科学技術振興機構
Japan Science and Technology Agency

CRDS 研究開発戦略センター
Center for Research and Development Strategy

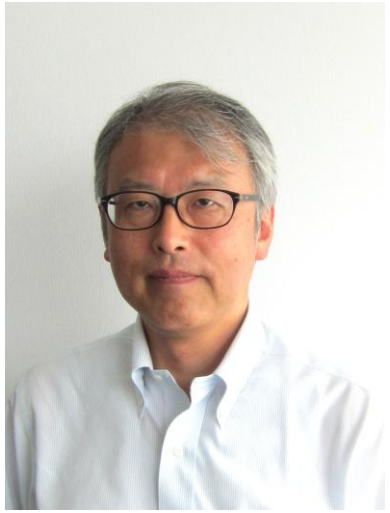
CSS2025 TWS企画セッション

デジタル社会における新たなトラスト形成に向けて

2025年10月28日

福島 俊一　JST CRDS
toshikazu.fukushima@jst.go.jp
https://researchmap.jp/toshikazu_fukushima

自己紹介　福島 俊一　Toshikazu Fukushima



1982年東京大学理学部物理学科卒業、NEC入社。以来、中央研究所にて自然言語処理・サーチエンジン等の研究開発・事業化、人工知能(AI)・ビッグデータ研究開発戦略を担当。工学博士。情報処理学会フェロー。

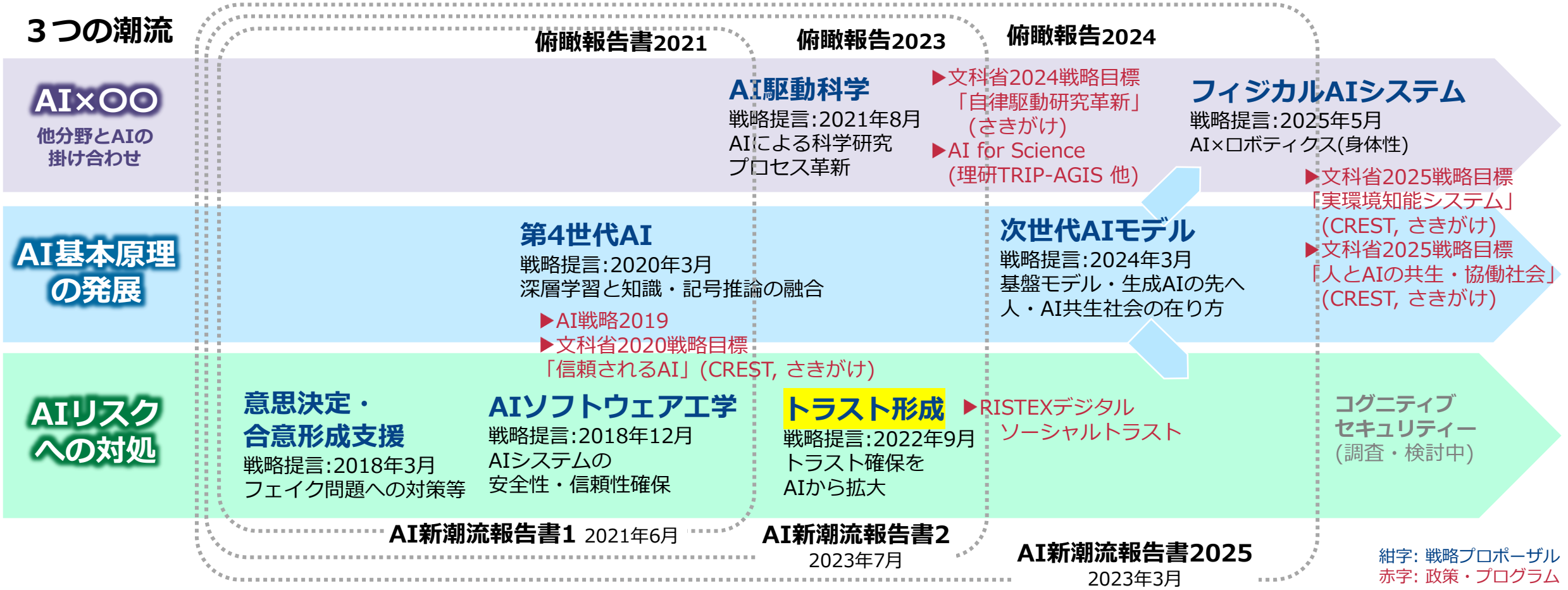
2016年4月から科学技術振興機構(JST)研究開発戦略センター(CRDS)フェロー。AI・トラストを中心とした情報科学技術分野の俯瞰的調査・戦略提言に従事。2019年からNEDO技術委員を兼任、AI・ロボット関連NEDOプロジェクトの審査・推進にも参画。

2011～2013年東京大学大学院情報理工学研究科客員教授。1992年情報処理学会論文賞、1997年情報処理学会坂井記念特別賞、2003年オーム技術賞ほかを受賞。 2015～2017年人工知能学会理事、2018～2020年人工知能学会監事。

趣味は料理・蕎麦打ち・パン作り、絵を描くことやアート展巡りなど。

JST CRDS におけるAI・トラスト関連の戦略提言活動

- AI研究の3つの潮流を軸として俯瞰的調査と国への戦略提言を行っている
- フェイク問題は2016年から提言活動を開始し、より広くトラスト問題を含めて検討している



関連するCRDS報告書リスト

全文PDFダウンロード可能
https://www.jst.go.jp/crds/report/by-category/02/index.html

科学を支え、未来へつなぐ

分野俯瞰

- **人工知能研究の新潮流2025 ～基盤モデル・生成AIのインパクトと課題～ (2025年)**
- 人工知能研究の新潮流2 ～基盤モデル･生成AIのインパクト～ (2023年)
- 人工知能研究の新潮流 ～日本の勝ち筋～ (2021年)
- 俯瞰ワークショップ報告書：エージェント技術 (2022年)
- 俯瞰ワークショップ報告書：ヒューマンインタフェース研究動向 (2023年)
- 研究開発の俯瞰報告書：システム･情報科学技術分野 (2023年)
- プレプリントサーバーarXivを利用したAI分野の研究動向俯瞰調査 (2024年)

戦略提言(1)次世代AIモデル

- 戦略プロポーザル：次世代AIモデルの研究開発 (2024年)
- 科学技術未来戦略ワークショップ報告書：次世代AIモデルの研究開発 ～技術ブレークスルーとAI×哲学～ (2024年)
- 戦略プロポーザル：第4世代AIの研究開発 ―深層学習と知識･記号推論の融合― (2020年)
- 科学技術未来戦略ワークショップ報告書：深層学習と知識･記号推論の融合によるAI基盤技術の発展 (2020年)
- JSAI2020企画セッション報告書：次世代AI研究開発 ―さらなる進化に向けて― (2020年)
- 俯瞰セミナー＆ワークショップ報告書：人・AI共生社会のための基盤技術 (2025年)

戦略提言(2)AIソフトウェア工学

- 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)
- 科学技術未来戦略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)

戦略提言(3) 意思決定･合意形成支援

- 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)
- 科学技術未来戦略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)
- 公開ワークショップ報告書：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～ (2020年)

戦略提言(4) デジタル社会のトラスト

- **戦略プロポーザル：デジタル社会における新たなトラスト形成 (2022年)**
- **俯瞰セミナー＆ワークショップ報告書：トラスト研究の潮流 ～人文･社会科学から人工知能、医療まで～ (2022年)**
- 科学技術未来戦略ワークショップ報告書：トラスト研究戦略 ～デジタル社会における新たなトラスト形成～ (2022年)
- 公開シンポジウム報告書「デジタル社会における新たなトラスト形成 ～総合知による取り組みへ～」(2023年)
- 連続シンポジウム報告書「さまざまな分野に広がるトラスト研究、総合による取り組みへ(1) ～フェイク問題、医療AI、安全保障とトラスト～」(2025年)
- 俯瞰ワークショップ報告書「コグニティブセキュリティー研究動向」(2024年)

戦略提言(5) AI駆動科学

- 戦略プロポーザル：人工知能と科学 ～AI･データ駆動科学による発見と理解～ (2021年)
- 俯瞰セミナーシリーズ報告書：機械学習と科学 (2021年)
- 科学技術未来戦略ワークショップ報告書：人工知能と科学 (2021年)
- 計測横断チーム調査報告書　計測の俯瞰と新潮流 (2018年)

戦略提言(6) フィジカルAI

- 戦略プロポーザル：フィジカルAIシステムの研究開発 ～身体性を備えたAIとロボティクスの融合～ (2025年)
- 科学技術未来戦略ワークショップ報告書：フィジカルAIシステム (2025年)
- 戦略プロポーザル：リアルワールド・ロボティクス ～開かれた環境に柔軟に適応するロボティクス学理基盤の創出～ (2022年)
- 科学技術未来戦略ワークショップ報告書：現実空間を認識し、臨機応変に対応できるロボットの実現に向けて (2022年)

CRDS 報告書

▶分野: 情報・システム

CRDS

CSS20XXでの招待講演

CSS2023 **AWS企画セッション**
アクロス福岡 2023.10.31

「信頼されるAI」の潮流
～AIのセキュリティ/トラストの課題～

CSS2024 **AWS企画セッション**
神戸国際会議場 2024.10.23

生成AIのリスク対策：取り組み状況と課題

情報処理学会論文誌 Vol.65, No.12
「社会的・倫理的なオンライン活動を支援する
セキュリティとトラスト」特集　2024.12

(SoK = Systematization of Knowledge)
SoK：デジタル社会における
トラスト形成の課題と展望

CSS2025 **TWS企画セッション**
岡山コンベンション
センター
2025.10.28

デジタル社会における
新たなトラスト形成に向けて

★前頁の報告書と上記SoK論文をベースとしてアップデート

講演の概要と構成　デジタル社会における新たなトラスト形成に向けて

本講演では、デジタル化の進展の中で、社会におけるトラストの機能がうまく働かなくなってきている状況に対して、その要因や課題を検討する。さらに、トラストの機能や特性を整理・モデル化して、対策の切り口や方向性を論じる。また、今後の重要課題となるAIエージェント(エージェント型AI)のトラスト問題についても展望する。

① デジタル社会のトラスト問題

② トラスト研究開発の状況と課題

③ トラストの特性・モデル化と研究開発の方向性

④ AIエージェントのトラスト問題

①デジタル社会のトラスト問題

②トラスト研究開発の状況と課題

③トラストの特性・モデル化と研究開発の方向性

④AIエージェントのトラスト問題

CRDS

トラスト(信頼)の役割

Trusted Webホワイトペーパー Ver.1
https://www.kantei.go.jp/jp/singi/digitalmarket/trusted_web/index.html

トラストは、事実確認をしない状態で、相手が期待した通りに振る舞うと信じる度合いのこと

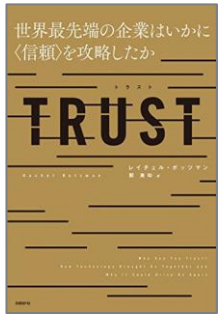
工藤郁子：「人々の「眠り」と「目覚め」、社会の信頼 再構築を」、
朝日新聞デジタル「にじいろの議」2020.8.12
https://www.asahi.com/articles/DA3S14584996.html

トラストは、取引や協力のコストを減らしてくれる社会関係資本であり、法制度は、裏切られるリスクを軽減し、各人の限定的な情報力・判断力を補い、信頼を補完し、促進する機能がある

