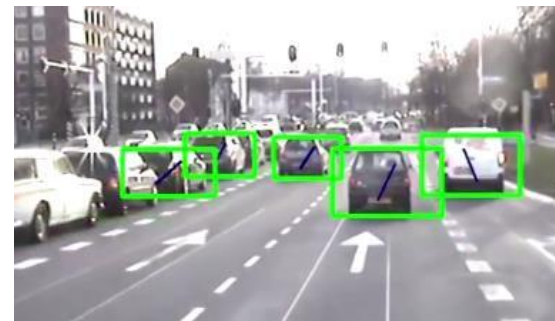


- 2016年3月 李世乭に4勝1敗
- 2017年5月 柯潔に3局全勝

[図の出典] <https://youtu.be/qDW0dhcOBhA>

様々なAI応用の実用化

判別・分類 映像監視、状況認識等



[図の出典] https://youtu.be/F_M_skebbpA

予測 需要予測、劣化予測等

異常検知 故障予兆検知、不正検知等

変換・生成 自動翻訳、自動着色等

第3次
AIブーム

AIブームは衰えず!?

[参考] <https://japan.zdnet.com/article/35126917/>

- AIに対する期待はピークを越え、幻滅期に入りつつあるとの見方もあるが、
- 国内外でAI関連学会の参加者数は依然として爆発的に増加

抽選

NeurIPS2019:
約13,000名

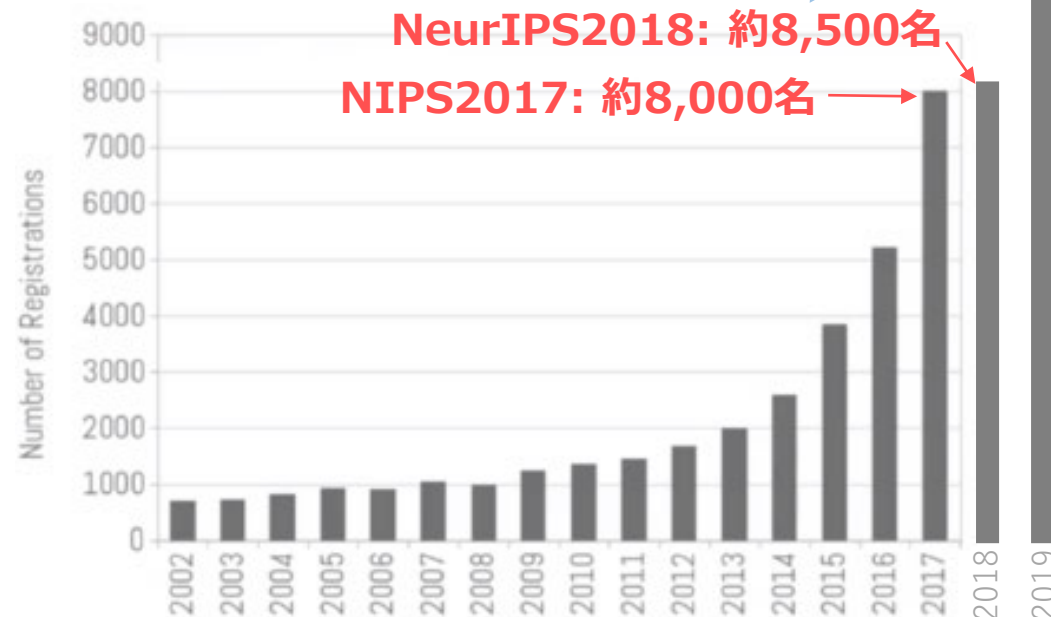
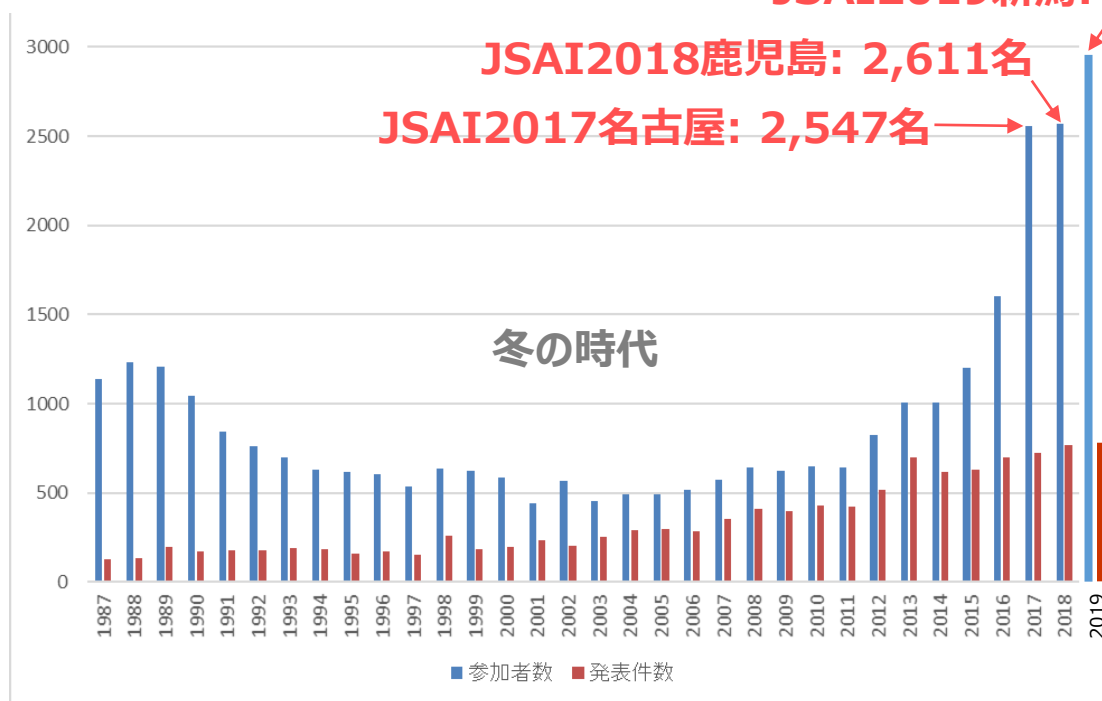
【国内】人工知能学会の全国大会(毎年6月頃に開催)

【海外】機械学習分野の最高峰の国際会議(毎年12月頃に開催)

人工知能学会全国大会の参加者数と発表件数

JSAI2019新潟: 2,905名

NIPS/NeurIPSの参加登録者数

11分38秒で
売り切れ

AIへの懸念

AI×社会に関わる問題、新たな技術課題

- AI技術の精度・性能向上によるご利益の一方、AI×社会に関わる問題が顕在化し、取り組まねばならない新たな技術課題が生まれている

第3次
AIブーム

- ✓ 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への懸念

- ✓ AI代替によって消える職業
- ✓ 監視社会、プライバシー不安
- ✓ AIの軍事利用、AIの悪用
- ✓ AIの判定結果を信じられるか
- ✓ 人間の意思決定の主体性を確保できるか

社会におけるAIのあるべき姿



AI社会原則

を国・世界レベルで議論・策定する動き

技術的取り組み強化が必要な
これらの問題については詳細後述

- ① AIのブラックボックス問題(説明責任)
- ② AIのバイアス問題(公平性)
- ③ AIの脆弱性問題
- ④ AIの品質保証問題
- ⑤ フェイク動画問題
- ⑥ ソーシャルメディアによる思考誘導問題

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

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 → 実現に向けて技術課題への取り組みが必要

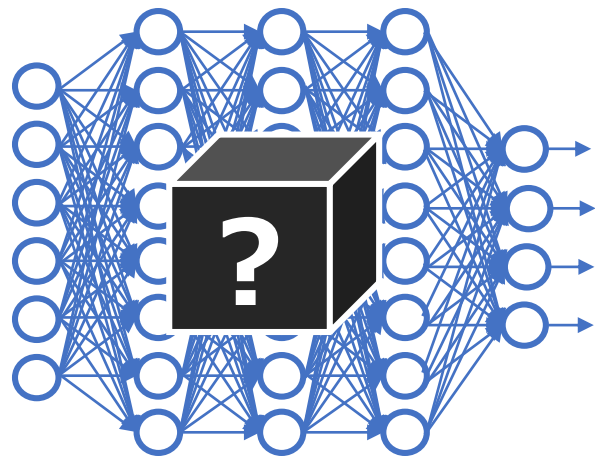
人間の尊厳が尊重される社会(Dignity)、多様な背景を持つ人々が多様な幸せを追求できる社会(Diversity & Inclusion)、持続性ある社会(Sustainability)という3つの価値を基本理念とし、「AI-Readyな社会」をビジョンに掲げる

