AIのtrustworthinessの標準化状況
-- バイアスを中心に --

2021/06/18

産総研 江川尚志
ISO/IEC JTC1/SC42専門委員会幹事

Trust, Trustworthinessの流行

- 欧州：Excellence and Trust in AIを大方針に
  - 2021/04に発表された欧州AI規制は「Building trust through the first ever legal framework on AI」が売り
- 米国：大統領令ほかで頻出
  - Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (2020/12)
- 内閣府：Data Free Flow with Trust (DFFT)
  プライバシーやセキュリティ・知的財産権に関する信頼を確保しながら，ビジネスや社会課題の解決に有益なデータが国境を意識することなく自由に行き来する，国際的に自由なデータ流通の促進を目指す，というコンセプト
  「デジタル時代の新たなＩＴ政策大綱」（IT総合戦略本部，2019/06）

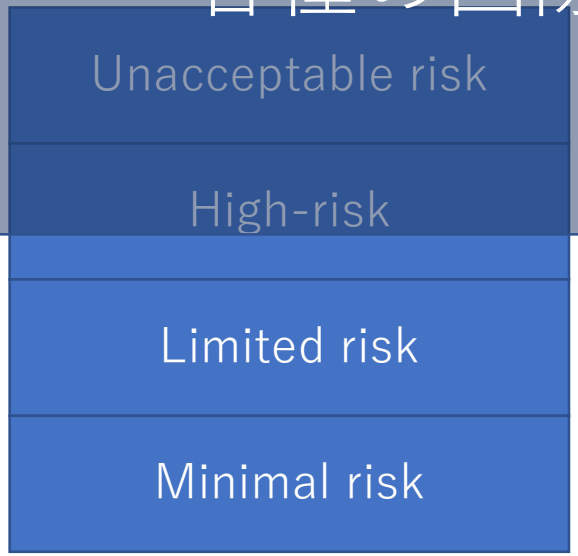
たぶん欧米には「人工知能」という概念がどうしても受け入れられない人々
→ Trustを強調。技術者も倫理 (ethics) と言われるよりは解る
日本も「信頼できないAIはダメ」に異論はない

日本語では信頼・トラスト、信頼性・トラストワージネス等と書く

欧州委員会 AI規制法案

- Proposal for a Regulation laying down harmonised rules on artificial intelligence
- 2021/04発表
- AIを規制する世界初の包括的ハードロー

おそらく、この実現に
各種の国際標準が使われる



禁止; 信用スコアリング、法執行機関による公共空間でのリアルタイムのバイオメトリクスに基づく人物識別 他

適合性評価など各種の法的義務; 医療機器等すでに規制されている用途 + 重要インフラや教育など新規分野

AIであることの明示による透明性の確保他; チャットボットやディープフェイク 他

特段の規制なし、自主行動計画等は奨励

Trustworthyの定義: 欧州

- human agency and oversight (人間の関与)
- technical robustness and safety (ロバスト、安全)
- privacy and data governance (プライバシーとデータガバナンス)
- Transparency (透明性)
- diversity, non-discrimination and fairness (多様性、非差別、公平)
- environmental and societal well-being and (環境と社会のwell-being(福祉))
- Accountability (アカウンタビリティ)

ALTAI - The Assessment List on Trustworthy Artificial Intelligence
High-level Expert Group on AI（2020/07)

そもそも論が大好きな欧州らしい定義

Trustworthinessの定義: ISO/IEC

TR 24028:2020 (overview of trustworthiness)での定義

ability to meet stakeholders' expectations in a verifiable way (ステークホルダーの期待に検証可能な形で答える能力)

Note 1 to entry: Depending on the context or sector, and also on the specific product or service, data, and technology used, different characteristics apply and need verification to ensure stakeholders expectations are met. (分野ごとに顧客の期待は異なる)

Note 2 to entry: Characteristics of trustworthiness include, for instance, **reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, usability.** (特性は様々)

Note 3 to entry: Trustworthiness is an attribute that can be applied to services, products, technology, data and information as well as, in the context of governance, to organizations. (対象も様々)

ISO/IEC JTC1/WG13 (trustworthiness) はTS 5723 'trustworthiness vocabulary'を策定中。
上記を出発点とするが、今後いろいろと変更される可能性