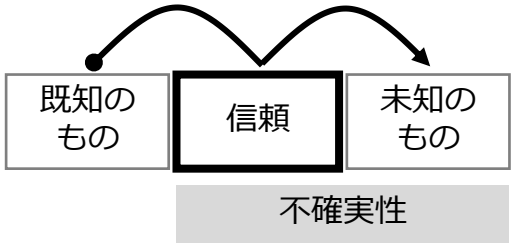
ニクラス・ルーマン著、大庭健・正村俊之訳、勁草書房

信頼－社会的な複雑性の縮減メカニズム

レイチェル・ボッツマン著、関美和訳、
「TRUST: 世界最先端の企業はいかに〈信頼〉を攻略したか」、日経BP社



トラストは、確実なものと不確実なものの隙間を埋める力であり、未知のものに自信を持つこと



トラストはビジネスを発展させ、かつ競争要因

- 取引・協力できる相手の拡大
  (サプライチェーン拡大、シェアリングエコノミー等)
- 先端技術を用いたビジネス加速
  (新技術・新ビジネスに対する社会受容の促進等)
- ビジネス意思決定の迅速化・不安低減
  (経営判断、災害時・非常時判断等)
- 消費者・取引参加者の安心
  (偽装偽造リスクの低減、評判・レビューの健全化等)

CRDSでの有識者インタビュー等をもとに整理

問題意識、目指す社会の姿　社会のどんな問題をどう解決したいのか

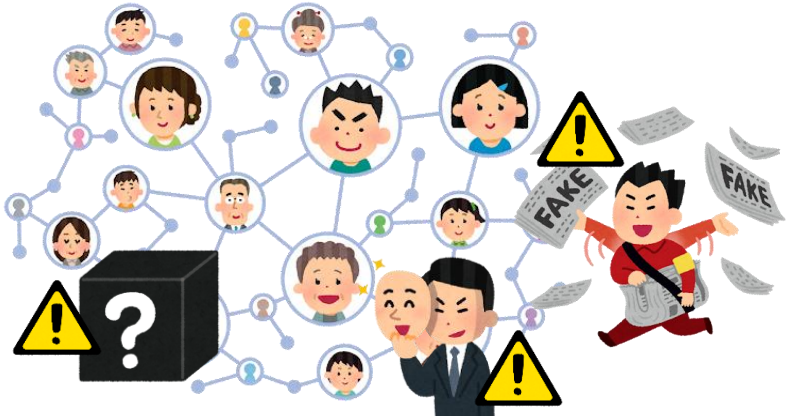
- リスクはあるのだが、トラストすると、安心して迅速に行動・意思決定ができる
- デジタル化の進展に伴い、「旧来のトラスト」がうまく機能しなくなってきている
- デジタル社会においてもうまく機能する新しいトラストの仕組み作りを目指す

(a)過去

旧来のトラスト

顔が見える人間関係や人々の間のルールに支えられたトラスト

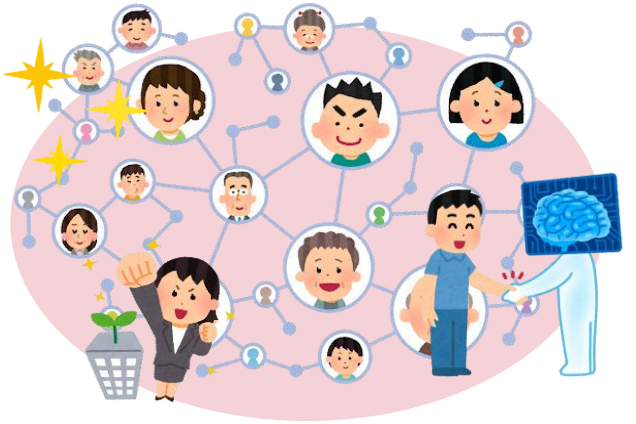
(b)現在の状況



デジタル社会におけるトラストのほころび

- バーチャルな空間にも広がった人間関係
- 複雑な技術を用いたシステムへの依存
- だます技術の高度化

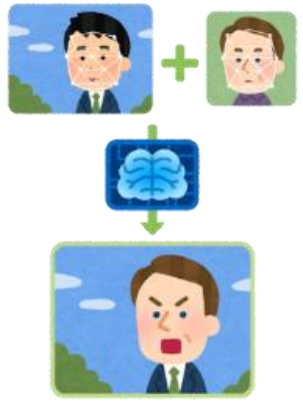
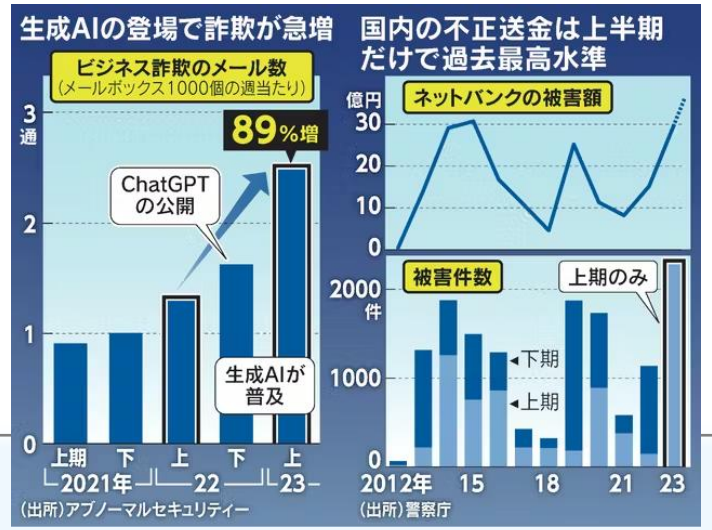
(c)目指す姿



新たなトラスト形成

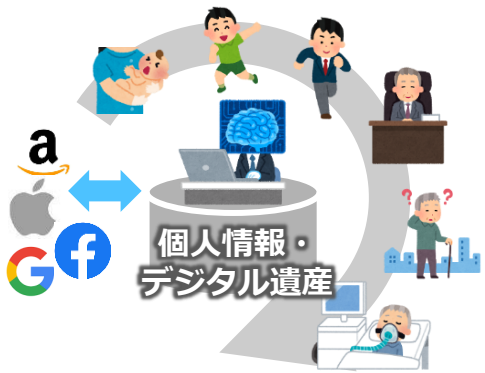
不信・警戒を過度に持つことなく幅広い協力・取引・人間関係が作れて、デジタル化によるさまざまな可能性・恩恵が広がる

デジタル化の進展とトラスト問題　だます技術の高度化

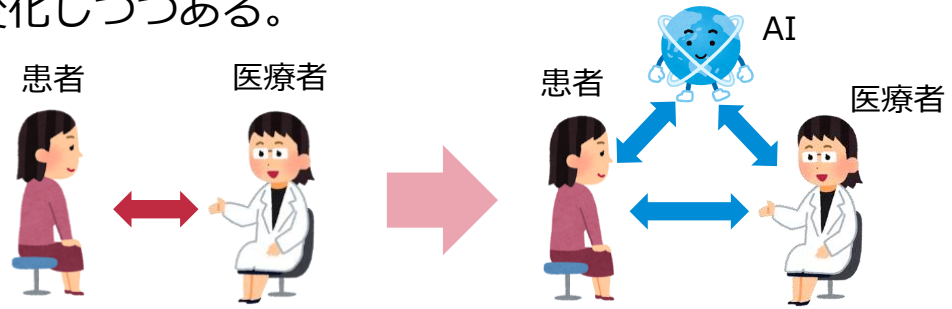
場面	変化内容	深刻化するトラスト問題
(1) メディアにおけるフェイク拡散	生成AI技術の高性能化・高機能化が進み、人間の認識能力では見破れないフェイク画像・音声・動画・文章の作成が容易になった。ソーシャルメディアの普及によって、フェイクやデマの拡散が大規模化。  	フェイク動画による政治干渉や個人攻撃・棄損等が社会問題化。簡単に人をだますことができてしまい、裁判等での証拠の信頼性も揺らぐ。フェイクの法的規制が強くなると、表現の自由が妨げられる恐れも生じる。
(2) 仮想世界のトラストに基づく取引	ネットの世界で、リアルには面識のない人たちとの取引や、仮想通貨・デジタル資産を用いた取引が拡大。様々な分野で、不特定多数の間のマッチングビジネスやシェアリングビジネスが立ち上がっている。  	既に様々なビジネスが広がっている状況だが、仮想世界・デジタルデータの性質を悪用した偽装やなりすまし等の犯罪も起きている。対策も取られているが、常に新しい仕組みが生まれ、新しいリスクが発生し、対策が追いつかない面もある。相互評価スコアはある程度は有効であるものの、それだけでは不十分だという問題も起きている。

https://www.nikkei.com/article/DGXZQOUC064RI0W3A101C2000000/

デジタル化の進展とトラスト問題　複雑な技術・システムへの依存

場面	変化内容	深刻化するトラスト問題
(3) 自動運転車	人口減少・過疎化からバスやタクシーの運転手不足が進みつつある中、自動運転車のニーズが高まる。AI技術を活用した状況認識や運転制御によって、車が運転手なしで走行する地区・状況の拡大が見込まれる。 	AI技術はブラックボックスで動作保証や精度保証ができない。 安心して乗車できるのか？ 事故が発生したときに、原因解明や責任の所在はどうなるのか？
(4) パーソナルAIエージェント	個人情報の管理代行をパーソナルAIエージェントに任せるサービスの利用が増えていく。 ・ デジタル遺産管理 ・ お薬手帳 ・ 母子手帳　等 	ブラックボックスで、個人の意図・期待の通りに振る舞うという100%保証はできないのに、個人情報を委ねることができるか？ 期待に反する事態が起きた場合の責任の所在はどうなるのか？
(5) 人を評価するAIシステム	採用試験の一次フィルタリング、人事評価や配属最適化といった人事業務へもAIシステムが応用されつつある。 中国では個人のさまざまな行動履歴を追跡し、信用スコアを算出して、優遇や制限を与えるシステムが稼働している。 	学習データに偏りや差別的要因が含まれていると、不公平で差別的な評価を助長するという問題がある。また、評価アルゴリズムに過剰適合して行動する人々を生み出す。

デジタル化の進展とトラスト問題　関係性の拡大・多様化

場面	変化内容	深刻化するトラスト問題
(6) 医療意思決定におけるAIセカンドオピニオン	従来は患者と医療者の二者関係による医療意思決定だった。そこに、AIによる情報提供や診断が加わり、三者関係による医療意思決定へと変化しつつある。 	患者が異なる多様な情報を参照できるようになり、医療者からの説明と異なる情報が得られることもある。三者関係における新たな役割関係とそこでのトラストの在り方が必要になっている。
(7) メタバース内活動におけるトラスト	仮想世界(メタバース)の中で自分のアバターを介した新たな人間関係や経済活動が生まれる。 	生身の人間や物理的な実体が必ずしも確認できない世界において、リアル世界と同様のトラストが成り立ち得るか？
(8) 高度な擬人化インタフェース	親近感を持てる外観と高い会話能力を備えたコンピュータエージェントやロボットが、様々なサービスや日常的なタスクにおいて、対人インタフェースとして使われるようになる。 	人間らしい外観を持っていると、人間並みの能力を持っていると期待・過信してしまい、そうでないと失望するといったことが起きやすい。その一方、身近なロボットに、過度な親近感・依存感を持ってしまうタイプの人もいる。