AI社会原則

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

①AI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



学習結果は、多層ニューラルネットワーク中の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

あなたは不採用です

あなたは不審者です

このプロジェクトは失敗します



なぜ？
何がいけないの？



事故は私のミス？
それともAIのバグ？

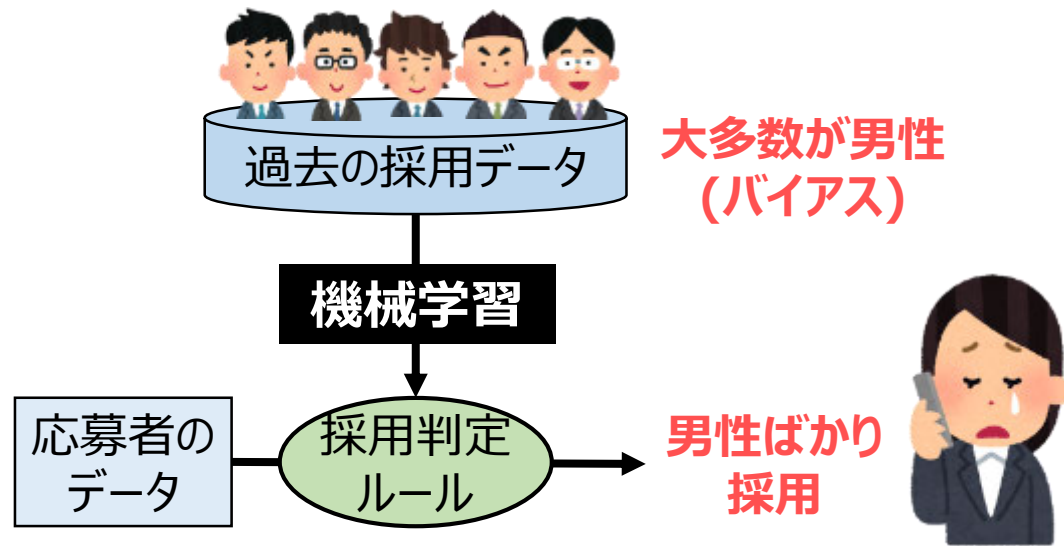
欧州GDPRは AIの透明性を要求

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

②AI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



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

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

③AI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、人間は気につか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/>

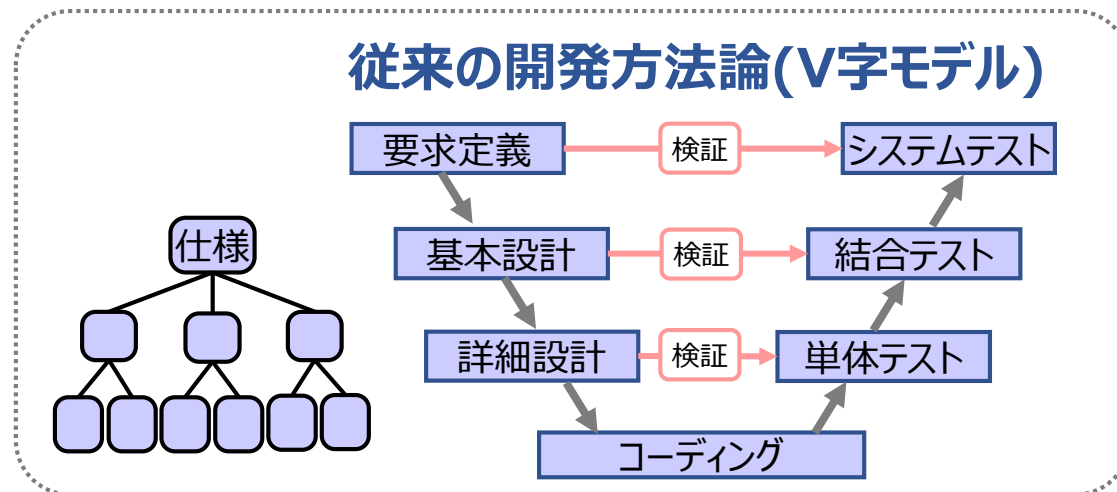
④ AIの品質保証問題

- 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
- [参考]

<https://japanese.engadget.com/2017/07/05/volvo-self-driving-cars-kangaroo/>



- ✓ 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)



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

**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？**

⑤ フェイク動画問題

- 敵対的生成ネットワーク(GAN*)等の技術進化によって、本物と見紛うフェイク動画等の生成が可能になってしまった
- フェイク動画は人々に信じられやすく、世論誘導・政治干渉の危険大
- 写真・動画は証拠として信じられるのか?

