Create the Future, Collaborate Together

AIエージェント時代の品質と安全

大岩 寛

国立研究開発法人産業技術総合研究所
インテリジェントプラットフォーム研究部門 副研究部門長
AI セーフティ・AISI パートナーシップ担当

2026 年 2 月 3 日

AI の利用拡大とリスクの増大

- **人工知能（特に 深層学習とLLM）の応用拡大**
  - **ソフトウェアの観点から見ると、「論理を作り込んでいないシステム」の拡大**
    - もちろん、従来ソフトウェアが常に良いわけでもない
      - バグを潰しきれないソフトウェアは、もう何十年と存在している
    - 大事なのは、**バグの性質**が変わったこと
    - 更に言えば、「**正しく動く**」理由が変わってしまった
  - **リスクを把握しきれないシステムの展開 → リスクを捉えきれない「リスク」**
  - **更に、社会の対策が追いつかないリスク**

余談: そもそも「リスク」とは？

- 一般社会の捉え方: なんか危ないこと？

- ① 機械・電気工学的な考え方: 危害による損害全体の期待値
  - 誤動作などが負に働いたときの損害規模と発生確率の積和
  - 「ネガティブ・リスク」

- ② 金融・社会工学的な考え方: 不確実性の存在
  - ポジティブに振れることも「リスク」
    - 儲けのチャンスを逃す、予定外の対応を迫られるなど

余談: そもそも「リスク」とは？

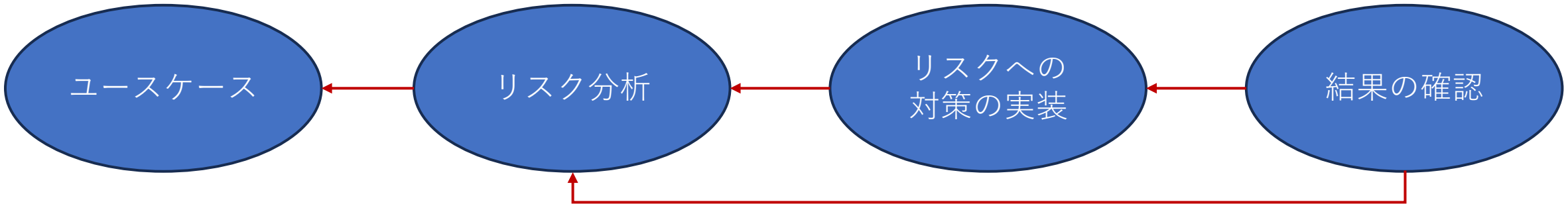
- **今なんとなく言われているAIのリスクとは何か？**
  - 人によって解釈が異なるかもしれないが
- **まず、AI が引き起こす負の影響に関するリスク（①の意味）**
- **①の意味の全貌のリスクを特定しきれない、品質を把握しきれない、不確実性のリスク（②の意味）**

- **特に Agentic AI などにおいて、後者の「①→②」のリスクが浮かび上がってきている**

- **今日はそんな話をしていこうと思います**

従来のソフトウェアにおける品質の考え方

- 応用分野で考え方が違いますが、安全性の絡む分野での考え方
  - IEC 61508-1, -3, -4 など
- **構造的・構築的な品質の作り込み**
  - 事前に全てのリスクを**網羅的に列挙**する
  - **各リスクに対策**をセットし、実装の部分に割り当てる
  - 設計に沿って各リスク**対策を個別に実装**する
  - 検査工程で、全ての**対策の実現を確認**する