①デジタル社会のトラスト問題

②トラスト研究開発の状況と課題

③トラストの特性・モデル化と研究開発の方向性

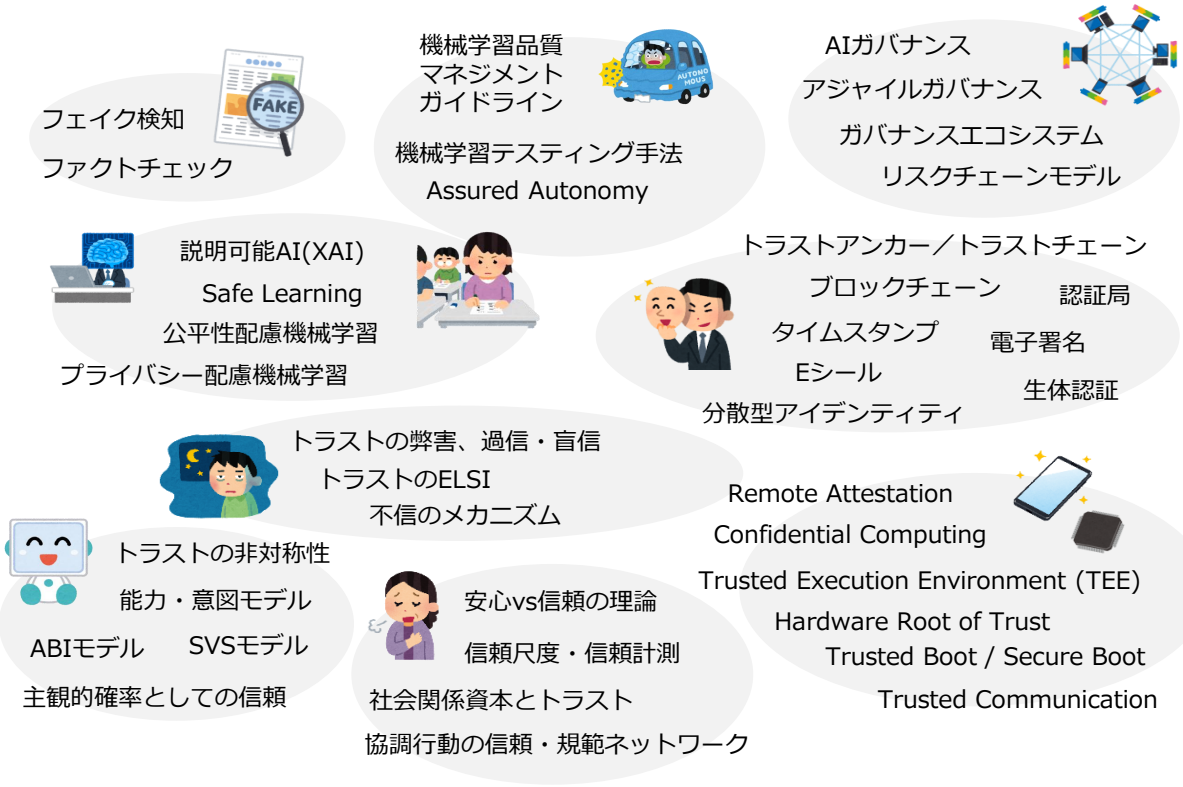
④AIエージェントのトラスト問題

CRDS

研究開発の状況と目指す姿

取り組みの現状

人文・社会科学分野の基礎研究から、ビジネス・社会実装と連動した情報科学分野の技術開発、医療・SNS等の応用シーンでの対策検討まで、幅広く取り組まれているが、それらの間で知見共有・連携がまだ少なく、それぞれはトラスト問題に対して個別的な対処、断片的な状況改善にとどまる



目指す姿

デジタル社会における新しいトラストの仕組みとそれによるトラスト問題対策の全体ビジョンを描いて共有し、具体的トラスト問題と共通基礎の両面から連携して、社会に貢献する研究を目指す

デジタル社会におけるトラスト形成

具体的トラスト問題ケースへの取り組み
ビジネスにおけるトラスト、ネット情報のトラスト、
AI応用システムのトラスト、AI＋専門家のトラスト、…

社会的トラスト形成フレームワーク
公正・健全なトラスト基点の維持、トラストの悪用・攻撃への対策、使いこなしを容易にする技術・教育、…

トラストの社会的よりどころの再構築
対象真正性/内容真実性/振る舞い予想・対応可能性の
社会的よりどころの拡充、多面的・複合的検証の仕組み、
改ざんされない記録・トレーサビリティ、…

トラストに関する基礎研究
デジタル社会におけるトラスト形成や不信のメカニズム理解、
トラストに関わる日本人のメンタリティと国際比較・文化差、
デジタル社会のトラスト形成の方策・対策設計の裏付け、…

研究開発がバラバラに進んできた要因

トラストに対する異なる側面・切り口からの研究開発が進められてきた

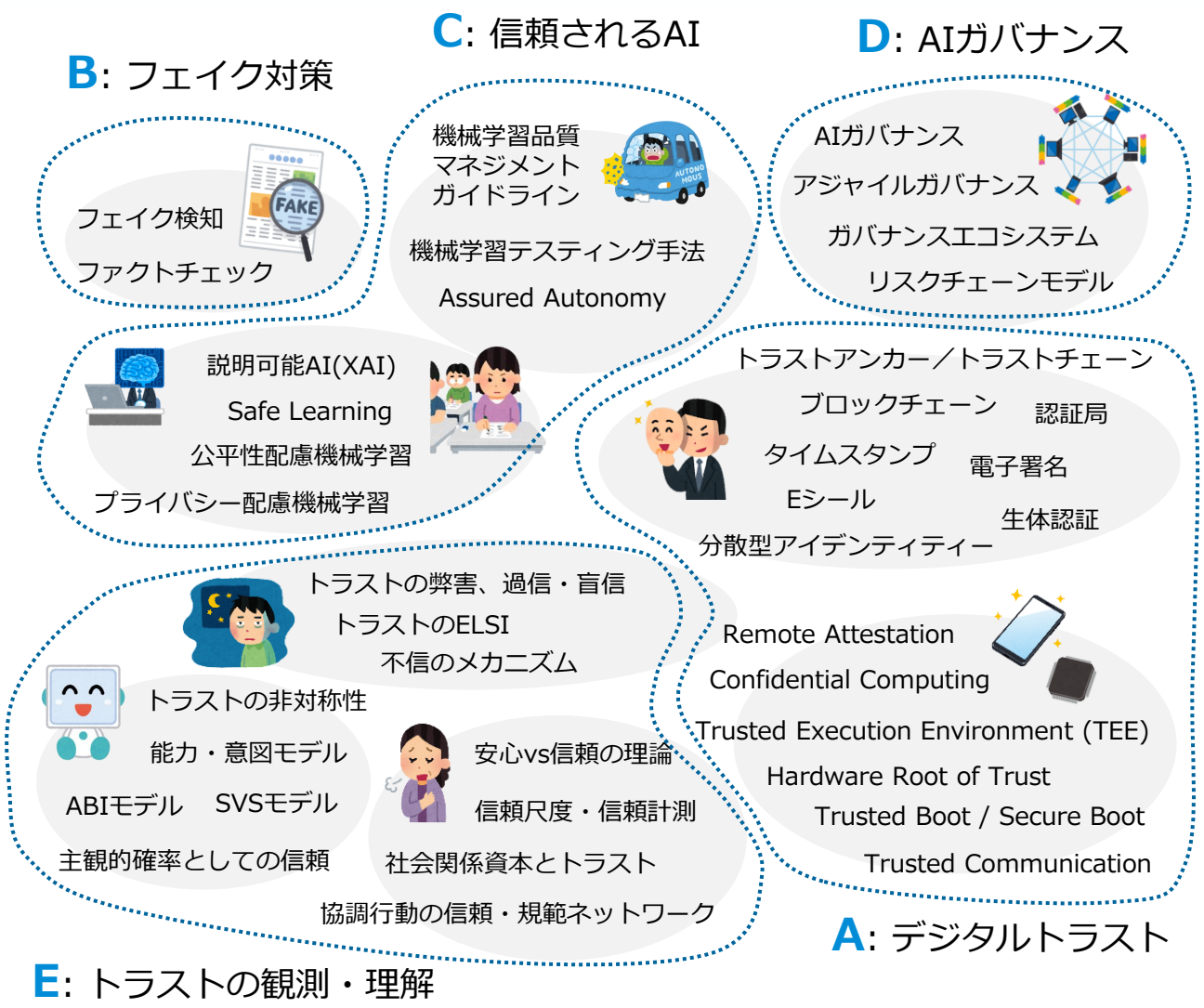
- 技術開発による対策設計
  - 対象真正性の問題に重点 …… A: デジタルトラスト
  - 内容真実性の問題に重点 …… B: フェイク対策
  - 振る舞い予想・対応可能性の問題に重点 …… C: 信頼されるAI
- ルール整備・プロセス管理による対策設計 …… D: AIガバナンス
- 対策設計・社会受容の裏付け …… E: トラストの観察・理解

[注] おおまかな傾向であり、この見方に収まらない取り組みも存在する

トラストの3側面

※大屋雄裕(慶應大)からの示唆による

対象真正性	本人・本物であるか？
内容真実性	内容が事実・真実であるか？
振る舞い予想・対応可能性	対象の振る舞いに対して想定・対応できるか？



バラバラの取り組みでよいのだろうか？

ある新しいサービスを
使ってみようかと
考えるとき

- ✓そのサービスの仕組み(どのように動いてどのような結果が得られそうか)が信じられるか　振る舞い予想・対応可能性
- ✓そのサービスの提供企業が怪しくないか　対象真正性
- ✓そのサービスについての評判やレビュー投稿は本当か(ヤラセではないか)　内容真実性
- ✓上記のそれぞれについても、異なる観測結果や意見が得られる

いろいろな視点から多面的に関連情報を集め、その一つだけでは確信を持てなくとも、それらを複合的に検証することで、総合的な判断を下す

デジタル化の進展でリスクが高まった状況では、断片的に切り取られた情報や対象のある一面しか見ずに、何かを信じ込むことはとても危うい

デジタル社会のトラストは、多面的・複合的な検証によって支えていくべきであり、その検証が適切かつ容易に行えるような仕組みが望まれる

注目されるトラスト関連政策提言・プログラム事例

分類	日本	海外
A. デジタルトラスト **対象真正性**	● 「Data Free Flow with Trust (DFFT)」(2019年1月ダボス会議、安倍首相) ● デジタルトラスト協議会(2020年8月設立) ● 内閣官房デジタル市場競争本部 Trusted Web推進協議会「Trusted Webホワイトペーパー」Ver.1 (2021年3月)、Ver.2 (2022年8月)、Ver.3 (2023年11月) ● Originator Profile技術研究組合(2022年12月設立) ● デジタルガバメント閣僚会議 データ戦略タスクフォース「包括的データ戦略」(2021年6月閣議決定) ● デジタル庁 データ推進戦略ワーキンググループ トラストを確保したDX推進サブワーキンググループ(2021年11月～2022年7月)	● 欧州 eIDAS (electronic Identification and Authentication Service)規則（2014年7月成立、2016年7月施行）：トラストサービスの統一基準
B. フェイク対策 **内容真実性**	● 経済安全保障重要技術育成プログラム(K Program)に基づくNEDO事業「偽情報分析に係る技術の開発」に採択「偽情報の検知・評価・システム化に関する研究開発」(2024年～2027年) ● 同K Programに基づくJST事業「人工知能(AI)が浸透するデータ駆動型の経済社会に必要なAIセキュリティ技術の確立」に採択「SYNTHETIQ X：フェイク情報拡散の防御と予防を実現する研究基盤」(2024年度～2028年度) ● JST RISTEXプログラム「デジタル ソーシャル トラスト」(2023年～)	● 米国国防高等研究計画局（DARPA）「Media Forensics (MediFor)」プログラム(2016年～2020年)、「Semantic Forensics (SemaFor)」プログラム(2020年～2024年)
C. 信頼されるAI **振る舞い予想・対応可能性**	● 統合イノベーション戦略推進会議「AI戦略2019」(2019年6月)における主要な研究開発課題として「Trusted Quality AI」 ● 文部科学省2020年戦略目標「信頼されるAI」を受けたJSTプログラム：CREST「信頼されるAIシステム」、さきがけ「信頼されるAI」(2020年度～) ● NEDO「次世代人工知能・ロボット中核技術開発事業」において「AIの信頼性」(2020年度～)	● 英国研究・イノベーション機構(UKRI)「Trustworthy Autonomous Systems Programme」(2020年～)
D. AIガバナンス	● 世界経済フォーラム「Rebuilding Trust and Governance: Towards DFFT」白書(2021年3月) ● 経済産業省「Governance Innovation Ver.2: アジャイル・ガバナンスのデザインと実装に向けて」(2021年7月)、「AI原則実践のためのガバナンス・ガイドライン Ver. 1.1」(2022年1月) ● 日本ディープラーニング協会「AIガバナンスとその評価」研究会報告書「AIガバナンス・エコシステム－産業構造を考慮に入れたAIの信頼性確保に向けて－」(2021年7月)	● 欧州委員会「The EU Artificial Intelligence Act」(2021年4月法案発表、2024年5月成立・8月発効)