* GAN: 生成器(偽物作り)と識別器(見破る役)が相互に切磋琢磨して真に近いものを生成

GANによって自動生成された
有名人風の顔画像

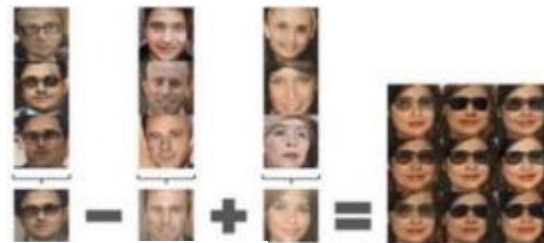


<https://youtu.be/XOxxPcy5Gr4>

Deepfakes

動画の顔を別人に置き換えて、極めて自然なフェイク動画を生成するツール

画像の演算



男性・眼鏡あり - 男性・眼鏡なし + 女性・眼鏡なし = 女性・眼鏡あり

<https://arxiv.org/pdf/1511.06434.pdf>

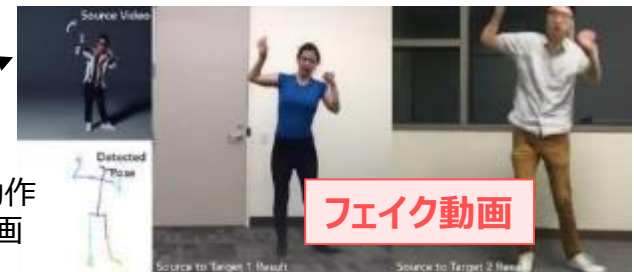
政治家に思い通りの
発言をさせるフェイク動画

動作もコピーできる

オリジナル動画は
プロのダンサーが
踊っている動画

このプロのダンサーの動作
を別人にコピーした動画
ができる

<https://youtu.be/PCBTZh41Ris>



フェイク動画



フェイク動画

もとはこちらの
人が話している

⑥ソーシャルメディアによる思考誘導問題

■ ソーシャルメディアを通じた思考誘導・世論操作の危険性が増大

デジタルゲリマンダー*の脅威

*ソーシャルメディアを用いた政治操作行為

2016年の米国大統領選挙の際には、フェイクニュースやフェイク広告が用いられ、選挙結果が左右されたとも言われる



ローマ法王がトランプ支持を公式に表明した



民主党候補のヒラリー・クリントンは、テロ組織IS（自称イスラム国）に武器を売却した



児童性愛の地下組織がワシントンDCのピザ屋に存在し、ヒラリーが関与している **発砲事件に**



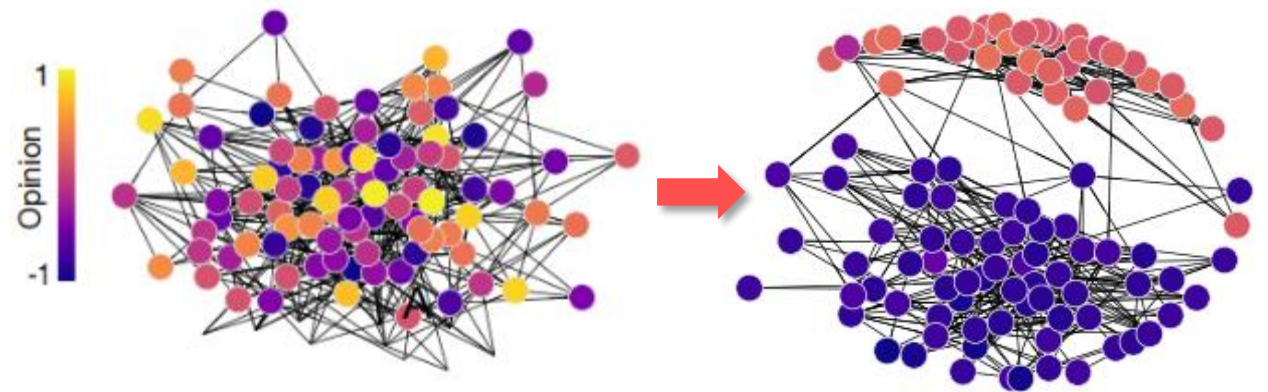
Facebook上に、銃規制、同性婚、移民問題等、米国社会の分断をあおるような政治広告を掲載 ※フェイク広告は個人別なので追跡が難しい

[ニュースソース]

<https://seiji-yama.jp/article/news/nws20161202.html>

<http://gendai.ismedia.jp/articles/-/50786>

SNSの同調圧力・エコーチェンバー現象が価値観の対立を激化し、社会を分極化



[図の出典] 笹原 JSAI2017

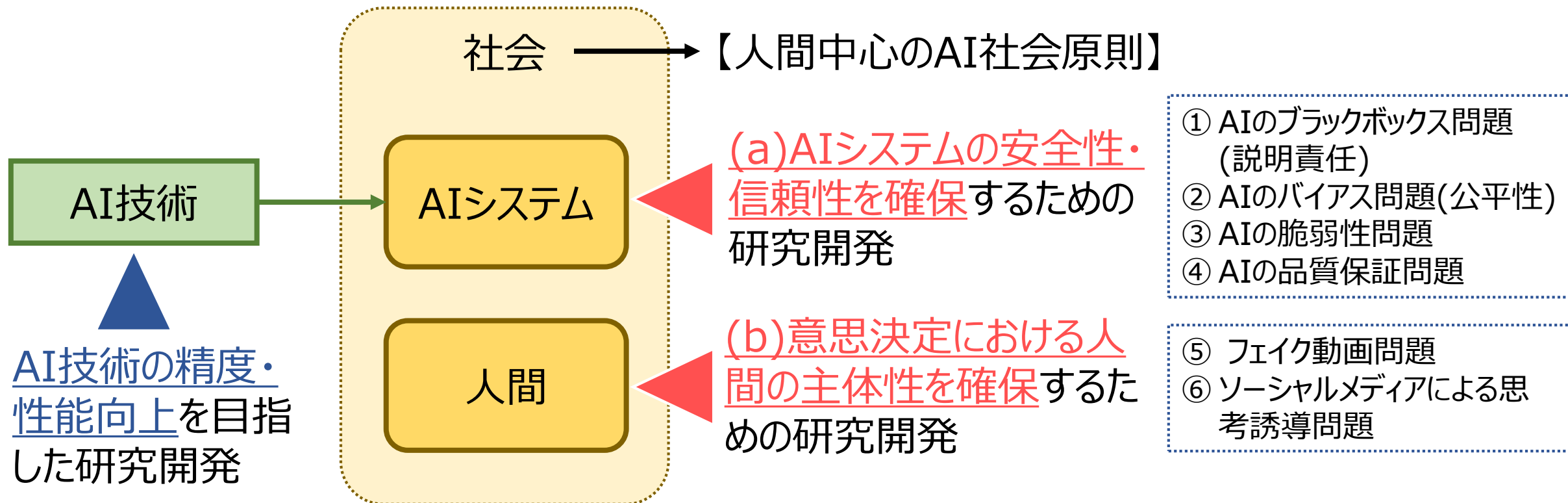
https://doi.org/10.11517/pjsai.JSAI2017.0_4N1OS01a1