国際標準の策定場所

国際標準策定機関を特定した条約は存在しない。血が流れる。実務的には

- ISO/IEC/ITU (公的な国際標準。たぶん)
  (米国的にはグローバルを目指すべき時に「国」という古臭い概念に引きずられた組織)
  - **ISO/IEC JTC1/SC42 (AI)**
    - WG1; 用語とマネジメント規格、WG2: データ一般、
      WG3: trustworthiness, WG4: ユースケースとアプリケーション、
      WG5: 計算論的アプローチ、JWG1: ガバナンス
  - ISO/IEC JTC1/SC7 (software)
  - ISO/IEC JTC1/WG13 (trustworthiness)
  - 個別分野には各々グループがある
- IEEE (欧州的には、米国の勝手な文書)
  (米国的には「志ある個人が世界中から集まり作った真にグローバルな文書」)
  - IEEE P70xxが有名だが、例えばP2976 (explainability) なども
- CEN/CENELEC/ETSI (欧州の法律に従う標準を作る機関。欧州規制を考えると外せない)
  - CEN/CENELEC JTC21 (AI)

国際標準の種類

- ISO, IEC, ISO/IEC JTC1: 例: ISO/IEC TR 24028
  - IS (International Standard), 国際標準
  - TS (Technical Specification), 技術仕様書。標準化を急ぎ査読を簡略化。将来的にIS化を想定
  - TR (Technical Report), 技術報告。強制力なし
- ITU: Recommendation; 例: ITU-T Y.3001
  - Recommendation (勧告), Supplement (補遺)
- CEN/CENELEC
  - ほぼISO/IECと同様。ISではなくEN (European Standard)
- IEEE
  - Standard (標準)
  - Recommended practice (推薦)
  - Guide (ガイド)
  - Trial-use documents (3年間の時限文書)

Trustworthiness関連標準の状況 (1/2)

- 全体概観: ISO/IEC TR 24028; 出版済
- Verifiability / Quality / Accuracy
  - ISO/IEC TR 29119-11; Testing AI-Based Systems; 出版済
  - ISO/IEC 25059; SQuaRE4AI—Quality model for AI systems; WD
  - ISO/IEC 5471; Quality Evaluation Guidelines for AI Systems; AWI
  - ISO/IEC TS 4213: WG5 Machine learning classification performance; DTS
- Availability/ Resiliency / Reliability / Robustness
  - ISO/IEC TR 24029: Robustness of NN
    Part 1: overview, 出版済, Part 2: 形式手法, WD
  - Maintenance: 活動なし
  - Calibration: 活動なし
  - ISO/IEC TS 8200: Controllability of automated artificial intelligence systems: 開始可否の投票中
- Bias
  - ISO/IEC TR 24027; Bias in AI systems and AI aided decision making; DTR
  - IEEE P7003; Algorithmic Bias Considerations

完成度はISO, IEC文書では IS: AWI → WD → CD → DIS → FDIS → IS (出版)
TS: AWI → WD → DTS → TS (出版)
TR: AWI → WD → DTR → TR (出版)

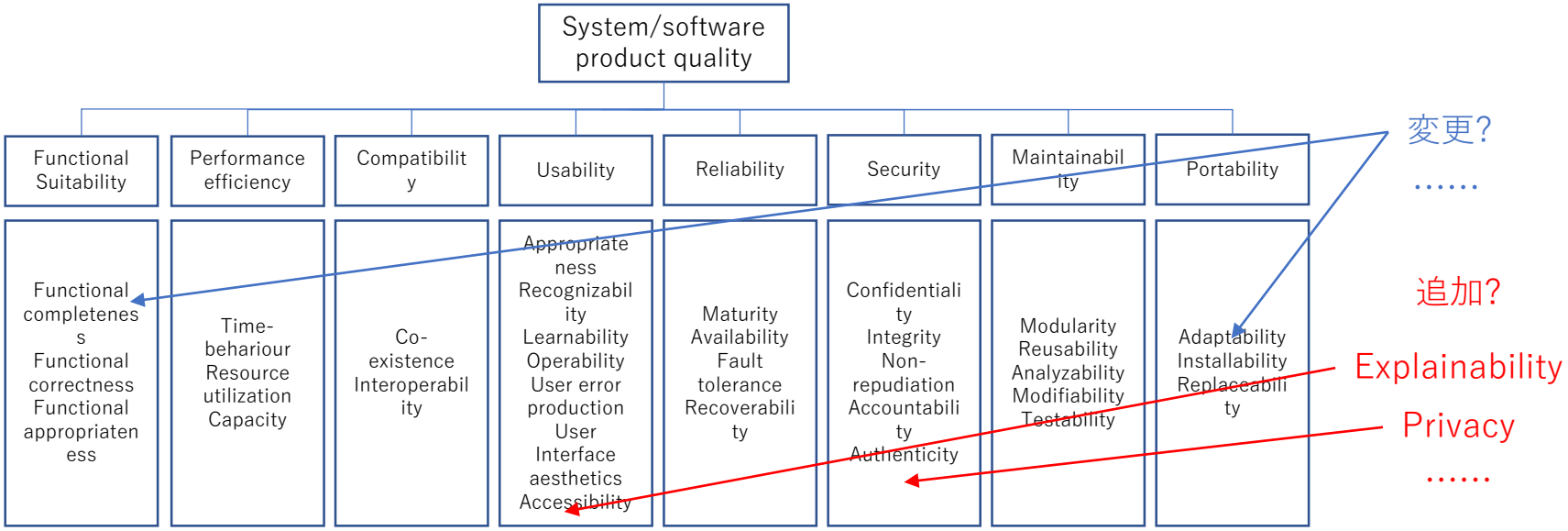
Trustworthiness関連標準の状況 (2/2)