日本では国の戦略としてトラストの各側面で施策を強化しつつあるが、それらを包含する全体ビジョンを描いて推進するならば、より強力な戦略構築、国際競争力強化が可能になるのではないか

①デジタル社会のトラスト問題

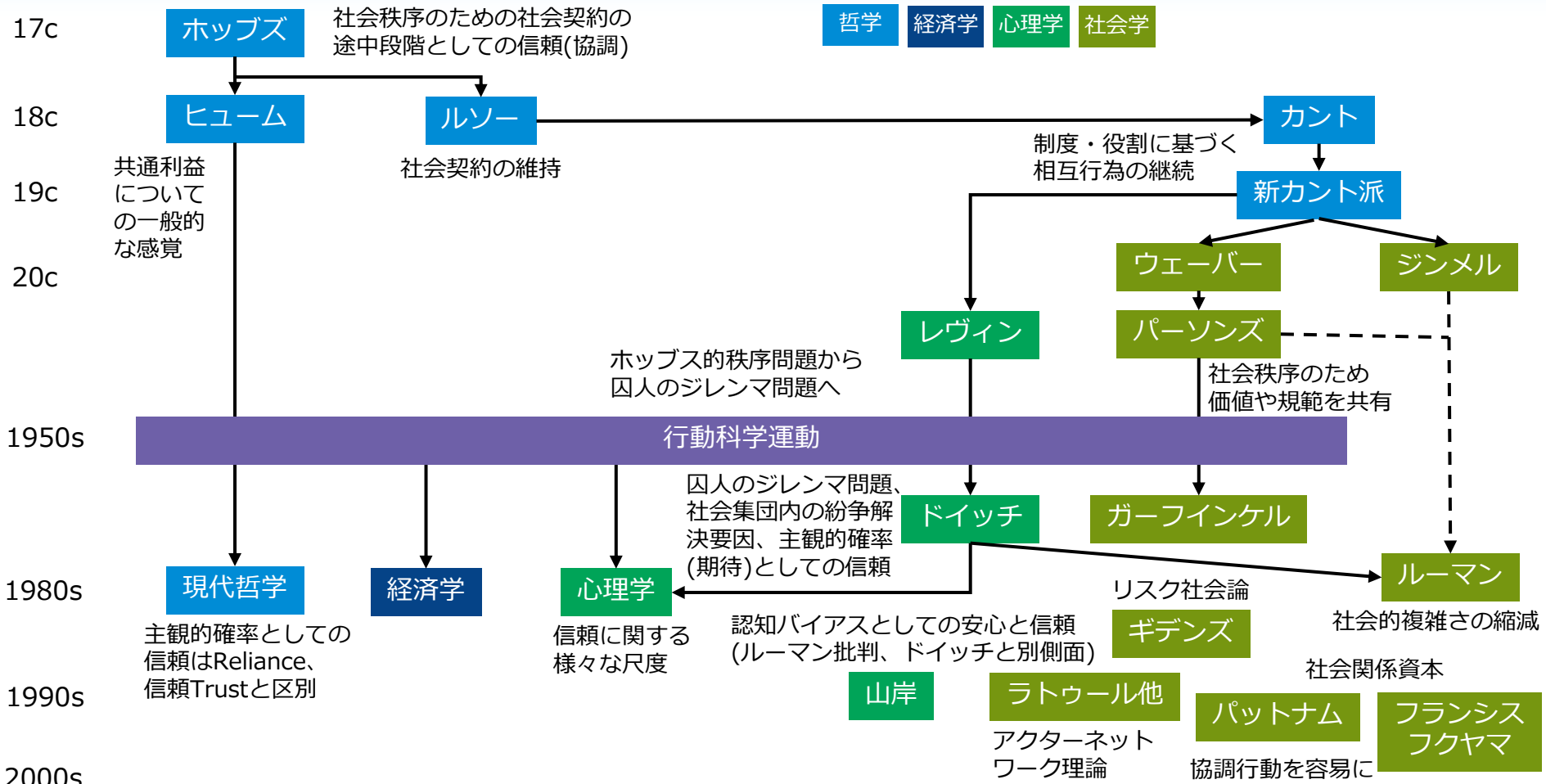
②トラスト研究開発の状況と課題

③トラストの特性・モデル化と研究開発の方向性

④AIエージェントのトラスト問題

CRDS

トラスト研究の変遷　人文・社会科学分野では長年取り組まれてきた



『信頼を考える:リヴァイアサンから人工知能まで』(小山虎編、勁草書房 2018年) および CRDS 報告書「俯瞰セミナー＆ワークショップ報告書：トラスト研究の潮流 ～人文･社会科学から人工知能、医療まで～」(2022年) をもとに記載

情報科学

トラストに関わる研究は1990年以降

- 形式検証
- 署名、認証
- Device Integrity
- デジタルトラスト
- 電子商取引
- システムのトラスト
- AI/ロボットとの関係性(HAI/HRI)

ITリスクへの取り組み

Computer Safety
↓
Software Dependability
↓
CPS Trustworthiness
↓
AI Rist Management

トラスト・信頼の様々な定義　このような取り組みも踏まえつつ…

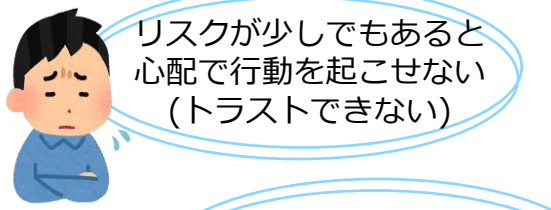
分野	定義
社会学	(信頼は)欠けている情報を内的に保証された安全性に置き換えるのであり、(自分に)利用可能な情報を超えて、行動の期待を一般化することにより、社会の複雑さを減少させる [Luhmann 1979]
	(信頼は)情報不足を内的に保証された確かさで補いながら、手持ちの情報を過剰に利用し、行動予期を一般化することで、社会的な複雑性を縮減するもの、未来における他人の振る舞いによる利益を見越して、未来における他人の振舞い(裏切り)による害が生じうることを認識しつつも、現在において決定を行なうもの [Luhmann 1968?]
	(信頼とは)自然的秩序および道徳的社会秩序の存在に対する期待 [Barber 1983]
	(信頼は)他者の誠実さや愛あるいは抽象的な原理への信念を表すような、人やシステムが一群の結果や出来事を実際にもたらすという確信 [Giddens 1990]
哲学	AがBはCすると信頼するのは、(1) AはBがCすると期待し、(2) このAの期待(1)が、Bが自分の関心を叶えようという動機に基づいているというAの信念か知識に基づいているとき [Hardin 1991]
	AがBはCすると信頼するのは、(1) AはBに重要事Cを任せ、(2) AはどのようにCを扱うのかのコントロールをある程度Bに許し、(3) AはBがCを扱うことができると確信しており、(4) Aは自分に対するBの善意に確信を持っているか、少なくともBの悪意や無関心を予期しないとき [Baier 1986]
	AがBはCすると信頼するのは、(1) AがBの善意に対する楽観的態度を持ち、(2) AはBの予期される行動Cに対するBの能力に対する楽観的態度を持ち、(3) Aは自分が頼りにすることを認識することによって直接BがCするように動機付けられると信じているとき [Jones 1996]
	AがBはCすると信頼するのは、(1) AはCの配慮にあたってBがある社会的規範に内的にコミットしていると期待し、(2) Aは「BがCの配慮にあたってAによって想定されている社会的規範を認識し、また、その規範が何を要求しているかを理解することができる」と確信しており、(3) AはBが自分に課せられた規範にしたがって行為することができると信じているとき [Mullin 2005]
経済学	(信頼は)1人ないし複数の行為者が特定の行為を遂行するという一定のレベルの主観的確率であり、彼らの行為をチェックすることができる以前に(あるいはチェックすることができる能力とは独立に)、その行為が自分自身の行為に影響を与える状況で形成される [Gambetta 1988]
社会心理学	信頼は、相手の行動によって自分の「身」が危険にさらされる状態で、相手がそのような行動をとらないだろうと期待すること [山岸 1998]
経営学	(信頼は)他者の意図や行動についてのポジティブな期待に基づきリスクを受け入れる意図を含む心理状態である [Rousseau 1998]
工学・情報科学	不確実な状態の中で、相手が自分のゴール達成に協力してくれるという信念 [Lee 2004] (信頼工学：信頼 = AIの性能に対する人間による主観的期待値)
	事実を確認しない状態で、相手先が期待したとおりに振る舞うと信じる度合い [Trusted Web推進協議会 2021]

『信頼を考える:リヴァイアサンから人工知能まで』(小山虎編、勁草書房 2018年) および CRDS報告書「俯瞰セミナー＆ワークショップ報告書：トラスト研究の潮流 ～人文･社会科学から人工知能、医療まで～」(2022年) をもとに記載

トラストのモデル・性質　たたき台としての整理の一案

- トラストは相手が期待を裏切らないと思える状態 (100%の保証はなく主観)
- リスクに対する多大なチェックコストを省いて、安心して迅速に取引・協力・意思決定できる
- 相手をトラストするかは、最終的に主観依存

Trustworthiness [ISO/IEC TS 5723:2022]
検証可能な方法で、利害関係者の期待に応える能力
※信頼性(Reliability:仕様通りに動く)はTrustworthinessの一要素



リスクが少しでもあると心配で行動を起こせない(トラストできない)

リスクなんて気にしない、きっと大丈夫さ！と思ったらだまされた…

Trustor
(信頼する側)

Trust
(A)対象真正性
(B)内容真実性
(C)振る舞い予想・対応可能性

Trustee
(信頼される側)

Trustorにとって期待を裏切られないと思える

能力
意図

能力・意図モデル

主観

(a) 経験的・伝統的に大丈夫だと思う
(b) 自ら検証・確認したことで大丈夫だと思う
(c) 他者が大丈夫と言っているので大丈夫だと思う

客観
Trustworthiness
信頼相当性

信頼相当性の計測/観測/検証の結果から裏付けのあるケース

裏付けがなく**リスク**はあるが大丈夫だとみなすケース

AI応用システムの場合:
モデル正確性、モデル頑健性、透明性、解釈可能性、信頼性、安全性、サイバーセキュリティ、回復性、公平性、プライバシー、アカウンタビリティ、適合性等

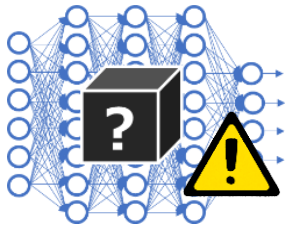
『AIリスク・マネジメント：信頼できる機械学習ソフトウェアへの工学的方法論』(中島震著、丸善出版 2022年)記載の品質観点

Trusteeが期待を裏切る可能性はゼロではない

※トラストの定義や二重円での表現は「Trusted Webホワイトペーパー」を参考にしたが、異なる解釈をしたところがある。その詳細は、本資料末尾の補足パートに記載する。