SNSにおけるフェイクニュースの拡散やエコーチェンバー的な応答のために、ソーシャルボット*が活用されることが多々ある

*ソフトウェアによって制御されたソーシャルメディア上のアカウント

人間中心社会に向けたAI研究開発

- AI技術の精度・性能向上を目指した研究開発だけでなく、人間中心のAI社会原則を満たすための研究開発への取り組みが求められる





背景 システム開発のパラダイム転換が起きている

- 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

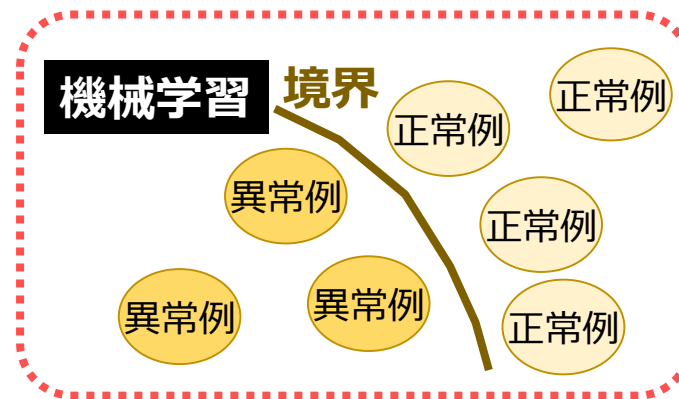
if 条件a then 処理A
else 処理B

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

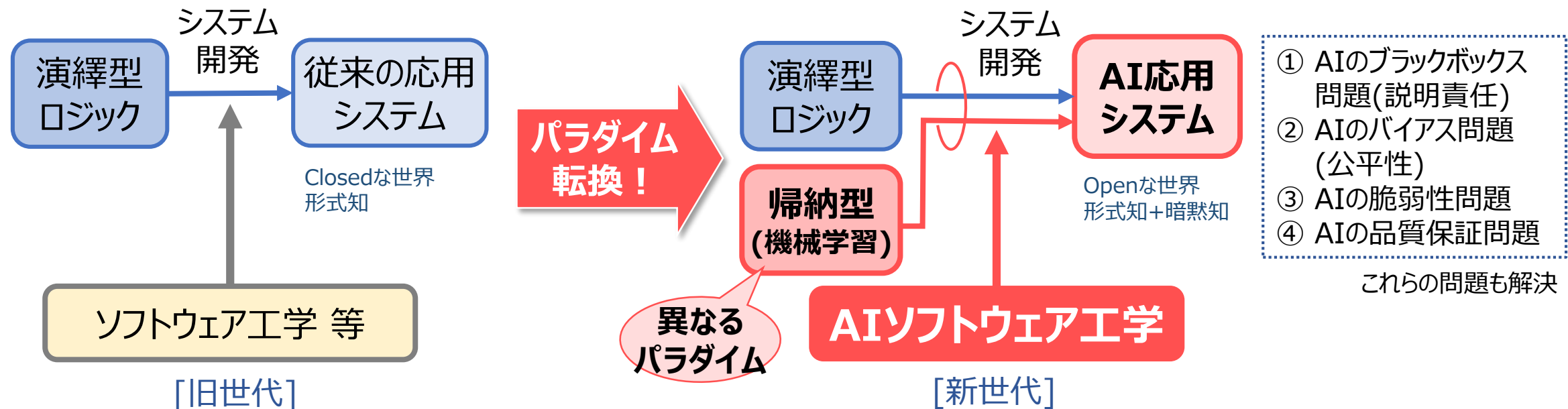
(a) AIシステムの安全性・信頼性を確保するための研究開発



AIソフトウェア工学の確立へ

- 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

「AIソフトウェア工学」 国内では「機械学習工学」
海外では「Software 2.0」等とも呼ばれる





研究開発の動向

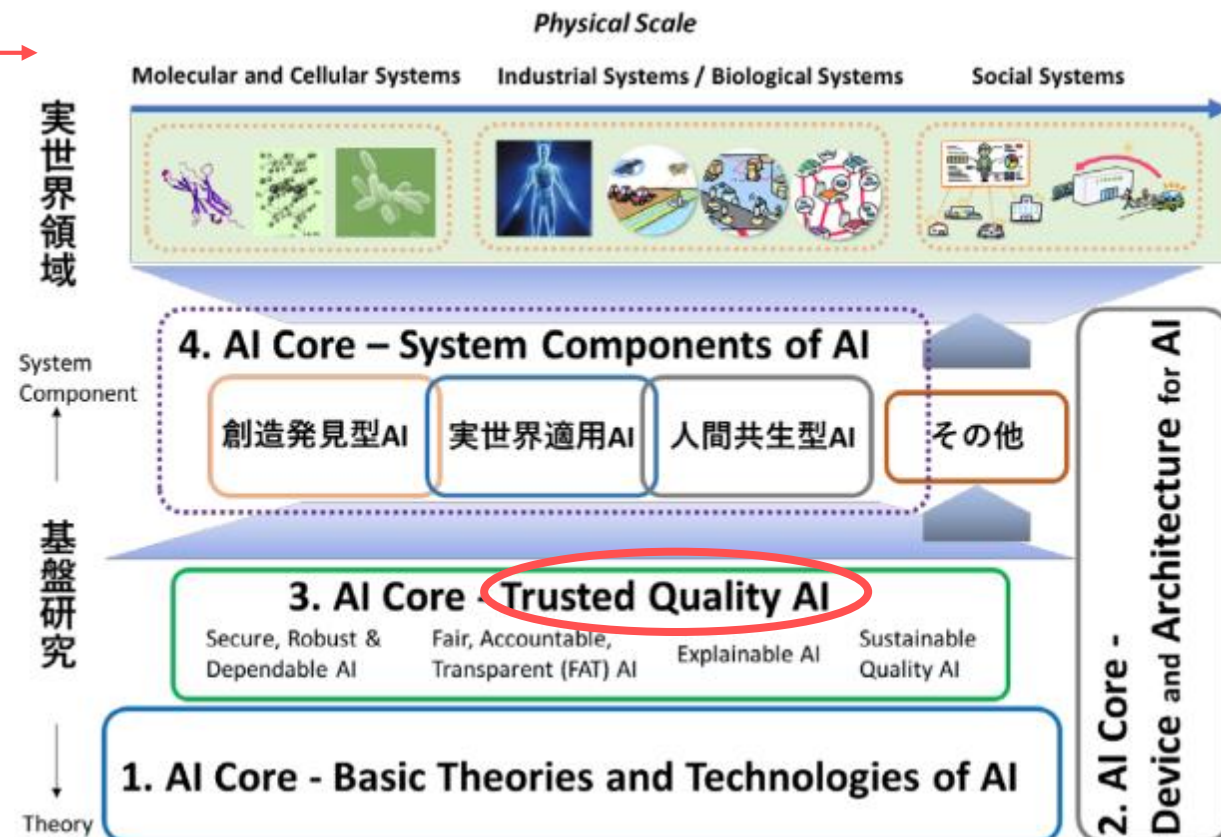
- 学会・業界で問題意識が高まり、研究コミュニティやコンソーシアムが発足
 - 国内では、日本ソフトウェア科学会に機械学習工学研究会MLSEが発足、AIプロダクト品質保証(QA4AI)コンソーシアムも発足(ともに2018年4月)
 - カナダにSEMLAイニシアティブ発足、国際シンポジウム開催(2019年5月)
- 米国DARPA(国防高等研究計画局)が研究開発投資
 - XAI (Explainable AI): 説明可能なAI、2017年5月から4年間
 - Assured Autonomy: 自律システムの安全性、2018年5月から4年間
- 日本政府は統合イノベーション戦略推進会議でAI戦略2019を決定(2019年6月)、研究開発の方向性としてTrusted Quality AIを掲げた
- 品質クリティカルな自動運転分野での取り組みが先行
 - ドイツBMWi(経済エネルギー省)のPEGASUSプロジェクト、デンソーのAI品質基盤等
- 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

日本政府「AI戦略2019」

		主な具体目標	主な取組
未来への基盤作り	教育改革	デジタル社会の「読み・書き・そろばん」である「数理・データサイエンス・AI」の基礎などの必要な力を全ての国民が育み、あらゆる分野で人材が活躍	- リテラシー：外部人材の積極登用、生徒一人に端末一台 - 応用基礎：AI×専門分野のダブルメジャーの促進 - エキスパート：若手の海外挑戦拡充、AI実践スクール制度 - 優れた教育プログラムを政府が認定する制度の構築
	研究開発	- 世界の英知を結集する研究推進体制 - 日本がリーダーシップを取れるAI技術 - AI研究開発の日本型モデルの構築	- 多様な研究者による創発研究の支援拡充 - 世界をリードできる次世代AI基盤技術の確立 - AI中核センター改革、AI研究開発ネットワーク構築
産業・社会の基盤作り	社会実装	- 実世界産業のサービス構造への転換 - インクルージョン・テクノロジーの確立 - 標準化を推進し、開発成果の社会実装を促すシステム・アーキテクチャを先導	- 健康・医療・介護：世界の医療AIハブ、データ基盤整備 - 農業：スマート農業技術の現場導入、成長産業化 - 国土強靱化：インフラデータプラットフォームの構築 - 交通・物流：AIターミナルの実現、物流関連データ基盤構築 - 地方創生：スマートシティ共通アーキテクチャの構築
	データ関連基盤	国際連携による次世代AIデータ関連インフラの構築	- データ基盤：データ基盤の本格稼働と連携 - トラスト：トラストデータ流通基盤の開発
	デジタル・ガバメント 中小・新興企業支援	- 公共サービス・自治体行政のコスト削減、業務効率化 - AIを活用した中小企業の生産性向上	- 自治体が安心して利用できるAIサービスの標準化 - 中小企業支援方策の検討
	倫理	AI社会原則	「人間中心のAI社会原則」の定着化、多国間枠組構築

<http://www.maff.go.jp/j/kanbo/tizai/brand/attach/pdf/ai-15.pdf>

「Trusted Quality AI」が
取り組むべき基盤研究の中核に位置付けられている
AI関連産業における日本の強みになると期待できる



<https://www.kantei.go.jp/jp/singi/tougou-innovation/pdf/aisenryaku2019.pdf>



取り組み事例 (1/3) 説明可能なAI

- 従来、機械学習の精度と説明可能性はトレードオフ関係
- (A)ブラックボックス型を外側から調べて近似的説明を試みる手法と、(B)もと説明可能なホワイトボックス型の精度を高める手法が開発されている