- Safety
  - ISO/IEC TR 5469; Functional Safety and AI systems; WD
  - IEEE P7009; Fail-Safe Design of Autonomous and Semi-Autonomous Systems
- Ethics
  - ISO/IEC TR 24368; Overview of ethical and societal concerns (TR 24368); WD
  - IEEE Ethically Aligned Design (EAD): 出版済
  - IEEE ECPAIS; Transparency, Bias, Accountabilityに関して試行的な認証基準を策定
- Transparency
  - ISO/IEC TS 6254; Objectives and approaches for explainability of ML models and AI systems; AWI
  - IEEE P7001; Transparency of Autonomous Systems; 第１回目の投票終了
  - IEEE P2976; XAI – eXplainable Artificial Intelligence
- Security
  - ISO/IEC JTC1/SC27/WG4で議論開始
- Privacy
  - ISO/IEC JTC1/SC27/WG5で議論開始

完成度はIS: AWI → WD → CD → DIS → FDIS → 出版
TS: AWI → WD → DTS → 出版
TR: AWI → WD → DTR → 出版

ISO/IEC 25059: AI品質モデル

- ISO/IEC 25059, Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality Model for AI-based systems
- 目的：ソフトウェア品質とは何かを規定するISO/IEC 25010を拡張し、AIのソフトウェアで必要となる品質の一覧を規定
- エディタ：Adam Leon Smith (英)
- 開始2020/05、完成目標2023/04
- 対になる文書: TS 5471, 新規・変更された品質について実現方法例を詳述



ベースとなるISO/IEC 25010の製品品質

ISO/IEC TR24028: 信頼性: 概観

- 名称: ISO/IEC TR 24028
  Information technology — Artificial intelligence (AI) — Overview of trustworthiness in artificial intelligence
- 発行: 2020/05
- 作成: ISO/IEC JTC1/SC42/WG3
  Editor: Orit Levin (US)
- AIにおける信頼性 (Trustworthiness) の概念を整理
  - 既存の信頼性の枠組み
  - 関係者
  - ハイレベルの懸念: 責任/アカウンタビリティ/ガバナンス、安全性
  - 脅威の所在: セキュリティ、プライバシー、バイアス、予測性、ブラックボックス、仕様記述、実装、利用、ハードウェア故障
  - 対策: 透明性、説明性、制御性、バイアス軽減、プライバシー確保、信頼性/耐故障性/ロバスト性、耐ハードウェア故障、機能安全、試験と評価、利用

ISO/IEC TR24028: 信頼性: 用語

- **Autonomy, Autonomous**: characteristic of a system governed by its own rules as the result of self-learning
- **Bias**: favouritism towards some things, people, or groups over others (何かの物、人、グループを他よりも好むこと)
- **Explainability**: level of understanding how the AI-based system came up with a given result
- **Harm**: injury or damage to the health of people, or damage to property or the environment (ISO/IEC Guide 51:2014)
- **Interpretability**: level of understanding how the underlying (AI) technology works
- **Trust**: degree to which a user or other stakeholder has confidence that a product or system will behave as intended (ISO/IEC 25010:2011)
- **Trustworthiness**: ability to meet stakeholders' expectations in a verifiable way

ISO/IEC TR24028: 信頼性: ハイレベルの懸念

- 責任/アカウンタビリティ (答責性) /ガバナンス
  多数のステークホルダー間で各々の責任 (responsibility) を定め、それを他のステークホルダーに説明 (accountability) できることが必要。
  このフレームワークとしてはISO/IEC 38500 (組織におけるITのガバナンス) や 38505 (データの管理) が有用
- 安全性
  有形のharmだけではなく、無形のharm (例えば社会環境や文化環境に対するもの) も含む。
  あらゆる製品、あらゆるシステムはリスクを持つ。それを許容範囲に収める必要がある（IEC Guide 51より）。
  これらはライフサイクル全体について考慮する必要がある