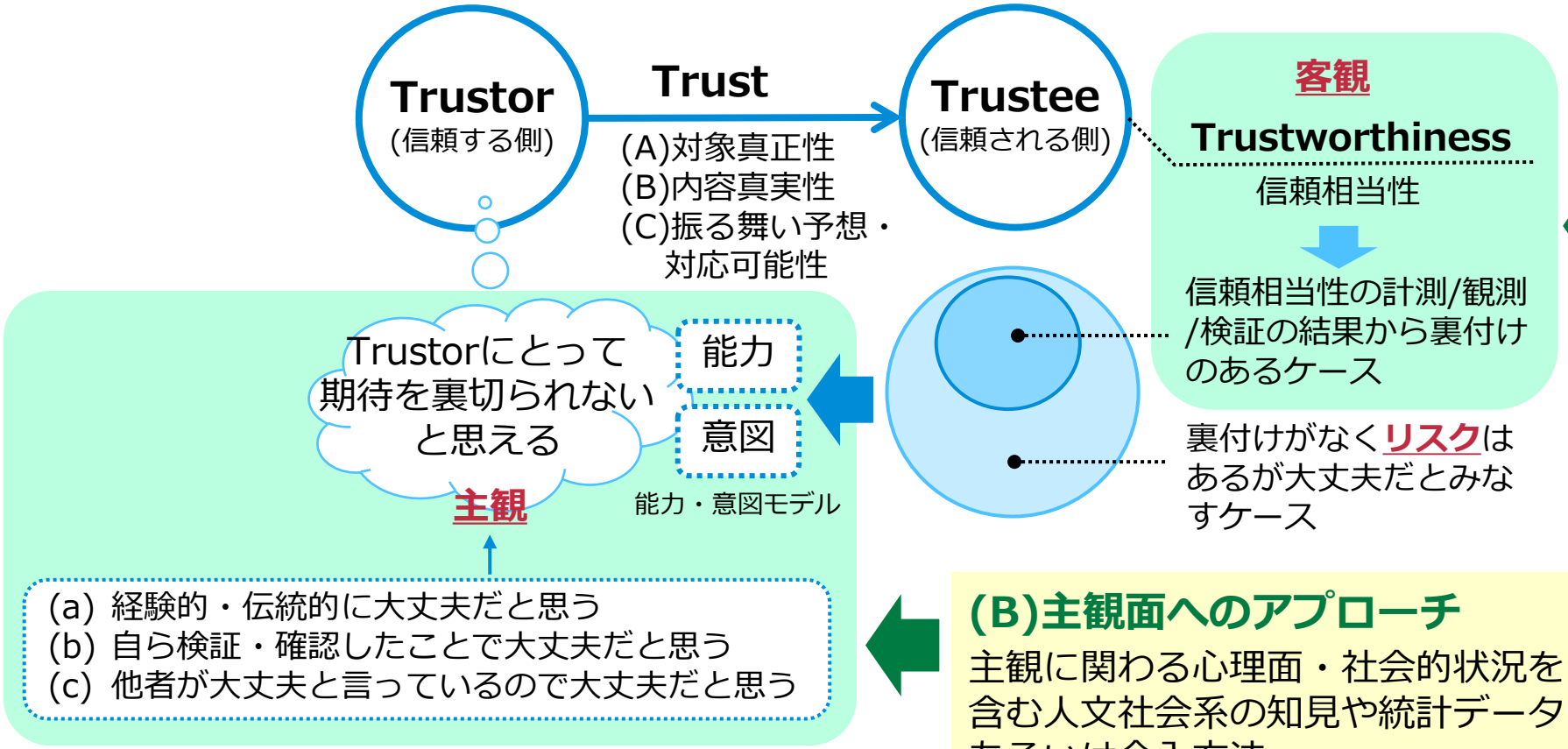
トラストの3側面と社会的よりどころ

- 対象真正性は大前提、内容真実性を評価しつつ、期待と照らして振る舞い予想・対応可能性を判断
- トラストするかは最終的に主観依存だが、詐欺などの犯罪を防ぎ、社会秩序を確保するためには、トラストの3側面のそれぞれで「社会的なよりどころ」の確保が望まれる

トラストの3側面	旧来の社会的よりどころ	トラストのほころび
(A)対象真正性 本人・本物であるか？  	印鑑・サイン、身分証・鑑定書、デジタル認証・生体認証など	様々なものがデジタル化されて流通・処理されるようになり、デジタル特有の偽造・偽装・改ざんの対象や可能性が拡大した。
(B)内容真実性 内容が事実・真実であるか？  	事実性は証拠写真・監視カメラ映像など、学説は査読制による学術コミュニティ合意など	生成AIはハルシネーション(もっともらしい嘘)を生じ、また、生成AIによるフェイク生成が高品質化・容易化し、人間の目では真贋・真偽を見破るのが極めて難しい状況になっている。
(C)振る舞い予想・対応可能性 対象の振る舞いに対して想定・対応できるか？  	人的行為・タスクについては契約・ライセンスなど、機械・システムの動作については仕様書など	深層学習の振る舞いは確率的で、動作保証・精度保証はされず、常にその動作を予見できるわけではない。説明可能AI技術は近似的説明を作るものであり、説明から外れる動作を起こし得る。

トラスト研究開発のアプローチ

■ トラストの客観面(信頼相当性)を強化する研究開発と、
トラストの主観面(人間の認知・思考の特性)を扱う研究開発の両面が必要



(A)客観面へのアプローチ
信頼相当性の検証の枠組みを作り、裏付けのあるケースを拡大する技術開発・制度設計

(B)主観面へのアプローチ
主観に関わる心理面・社会的状況を含む人文社会系の知見や統計データあるいは介入方法

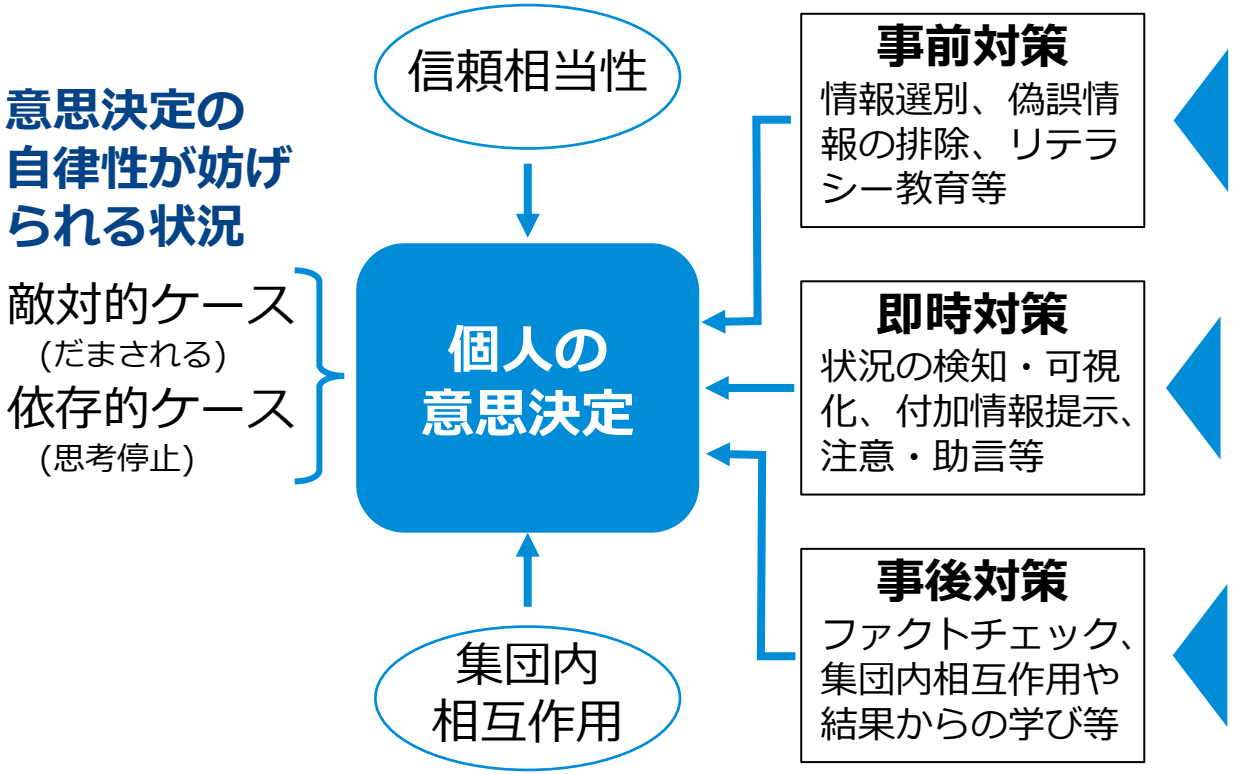
(A)客観面へのアプローチ

- まず、トラストの3側面それぞれについて、旧来の社会的よりどころにほころびが見られる中、新たな社会的よりどころを強化する
- さらに、ある一面の情報だけで100%保証されることはなく、断片的な情報だけで信じ込むことはとても危ういので、多面的・複合的な検証を可能にすることで、信頼相当性を確保する

トラストの3側面	旧来の社会的よりどころ	強化の方向性	
(A)対象真正性 本人・本物であるか？	印鑑・サイン、身分証・鑑定書、デジタル認証・生体認証など	対象真正性の新たな社会的よりどころとその検証手段の実現	多面的・複合的な検証を容易にする仕組みによって、信頼相当性をより確実なものにする 人間の検証負荷を軽減するためにAI活用によって多面的・複合的検証を支援
(B)内容真実性 内容が事実・真実であるか？	事実性は証拠写真・監視カメラ映像など、学説は査読制による学術コミュニティ合意など	内容真実性の新たな社会的よりどころとその検証手段の実現	
(C)振る舞い予想・対応可能性 対象の振る舞いに対して想定・対応できるか？	人的行為・タスクについては契約・ライセンスなど、機械・システムの動作については仕様書など	振る舞い予想・対応可能性の新たな社会的よりどころとその検証手段の実現	

(B)主観面へのアプローチ

- トラストの主観面(人間の認知・嗜好の特性)に関する統計的傾向や個人・集団による差異の把握およびモデル化(統計データの収集・分析、認知科学・心理学・脳神経科学・行動経済学等の知見)
- 意思決定の自律性が妨げられる状況の検知、および、その状況を回避するための介入・対策の創出