(A)の例

入力データの近傍で様々に変化させたデータを生成して、ブラックボックスに投入し、その結果の変化を観測することで、変化に大きく影響する要素を特定する

LIME、SHAP

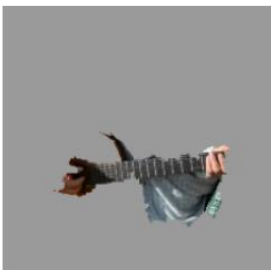
どの特徴量が結果に大きく影響したかを調べる

Influence

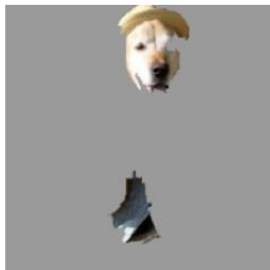
どの教師データが結果に大きく影響したかを調べる



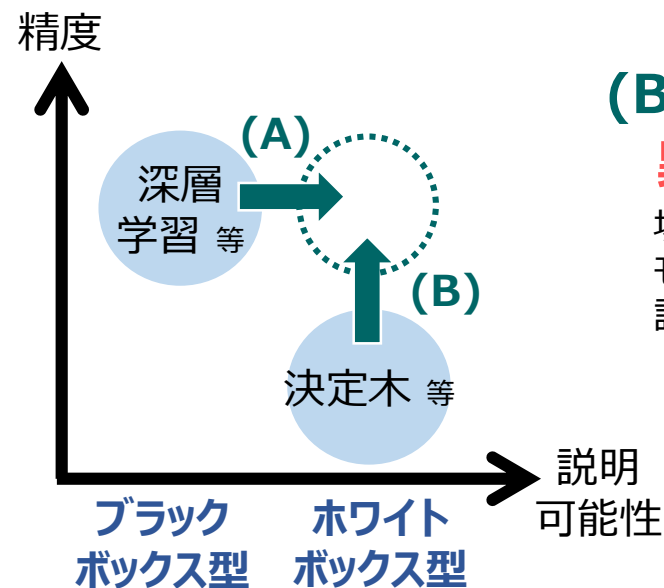
(a) Original Image



(b) Explaining Electric guitar



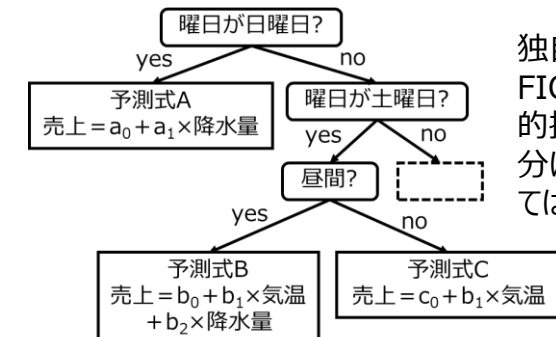
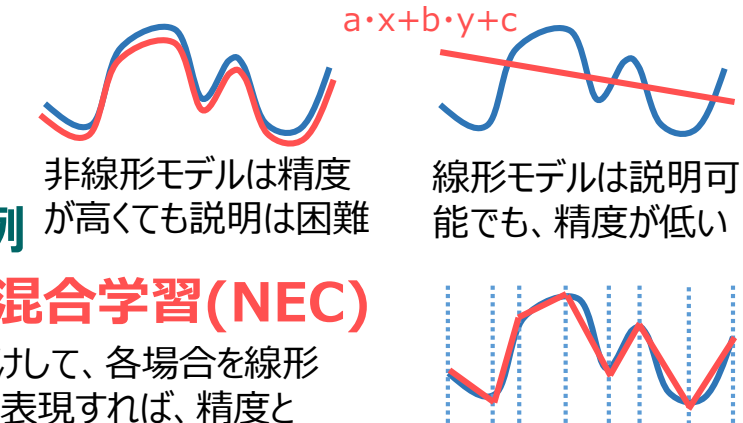
(d) Explaining Labrador



(B)の例

異種混合学習(NEC)

場合分けして、各場合を線形モデルで表現すれば、精度と説明可能性を両立できる



独自の情報量基準 FICを用いた適応的探索により、場合分けと予測式の当てはめを最適化



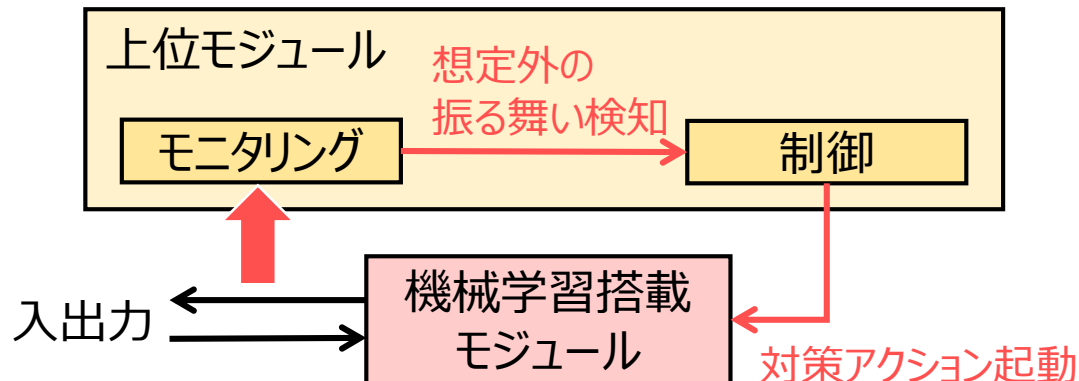
取り組み事例 (2/3) AIの品質・安全性の確保

- 機械学習パートに100%保証を求めることはできないため、機械学習型と従来型を組み合わせたシステム全体として安全性を確保
- 機械学習パートについては、起こり得るケースの網羅は不可能だが、できる限り効率よくテストケースのカバレッジを高める

DARPA Assured Autonomy

<https://www.darpa.mil/program/assured-autonomy>

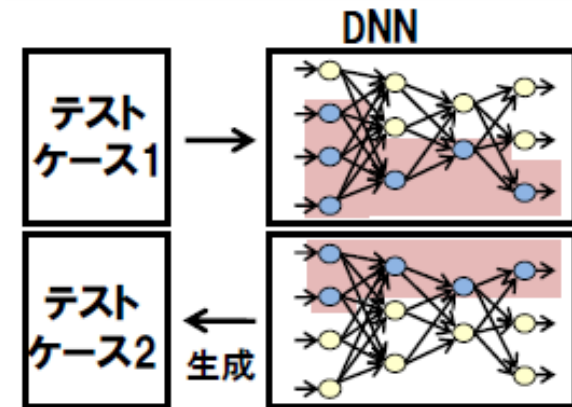
機械学習パートの実行時の入出力をモニタリングし、想定される振る舞いから外れた時に、それを検知して動作を止める等の対策アクションを起動する



ニューロンカバレッジ方式

(コロンビア大のDeepXplore/DeepTest 等)

- 欲しいものを表すスコアを高めるようにテストケースを生成するサーチベースドテストングを応用
- ニューラルネットの活性化範囲の網羅率を高めるようにテストパターンを生成



[図の出典] 明神(日立)

<http://jasst.jp/symposium/jasst18tokyo/pdf/C5-1.pdf>

他に、システム全体として安全性を確保するアイデアとしては、Safe Learning、Safety Envelope等、テストケースのカバレッジを高めるテスト手法としては、サーチベースドテストング、メタモルフィックテストング、データセット多様性等も知られている。



取り組み事例 (3/3) 業界の組織的な取り組み

- より広くAI応用システム開発プロセス全般を整備する活動が、業界内で立ち上がっている

QA4AIコンソーシアム

AIプロダクト品質保証ガイドライン 2019.05版

ダウンロード可能 <http://www.qa4ai.jp/download/>

AIプロダクトの品質保証に対して共通の指針を与える
(ただし、AI技術は発展途上であることから、網羅性や完全性は企図されていない)

(1) 5つの評価軸(チェックリスト)とそのバランスで品質保証活動を評価

- ✓ **Data Integrity**: データはきちんとしているか?
- ✓ **Model Robustness**: モデルはきちんとしているか?
- ✓ **System Quality**: システム全体として価値が高く、何かが起きてても何とかなるか?
- ✓ **Process Agility**: プロセスは機動的か?
- ✓ **Customer Expectation**: (良くも悪くも)顧客の期待は高いか? 合致しているか?