ISO/IEC TR24028: 信頼性: 脅威の所在

AI特有の脅威として下記を列挙

- セキュリティ; 通常のITの脅威 (バグ他) に加えdata poisoning, adversarial attacks, model stealing, ハードウェアを狙ったconfidentialityやintegrityへの攻撃
- プライバシー; データ収集、プリプロセスとモデル化、モデルへの問い合わせの各段階でプライバシー問題が発生し得る
- バイアス; バイアスを「特定の物、人、グループを好むこと」としたうえで、指標を定義することの重要性、不注意な指標の害を記述
- 予測性; 人がシステムのふるまいを予測できることは望ましく、時に必須
- ブラックボックス (opaqueness); モデル自体の解釈が難しい場合もあれば、データやアルゴリズムの透明性不足によるブラックボックス化もある
- 仕様記述; 問題の大半は仕様策定に原因、遠因がある、、、
- 実装; データ収集とpreparation, モデル作成、モデルアップデート、ソフトのバグが各々脅威の原因に
- 利用; 不適切な利用 (誤用、悪用、不使用他)、AIを人と誤認させる
- ハードウェア故障; 多様な問題を引き起こす

ISO/IEC TR24028: 信頼性: 対策

AIへの脅威に対する対策として下記を列挙

- 透明性; 特徴、構成要素、処理などが外部から分かること。各種情報が外部監査に対して提供され、ステークホルダーは要求が満たされていることを確認可能に。ラベリングなども時に有益
- 説明性; 説明性の目的 (因果関係/原因の認知/弁明), 事前 (システムのふるまいの理解) /事後 (原因追及)、現状、レベル分けなどを概観
- 制御性; オペレータが制御を引継ぎHuman-in-the-loop他を実現
- バイアス軽減; 法など外部環境の分析、データ由来のリスクの分析、モデル学習中の技術的対策、試験、トライアルなど様々な方法が存在
- プライバシー確保; 基本は匿名化だが限界も
- 信頼性/耐故障性/ロバスト性; 広いカバー範囲を持つデータで訓練他
- 耐ハードウェア故障; 従来手法を様々に活用、ライフサイクル全体で実現
- 機能安全; 安全性確保のための機能を追加
- 試験と評価; 形式手法、経験則、人との比較、シミュレーションの活用、フィールドトライアル他
- 利用

バイアス標準の状況

- ISO/IEC TR 24027: Bias in AI systems and AI aided decision making
  - エディタ: 英DragonFly
  - 完成目標: 2021/01
  - ステータス: DTR
  - 内容: バイアスの定義、分類、回避方法の列挙
- IEEE P7003
  - 議長: ノッティンガム大
  - 完成期限: 2021/12
  - ステータス: ドラフト作成中
  - 内容：ライフサイクルの各ステップ毎の要求条件を規定
- IEEE ECPAIS
  - 認証に関するパイロットプロジェクト
  - Transparency, Bias, accountabilityについて試験的な基準策定済
  - BiasのリーダーはP7003議長

バイアスの定義: ISO/IEC 24027

- ISO/IEC TR 24027; 2021/01 DTR投票版
  systematic difference in treatment of certain objects, people, or groups in comparison to others
  Note 1 to entry: Treatment is any kind of action, including perception, observation, representation, prediction, or decision
- 一般にバイアスは下記の３種類の意味を持つ
  1. 目標からの偏り
  2. 悪いもの
  3. 人間の認知機能に含まれる偏り（偉い人は常に上、王子さまは常に白馬、魔王は闇魔法）

(1)として定義。IEEE P7003も同様

ISO/IEC TR 24027: バイアス

- 名称：ISO/IEC TR 24027, Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making
  (技術報告書: AIシステムおよびAIに助けられた意思決定におけるバイアス)
- 策定: JTC1/SC42/WG3
- エディタ: Adam Leon Smith (英 DragonFly)
- 開始: 2018/12; 現在DTR; 完成目標: 2021/01
- Scope: バイアスを評価するための計測手法を記述し、バイアスに起因する脆弱性への対処を目指す。AIシステムの全ライフサイクル、データ収集から利用までを対象とする
- TRであり「～せよ (shall)」「～すべき (should, recommend)」等々の言い回しは厳しく排除されている。あくまでも「こんな風に考えるという方法があります」止まり
- TRが落ち着いたら、TS (Technical Specification) やIS (International Standard) の作成の必要が議論されよう

ISO/IEC TR 24027: 章構成

5. An overview of bias (概論)
6. Fairness (公平さ); バイアスはfairnessを損なうものの一つ
7. Sources of bias in AI systems (バイアス源)
   Human cognitive biases, Data bias, Machine learning model architecture bias, Bias in rule-based system design, Requirements bias
8. Assessment of bias and fairness in AI systems (評価)
   Classification systems, Reinforcement learning systems, Confusion matrix, Equalized odds, Equality of opportunity, Parity, Predictive equality
9. Treatment of bias throughout an AI system life cycle (ライフサイクルの中での扱い)
   Inception, Design and development, Verification and validation, Deployment

Annex A (informative) Related open source tools (OSSを列挙)

Annex B (informative) ISO 26000 – mapping example (26000, 社会的責任、とのマッピング)

Questions?