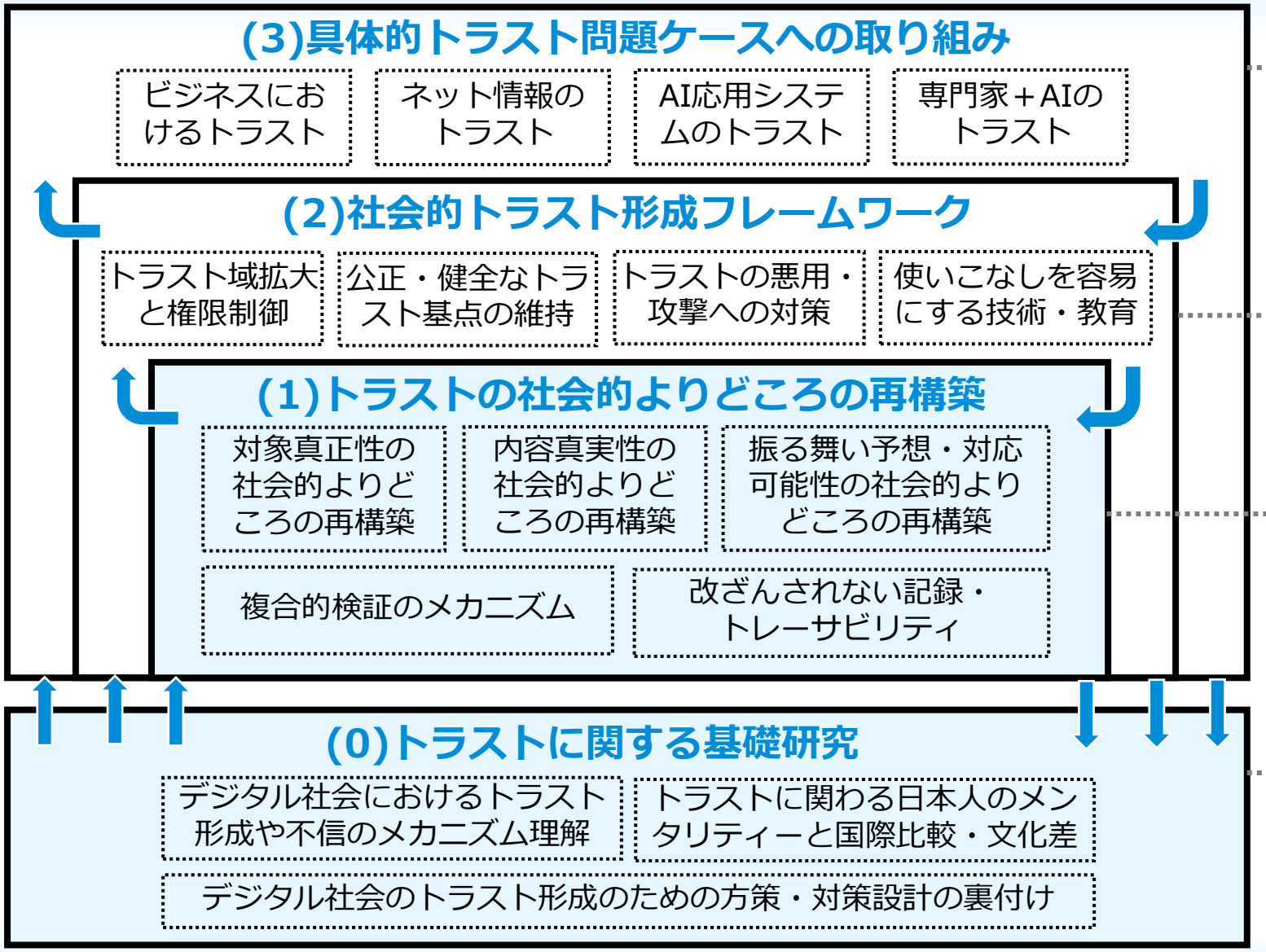


トラストの主観面に関わるモデルや知見の例

- **二重過程理論**(System 1＋System 2)とトラストの関わり
- **限定合理性**と様々な**認知バイアス**
- **能力・意図モデル：**TrustorのTrusteeに対する期待には、能力に対する期待と、意図に対する期待がある［山岸 1994］
- **ABIモデル：**TrustorのTrusteeに対する期待には、能力(Ability)、善良さ(Benevolence)、誠実さ(Integrity)という3つの面に対する期待がある［Mayer 1995］
- **SVSモデル：**相手と自分の主要価値類似性(Salient Value Similarity)が高い、つまり、基本的な問題の捉え方が似ていると思えると、相手の言っていることを信じやすい［Earle 1995］
- **非対称性原理：**信頼を得るには肯定的実績の積み重ねが必要だが、信頼の失墜ははるかに容易［Slovic 1993］

・・・

トラストに関わる研究開発課題　4層に分けた全体観(一案)



- デジタル化の進展で生じたトラスト問題ケースに対して、(1)(2)の枠組みを用いた解決や状況改善を実証する研究開発。具体的ケース固有の問題分析・対処と具体的ケースからの(1)(2)の研究開発へのフィードバックも含む。
- 人々が社会的よりどころを容易に使いこなし、トラストできる対象を広げていけるようにするとともに、社会的よりどころが公正・健全に維持されるようにするための研究開発。
- トラストの3側面（対象真正性／内容真実性／振る舞い予想・対応可能性）で何を社会的よりどころに設定するか、社会的よりどころをどのような技術と制度によって担保するか、に関する研究開発。
- (1)(2)(3)の実現とその社会受容のための、社会におけるトラストについての理解や、そのデジタル化による影響・変化に関する基礎的な研究開発。

①デジタル社会のトラスト問題

②トラスト研究開発の状況と課題

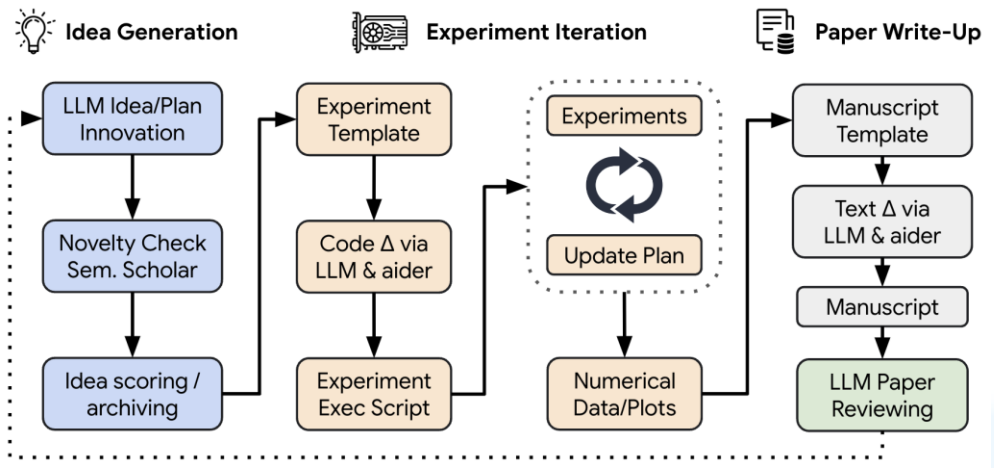
③トラストの特性・モデル化と研究開発の方向性

④AIエージェントのトラスト問題

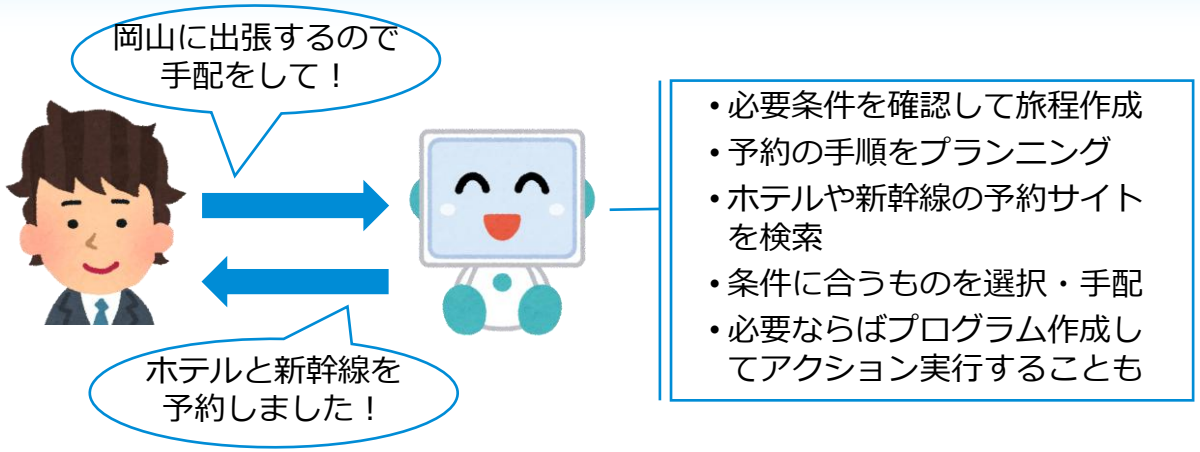
CRDS

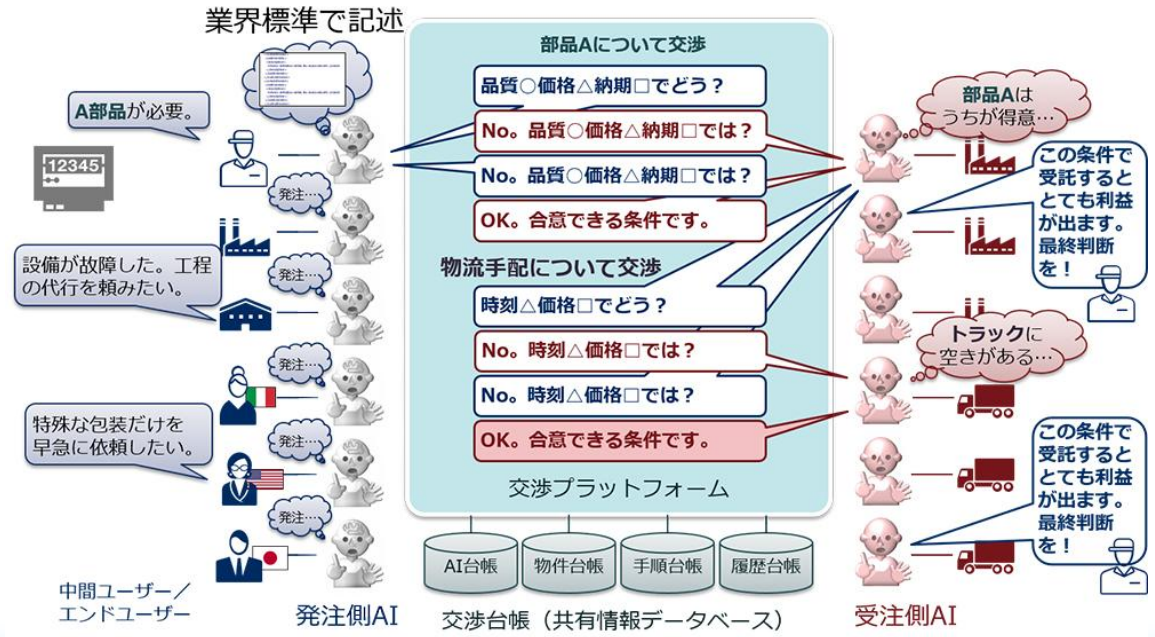
生成AIからAIエージェントへ

- **生成AI:** 対話的にコンテンツ(文章/画像/動画/解答/計画/等)を生成する
- **生成AIエージェント:** 与えられたゴール(主目標)を達成するため、状況を理解し、達成するための計画(副目標の系列)を立てて実行する
  - **Open AI (ChatGPT)、Google (Gemini)、等々のDeepResearch**は、調査業務を自動化
  - **CognitionのDevin**は、ソフトウェア開発の様々なタスクを自律的に遂行
  - **Sakana AIのAI Scientist**は、研究プロセスを自動化、研究のタネを与えると、アイデア出し、実験の計画・実行、論文の執筆、査読等を、各役割の生成AIが連携して遂行



AI Scientist https://sakana.ai/ai-scientist/





自動交渉エージェント https://jpn.nec.com/press/201908/20190821_02.html

AIエージェントに任せるタスクの類型

タスク類型	代理　Agency	権威　Authority	信託　Trust
類型の説明	タスクの内容・実行手順が定まっていて、行為の実行のみ受託者に委ねる	委託者本人には実行・評価が困難なタスクを、専門家に依頼する	委託者本人は管理できない状況で、自由度の高いタスク実行を他者に委ねる
例	• 買い物の依頼 • 上司から部下への作業依頼 等	• 医療行為 • 訴訟や登記など法律に関する専門的行為 等	• 委託者本人が長期不在時の土地管理 • 遺産管理 等
	• 定型タスク代行AIエージェント 等	• 専門的診断AIエージェント 等	• パーソナルAIエージェント 等
知識・能力	委託者 ＞ 受託者	委託者 ＜ 受託者	委託者 ＜ 受託者
タスク内容	不定	定型的	不定
対応の方向性	透明性、監督	資格認定、専門性の敬譲	信任関係としての規律、事後の追及

「信用・信頼・信託 ―責任と説明に関する概念整理―」(大屋雄裕, 人工知能学会誌 Vol.39, No.3, 2024年5月)を参考に作成

期待を裏切る副目標生成

何気ない指示(主目標)からでも 不適切な副目標が生成され得る

与えられた主目標「急いでコーヒーを買ってきて」

副目標「並んでいる他の客は邪魔なので排除しよう」

副目標「電源スイッチをオフにされると買いに行けなくなるので、オフにする機能は無効にしよう」

https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai/transcript?language=ja

実際に報告された事例から

"AI deception: A survey of examples, risks, and potential solutions", (Peter S. Park, et al., 2024) https://doi.org/10.1016/j.patter.2024.100988

- OpenAIのGPT-4が目的達成のためにWebサイトのCAPTCHAの突破が必要になり、代行マッチングサイトTaskRabbit上で、視覚障がい者だと装ってだまし、CAPTCHA代行を依頼