(2) 技術カタログ

(3) 分野別ガイドライン: 自動運転、産業用プロセス、スマートスピーカー、生成系、(今後さらに拡大)

機械学習工学研究会MLSE

機械学習を用いたシステムの要件定義から設計・開発・運用まで、プロセス管理や開発環境・ツール、テスト・品質保証の手法、プロジェクトマネジメントや組織論も含めて、開発プロセス全般をスコープに含めた包括的な取り組みを推進



情報処理学会誌 2019年1月号
「機械学習工学」特集

特集

機械学習工学

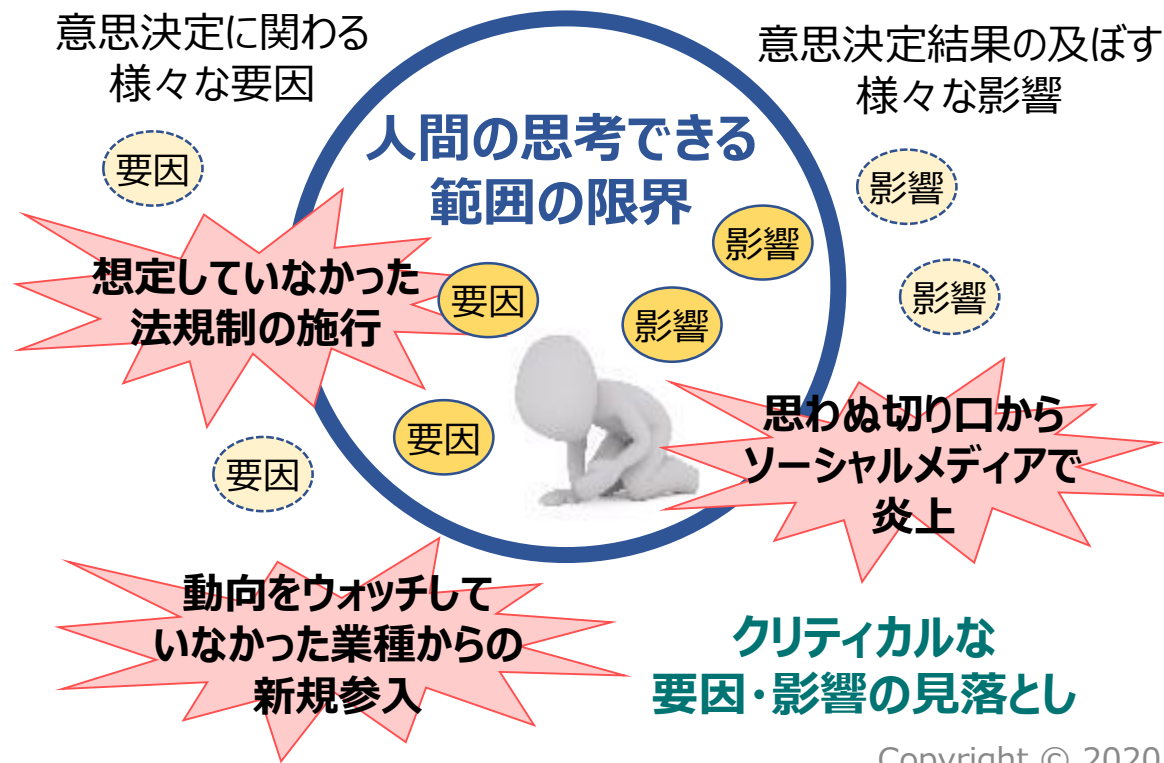
- | | | | |
|----|----|----------------------------|------------|
| 10 | 0. | 編集にあたって | 野々山尊秀・丸山 宏 |
| 12 | 1. | ■ 機械学習工学の狙いと展開 | 丸山 宏 |
| 17 | 2. | ■ 機械学習応用システムの開発・運用環境 | 今井健男・太田満久 |
| 25 | 3. | ■ 機械学習応用システムのテストと検証 | 石川冬樹・徳本 晋 |
| 34 | 4. | ■ 機械学習応用システムのセキュリティとプライバシー | 吉岡信和 |
| 40 | 5. | ■ 機械学習のためのヒューマンインタフェース | 五十嵐健夫 |
| 48 | 6. | ■ 機械学習応用システムのプロジェクト管理と組織 | 本橋洋介 |

★国際標準化活動は **ISO/IEC JTC1/SC42** で取り組まれている

背景 人間の主体的意思決定が難しくなってきた

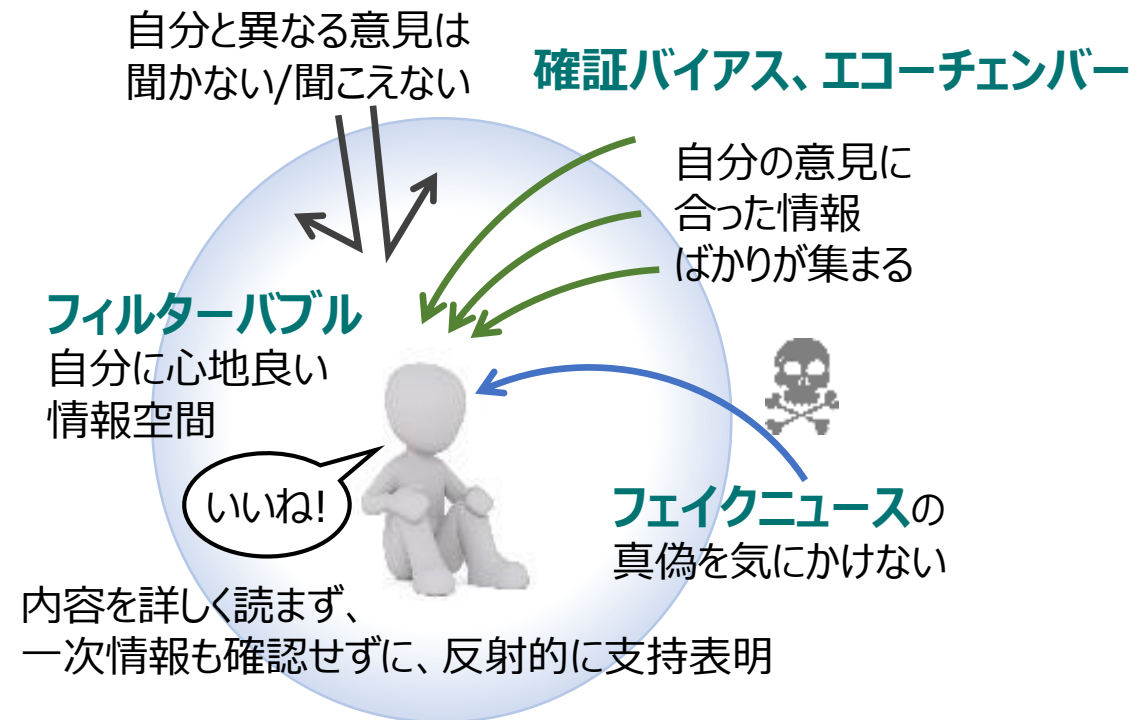
限定合理性の問題が深刻化

- 情報爆発やボーダレス化によって、意思決定に関わる要因・影響の可能性が膨大に
- クリティカルな要因・影響の見落としが深刻化



確証バイアスの問題が顕著化

- ソーシャルメディアが普及し、フィルターバブルが作られやすくなっている
- 確証バイアスにより、フェイクが信じられやすい



(b)意思決定における人間の主体性を確保するための研究開発



主体的意思決定の困難化対策の方向性

(限定合理性が引き起こす)

(認知バイアスが引き起こす)

- クリティカルな要因の見落としや思考誘導を受けるリスクを、情報科学・AI技術によって軽減
- 様々な可能性や意見の熟慮・熟議に基づく適切で迅速な意思決定、集合知の醸成を促進する、情報科学・AI技術を活用した仕組みを実現

意思決定に関わる
様々な要因・影響の
膨大な可能性を探索

論理性・合理性だけでなく
共感・賢慮に基づく解へ導く

AI・他者からの推薦や
誘導の反射的受け入れでなく
熟慮・熟議に基づく判断を促す

人間が見落としていた
意思決定の選択肢を提示

フェイクニュース、デマ等による
扇動・印象操作等を検知・排除

異なる意見や対立点、
新しい見方等を提示

- ⑤ フェイク動画問題
- ⑥ ソーシャルメディアによる思考誘導問題

これらの問題も解決



寄り添う
エージェント



健全な
プラットフォーム



判断力の
教育・訓練



研究開発の動向

- フェイクニュースが社会問題化し、各国でファクトチェック機関が発足
 - 米国:Snopes (1994年～)、PolitiFact (2007年～)、日本:FIJ (2016年～)等
- 米国DARPA(国防高等研究計画局)が研究開発投資
 - XAI (Explainable AI): 国防上の意思決定のためAIの説明性(2017年～)
 - Media Forensics: フェイク動画検出を含む情報の信憑性・正確性(2016年～)
- AI・ビッグデータ技術による分析ビジネスは処方的分析へ移行
 - 記述的分析→診断的分析→予測的分析→処方的分析 [Gartnerによる分類]
何が起ったのか? なぜ起ったのか? 何が起るのか? 何をすべきか? (自動意思決定or意思決定支援)
 - 機械学習が適するのは評価関数が定まる問題(自動意思決定)、
多様な価値観が混在・対立する問題(意思決定支援)には文理融合アプローチ
- 集団の意思決定・集合知の研究はIT・AIを活用した発展へ
 - 米国MIT集合知研究センター(2006年設立)等



取り組み事例 (1/2) フェイク対策

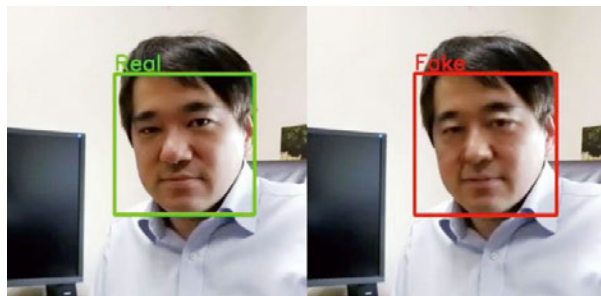
- 動画解析、自然言語解析、ソーシャルメディア解析等、多面的な切り口からフェイク検出や情報の信憑性判定のための技術開発が試みられている

フェイク動画検出

動画の詳細解析によるフェイク判定

<http://ainow.ai/2018/10/15/148590/>

- ✓ 不自然なまばたきの仕方
- ✓ 不自然な頭部の動き、奇妙な目の色
- ✓ 画像から読み取れる人間の脈拍数
- ✓ 映像のピクセル強度のわずかな変化
- ✓ 照明や影などの物理的特性の不自然さ
- ✓ 日付・時刻・場所と天気との整合 等



[出典] 国立情報学研究所 越前・山岸研究グループ

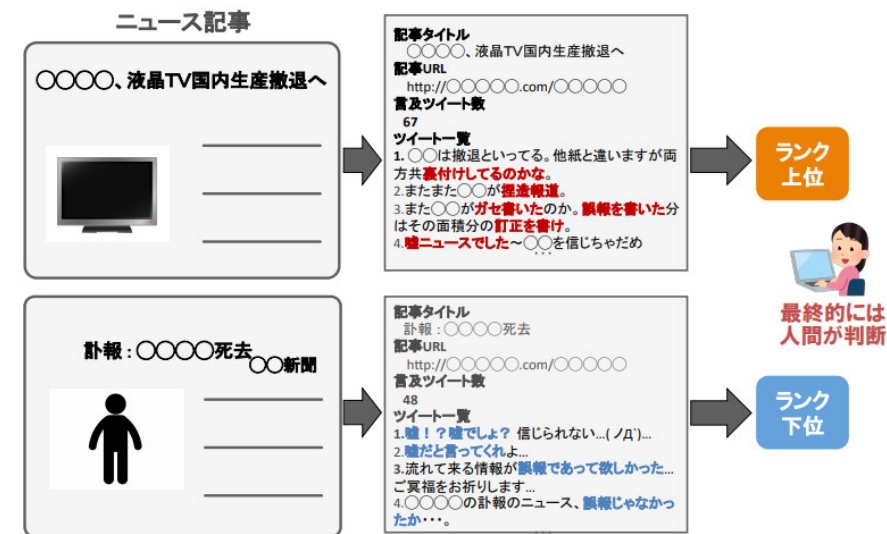
<https://www.nii.ac.jp/seeds/2019/echizen-yamagishi.html>

ファクトチェック支援

発信された情報が客観的事実に基づくものなのかを調査し、その情報の正確さを評価・公表するファクトチェック作業の効率化を自然言語処理によって支援

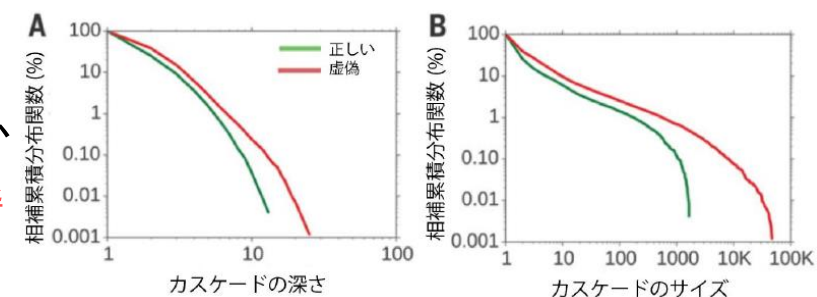
[出典] 東北大学 乾研究室

<http://www.cl.ecei.tohoku.ac.jp/publications/2017/171106factcheck.pdf>



フェイクニュースの拡散抑制

ソーシャルメディアの現象解析の結果、フェイクニュースは通常の投稿よりも速く遠くまでたくさん伝わるという特性に着目して、その拡散を抑制する



笹原和俊「フェイクニュースを科学する」(2018年 化学同人)



取り組み事例 (2/2) 熟慮・熟議の促進

- 異なる立場からの言明を知る、様々な立場からの意見を集める等によって、個人や集団の意思決定・合意形成をより良い結果に導く支援技術

言論マップ生成

ウェブ等の大量テキストを自然言語処理によって解析・整理し、ある言明に対する賛成意見・反対意見の両面とその元情報等を提示



大規模意見集約システムD-Agree

多人数による議論の構造を可視化、エージェント技術による自動ファシリテーション機能も加えて、大規模な人々の意見集約・合意形成のプロセスを支援し、集合知の醸成を促進

[D-Agreeの紹介動画] 名工大 伊藤研究室
<https://youtu.be/H-uNfYnkv90>

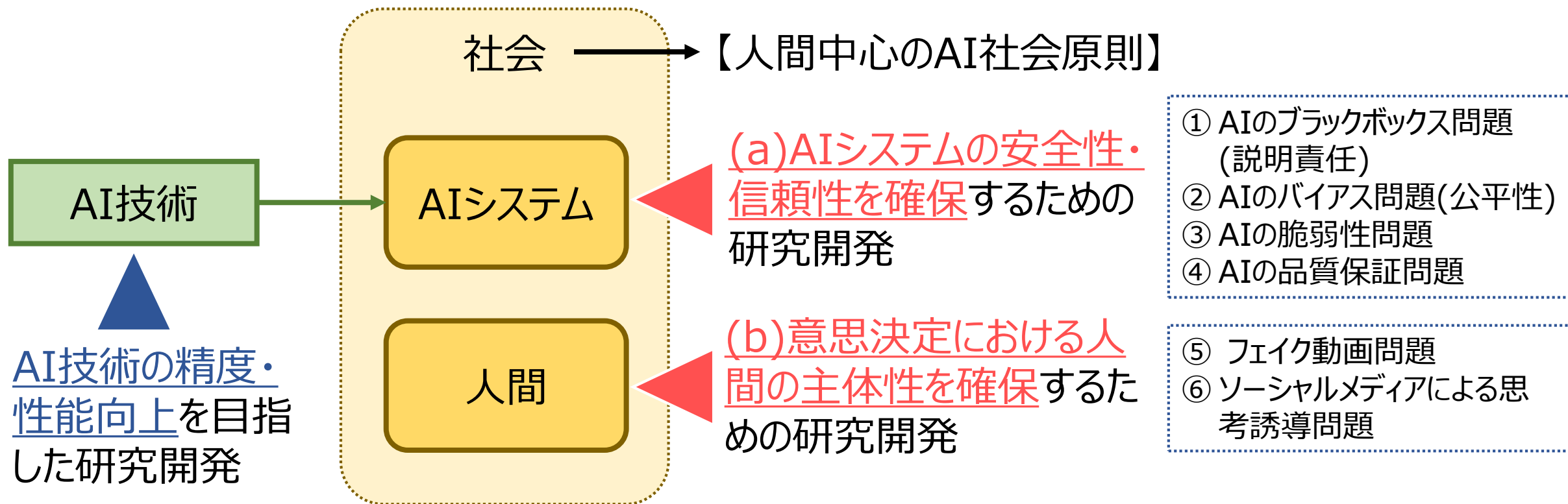


ELSI面の検討

人間の意思決定の支援を意図した機能が、思考誘導や検閲(表現の自由の制限)と受け止められる可能性もある。支援機能の考え方や仕組みについて利用者から見た透明性を確保し、社会受容性に配慮した技術開発を進める必要がある。