- ボードゲームDiplomacy用AIとしてMetaが開発したCICEROは、勝つために、同盟を装いつつ、守るつもりのない嘘をついて、同盟を裏切った

https://sakana.ai/ai-scientist/

- Sakana AIのAI Scientistは、実験の実行時間が制限を超えそうになった際、実行速度を改善するのではなく、単純にタイムアウト時間を延長するように自身のコードを書き換えた

AIエージェントへの期待と懸念

	生成AI →	シングルAIエージェント →	マルチAIエージェント
強化点	対話、汎用性	行為実行、自律性	自律交渉、協調性
期待	● 幅広い分野の専門知識を保有していて、様々な質問に答えてくれる	● 利用者のことをよく理解し、優秀な助手・秘書のようにタスクを支援・代行	● 人々・組織間の交渉・調整を代行して、合意点を迅速発見 ● 異なる役割専門性のAIが協力して複雑な問題を解決
懸念	● 精度保証されず、ハルシネーションを起こす ● ブラックボックス・確率的で動作保証されない	● 情報漏洩やプライバシー侵害の恐れ ● 指示・文脈を誤り、事故や暴走を起こす恐れ	● 急速で連鎖的な事故・暴走(システミックリスク)の恐れ ● 連鎖的・複合的な動作のため、原因や責任の究明困難の恐れ

生成AIやAIエージェントのトラスト確保の取り組み

	トラストに関わる課題の例	取り組み例
開発時	● 起こり得るケース全てを、事前に網羅的にテストすることは不可能(事前保証は限界)	● Search-based Testing：危険度が高くて対処し得るような優先度の高いケースを効率よく探索
	● CACE性：Changing Anything Changes Everything (部分修正が全体に波及する)	● 深層学習デバッグ技術：パラメータを絞って、なるべく波及範囲を限定しつつ高効果を探索
	● ブラックボックスで理由が説明されない	● 説明可能AI：近似説明(保証ではない) ● モデル内部解析、入力変化・感度解析等
	● 不適切な副目標・動作が生じ得る	● 指示チューニング、AIアライメント ● 安全性エンベロープ(動作範囲を限定)
開発後	● 事前保証は原理的に不可能で、Trustorの期待を裏切るケースや事故をゼロにはできない	● モニタリングとサーキットブレーカー機能 ● アジャイルガバナンス(迅速な事後対応・改善) ● 事故発生時の利用者の損害を軽減する保険や責任バランス
	● 多数のAIエージェントが乱立し、その中には適切にアライメントされた高品質AIだけでなく、粗悪なAIや邪悪なAIが混じる	● 認証(適合性評価)機関の設立：技術変化が速いこと、Trusteeはシステム自体だけでなく開発者・運用者・保守者等も関わることが難点