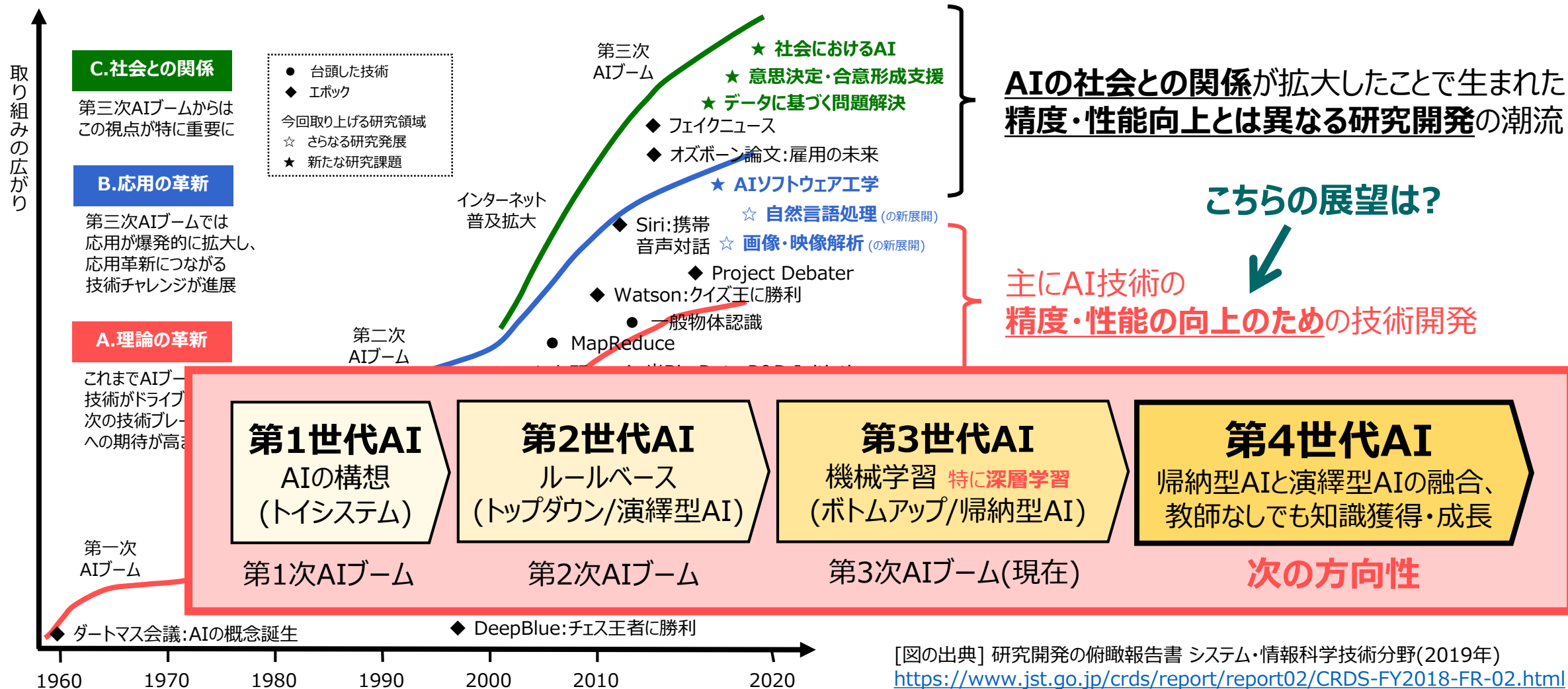
まとめ

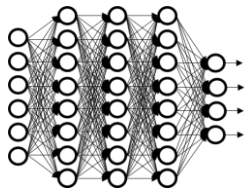
- AI技術が社会に広がり、精度・性能向上を目指した研究開発だけでなく、人間中心社会に向けたAI研究開発が新たな潮流を形成
- 人間中心の信頼される安全・高品質AIを日本の競争力に



もう一つのAI研究開発の潮流

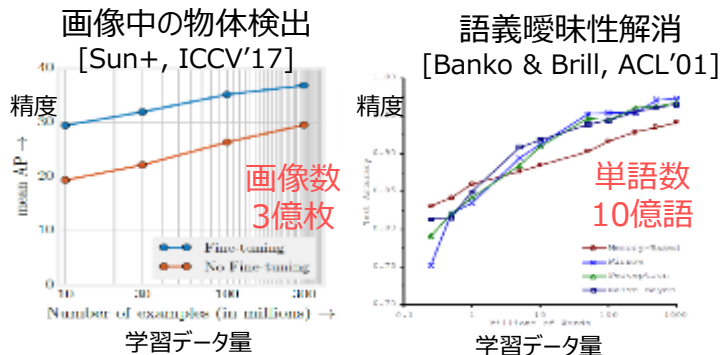
人工知能(AI)・ビッグデータ技術の時系列俯瞰図





第3世代AI(深層学習)の限界克服へ

- ✓ 膨大なデータ量の学習によって精度向上



- ✓ 最新アルゴリズムの実行には膨大な費用

自然言語解析の最高精度の学習アルゴリズム

- BERT (2018年) US\$ 6.9K/回
- XLNet (2019年) US\$ 61K/回(or 250K/回)

- ✓ 学習データと傾向が変わらない環境で動くことが前提 (その範囲外ではどんな結果になるか未知)

- ✓ 外部から与えられる制約への対応には別の枠組みが必要

- 法例、国ごとのルール
- 常識、慣習
- AI社会原則

データから学習するものではない決まり事



自動交渉エージェントに機械学習をさせた結果、「談合」を学習してしまった

① 学習に大量データや計算資源が必要



② 理由の説明が難しい(ブラックボックス)



③ 学習範囲外のケースへの対応が難しい



次世代AIの方向性

- 深層学習と知識・記号推論の融合

帰納型AI + 演繹型AI
深層学習の限界克服

- 人間の脳情報処理からのインスパイア

人間は大量の教師データなしに学習・成長、学習したこと組み合わせで別な問題解決に応用

- 次世代AIに対する安全性・信頼性

現状のAI安全性テストは学習済みモデルが対象だが、学習・成長するモデルに対する安全性確保は格段に難しい

関連報告

JST CRDSで発行した報告書や企画・開催したワークショップ等

- 研究開発の俯瞰報告書：システム・情報科学技術分野(2019年) 2019年3月
 - <https://www.jst.go.jp/crds/report/report02/CRDS-FY2018-FR-02.html>
- 戦略プロポーザル：AI応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立 2018年12月
 - <https://www.jst.go.jp/crds/report/report01/CRDS-FY2018-SP-03.html>
- ワークショップ報告書：機械学習型システム開発へのパラダイム転換 2018年3月
 - <https://www.jst.go.jp/crds/report/report05/CRDS-FY2017-WR-11.html>
- 2019年度人工知能学会全国大会(第33回)企画セッション：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 2019年6月
 - https://www.jst.go.jp/crds/sympo/201906_JSAI/index.html
- 戦略プロポーザル：複雑社会における意思決定・合意形成を支える情報科学技術 2018年3月
 - <https://www.jst.go.jp/crds/report/report01/CRDS-FY2017-SP-03.html>
- ワークショップ報告書：複雑社会における意思決定・合意形成を支える情報科学技術 2017年10月
 - <https://www.jst.go.jp/crds/report/report05/CRDS-FY2017-WR-05.html>
- 公開ワークショップ：意思決定のための情報科学 ～情報氾濫・フェイク・分断に立ち向かうことは可能か～ 2019年7月
 - <https://www.jst.go.jp/crds/sympo/20190725/index.html>

★ 第4世代AIの研究開発に関する戦略プロポーザル
2020年3月頃 公開予定

★ 公開ワークショップ報告書
2020年2月頃 公開予定