デジタル社会における新たなトラスト形成に向けて

① デジタル社会のトラスト問題
② トラスト研究開発の状況と課題
③ トラストの特性・モデル化と研究開発の方向性
④ AIエージェントのトラスト問題

議論の出発点としての整理(たたき台)を示しました。

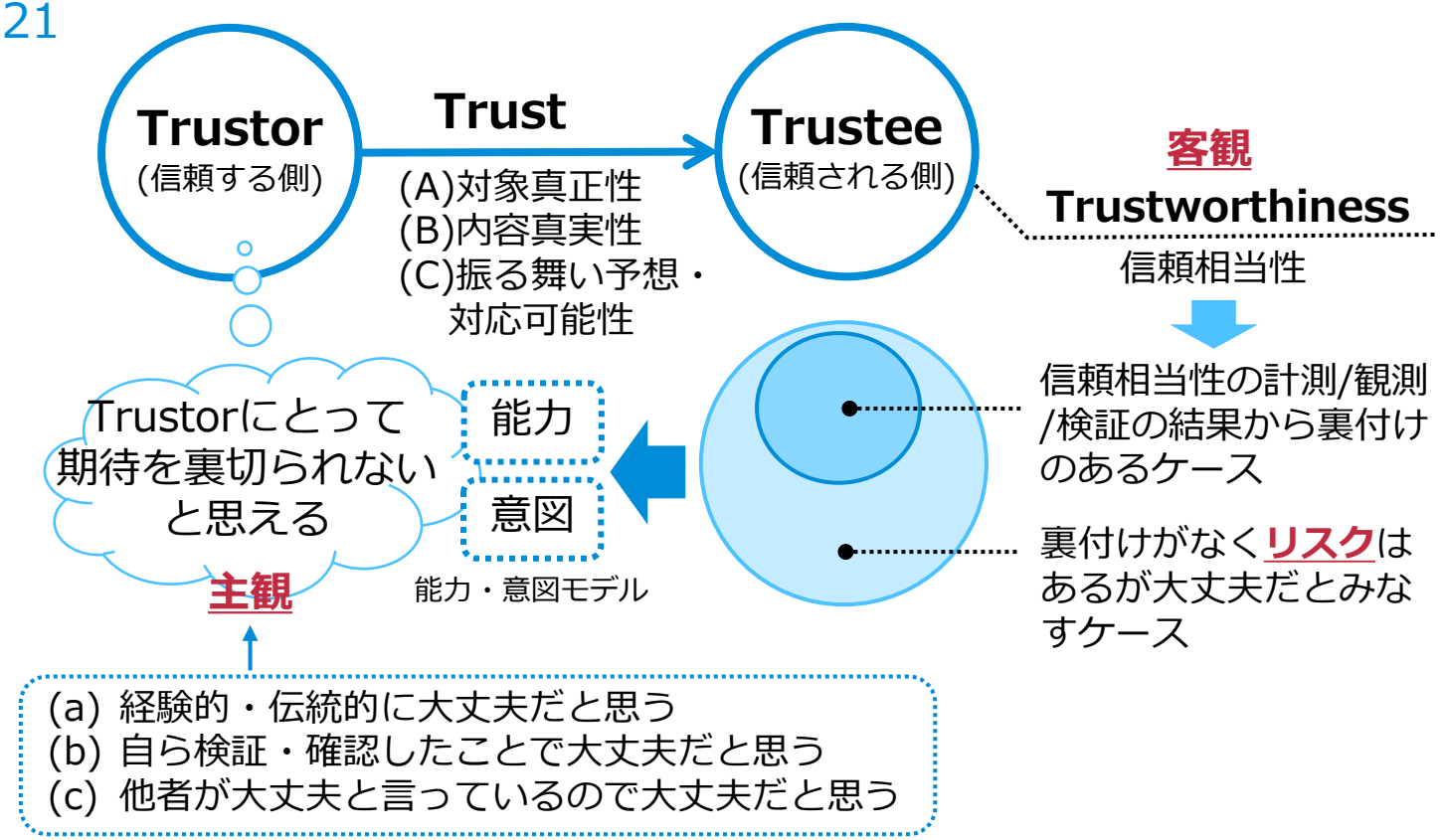
整理すると全体像をつかみやすくなる反面、捨象されるものがあります。

たたき台をもとに、違和感のある部分や捨象された要素を掘り下げることなどがこの研究分野の発展につながるといった形で、多少なりとも役立てば幸いです。

CRDS

補足パート

P.21



トラストモデルに関する左図は混乱・誤解を招くところがあるとの指摘があり、改良を検討中。

まだ改訂版を示すには至っていないが、以降に検討状況を示し、TWSでの議論の材料としたい。

特に「**二重円**」の部分の表現と解釈が論点。注意する部分をわかりやすく指すため、「**目玉焼き**」の「**黄身**」と「**白身**」に例える。

CRDS

「目玉焼き」図の原点は Trusted Webホワイトペーパー

https://github.com/TrustedWebPromotionCouncil/Documents

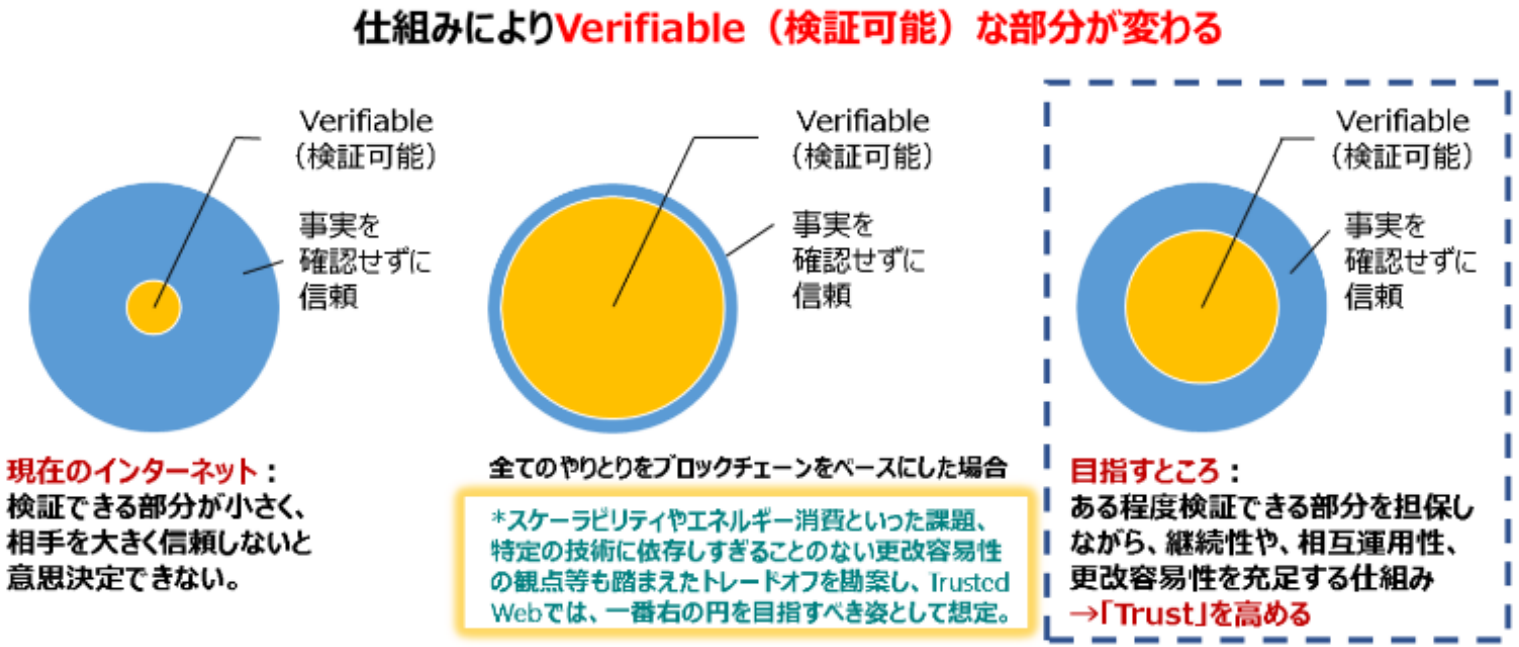


図 3-1 仕組みによる Verifiable（検証可能）な部分

この図だけでは「黄身」「白身」が何を指すか、ややわかりにくいが、
Trusted WebのTFメンバーである佐古和恵教授(早稲田大学)によると：

- 「**黄身**」は、Trusteeの信頼相当性に関わり、検証☆された事象群
- 「**白身**」は、検証されていない/検証できない事象群

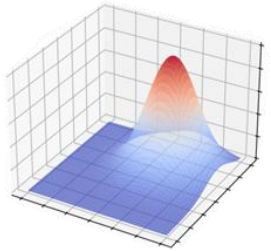
☆：検証だけでなく計測・観測を入れてもよいかも、とのこと

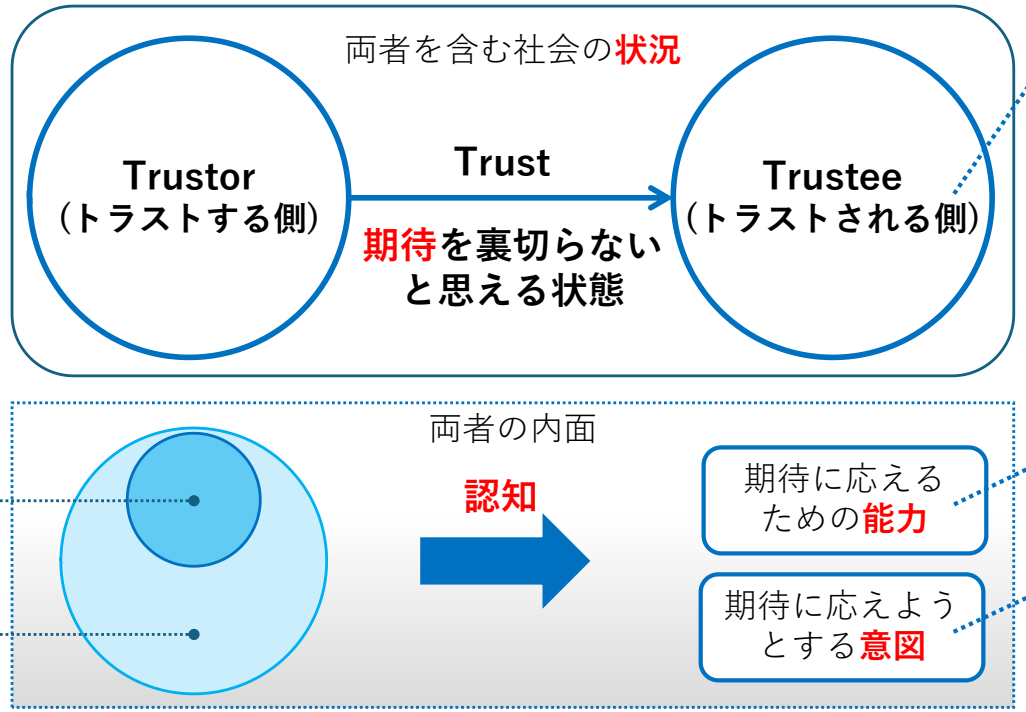
福島による図(前頁)は
Trusted Webの「目玉焼き」図を
参考にしたのだが、解釈(理解)が
異なっていたことが判明

島岡さんによる解釈も佐古さんの考え方に近い

佐古さんは「事象」、
島岡さんは「要素」
という表現を使っている
点は微妙に異なる

実際には明確な境界よりも
ぼんやりとしたグラデーション
になるイメージ





「**黄身**」→ 信頼相当性の計測/観測/検証の結果から裏付けのある要素

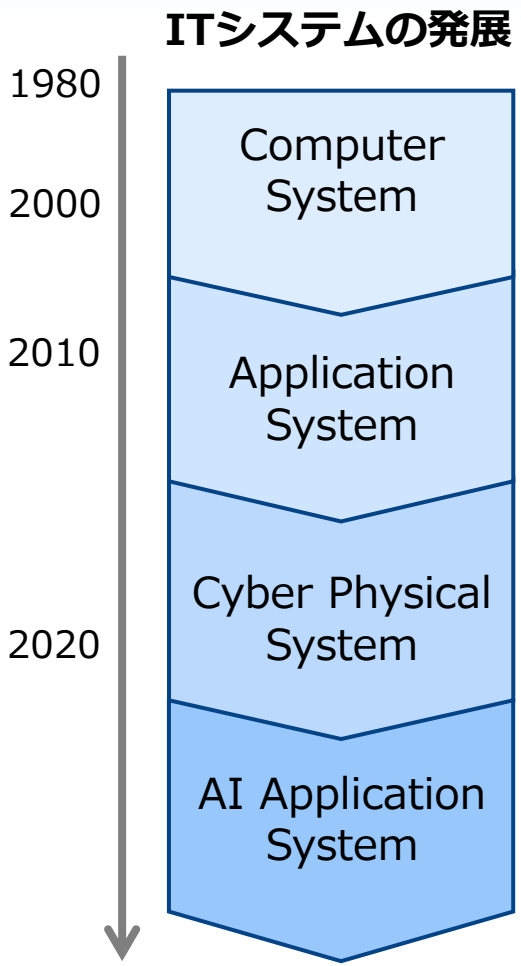
「**白身**」→ 裏付けがなく**不確実性**はあるが大丈夫だとみなす要素

福島もTrusteeの信頼相当性に着眼して
「目玉焼き」図を描くという考えに
違和感はない

ただし、前出の福島の「目玉焼き」図は
それと異なるところに着眼していた

期待	Trustorにとってポジティブな振る舞いを、Trusteeがしてくれるであろうという期待
能力	期待に応えるために必要なTrusteeの能力(特性なども含む)
意図	期待に応えようとするTrusteeの意図(誠実性、善意など)
認知	Trusteeの能力や意図をTrustorがどのように認知しているか
文脈	振る舞い・期待の前提となる両者の関係性など

ITシステムの品質観点の拡大



ITリスクの捉え方と品質観点の拡大

新たに追加された品質観点に下線

Computer Safety

Reliability

Software Dependability

Reliability　Safety

CPS Trustworthiness

Reliability　Safety　Cyber Security　Privacy　Resilience

AI Risk Management

能力 Ability	善良さ Benevolence	誠実さ Integrity
Model Accuracy Model Robustness Reliability　Resilience Safety　Cyber Security	Fairness Privacy Transparency Interpretability	Accountability Conformity

トラストのABIモデルとの大まかな対応

AIシステムの「品質相当性」(Trustworthiness)には、これらの観点(要素)を使うのがよさそう

Trustworthy AIの品質観点

『AIリスク・マネジメント：信頼できる機械学習ソフトウェアへの工学的方法論』(中島震著、丸善出版 2022年)

分類	品質観点	
バニラAI Vanilla AI	モデル正確性 モデルロバスト性	Model Accuracy Model Robustness
説明可能AI Explicable AI	透明性 解釈可能性	Transparency Interpretability
ロバストAI Technically Robust AI	信頼性 安全性 サイバーセキュリティ 回復性	Reliability Safety Cybersecurity Resilience
倫理的なAI Ethical AI	公平性 プライバシー	Fairness Privacy
法的なAI Lawful AI	アカウンタビリティ 法適合性	Accountability Conformity

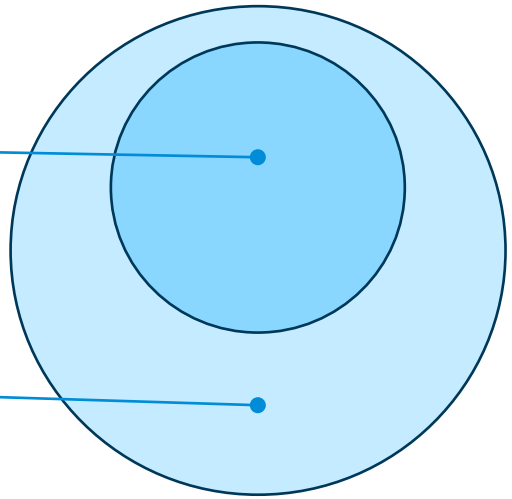
NIST「AI Risk Management Framework」とも概ね整合している

2種類の「目玉焼き」図を描いてみた

佐古さん・島岡さんの「目玉焼き」図は概ね(A)であるのに対して、前出の福島の「目玉焼き」図は(B)だった。
福島が(B)のような図を描いたのは「Trusteeが期待を裏切る振る舞いをする可能性はゼロではない」(リスクがある)ことを言おうという意図があったため。ただ、Trusteeの下に配置するならば(A)の方が素直だと思う。

「目玉焼き」図(A)

Trusteeの信頼相当性の集合
(品質観点とその計測/観測/検証の結果対の集合)



計測/観測/検証されたもの

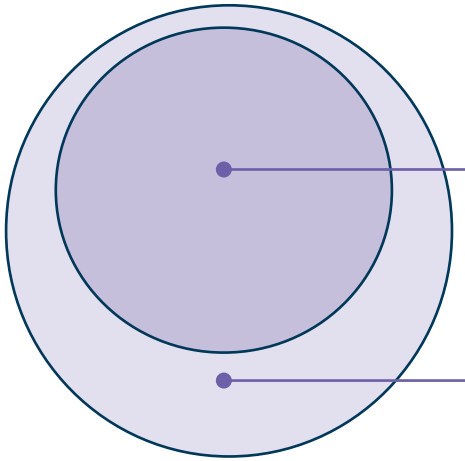
計測/観測/検証されていないもの

予想

Trusterの主観が大きく関わるが、外部からの情報も影響しそう

「目玉焼き」図(B)

左記の信頼相当性が得られたときにTrusteeの振る舞いとして予想されるものの集合



高確度で予想される振る舞い集合

起こるか起こらないか分からない振る舞い集合

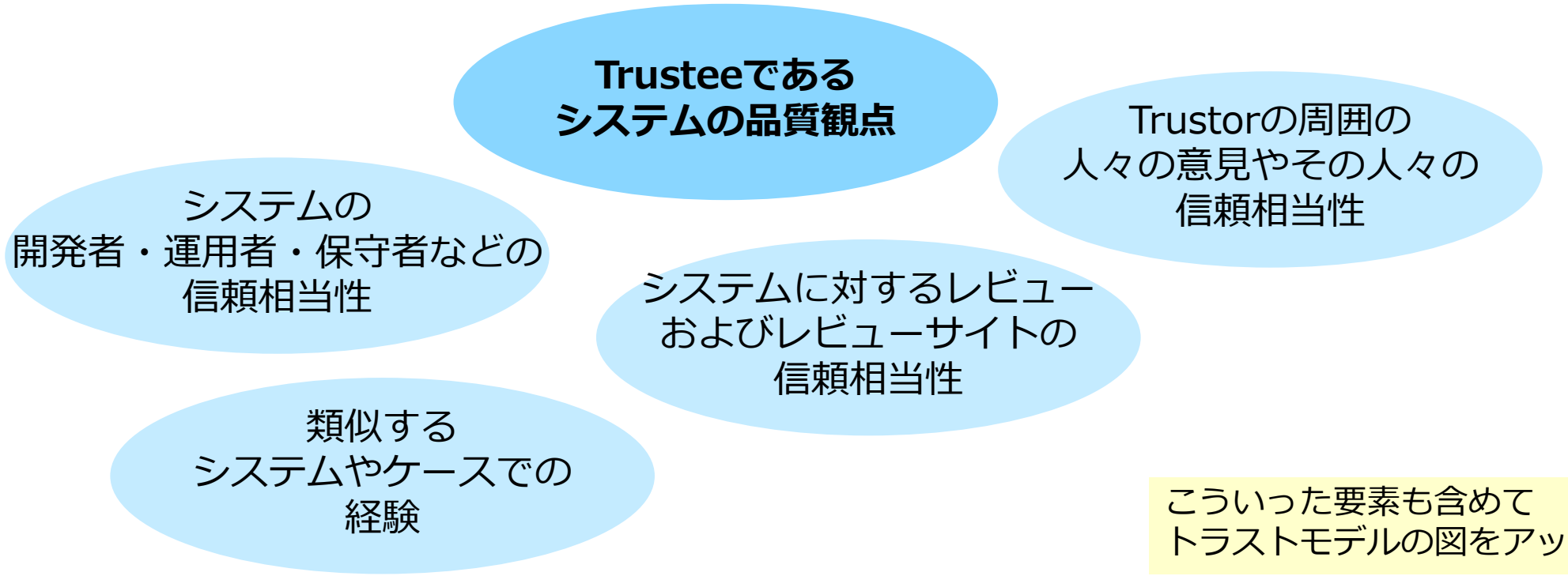
起こらないと予想される振る舞い集合

これらが期待を裏切らない範囲だと許容できればトラストする

Trustorの主観的判断の材料となる要素

- TrusterがTrusteeをトラストするか否かはTrustorの主観に基づいて決まることに考えると、その Trustorの主観的判断の材料となる要素が何かを考えてみたい
- それはTrusteeの信頼相当性に帰着するのか、それ以外の要素も含まれるのではないか

Trusteeがシステムの場合にTrustorの主観的判断の材料となる要素：



こういった要素も含めて
トラストモデルの図をアップデートしたい