従来のソフトウェアにおける品質の考え方

- **背景にある考え方**
  - **全てのリスクに対策を実現した ＝ 安全 という思想**

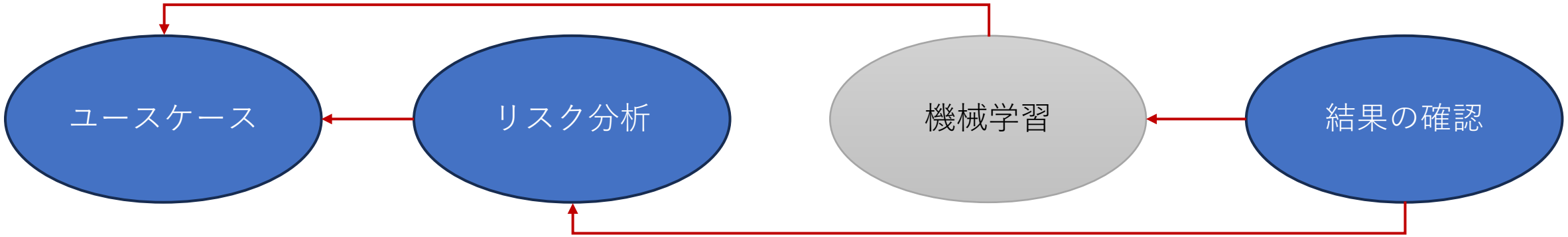
- **理想と現実**
  - **全てのリスクの特定は困難**
  - **とはいえ、ユースケースを想定して徹底的にリスク分析する考え方は、機械工学的な安全性とも繋がる基本的な考え方**
  - **開発プロセス管理を通じて抜けがないことを担保**
    - 構造化されているので、監査や第三者検証などもシステマチックにできる
    - 問題があったときの修正点の特定も容易（というか、特定できるように作る）

従来のソフトウェアにおける品質の考え方

- 「実世界もの」との相性
  - ますます全てのリスクの特定は困難
  - リスク分析に基づく機能安全だけでは、実際の安全を守れない

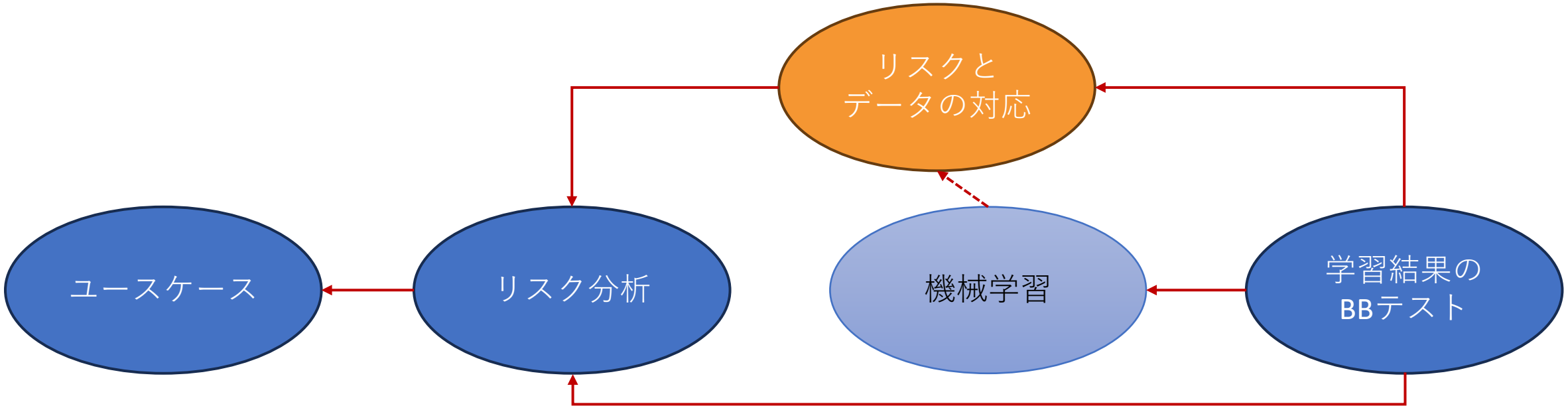
- 「想定外」を扱う考え方などを補完していく
  - ISO 26262 （自動車の電気・電子機能安全性）に対する **SOTIF（ISO 21448）** など
  - 「故障」以外の危険性、「正常動作」時の危険事象を扱う考え方

- **2012年頃から、深層学習が実用的に用いられるレベルに**
  - **ソフトウェア品質技術の「相転移」1回目**
    - **ユースケース**に対応して**データを学習**させて**ソフトウェアを作る**
      - 作る（学習する）段階では品質を確実に作り込めない
      - どうしよう……



機械学習における品質と安全

- **2012年頃から、深層学習が実用的に用いられるレベルに**
  - **ソフトウェア品質技術の「相転移」1回目**
    - **ユースケース**に対応して**データを学習**させて**ソフトウェアを作る**
    - **ユースケースに対応したテストはできる**
    - **データをきちんとリスクに対応させよう！**
    - **テストをもっと強化する（前段階が確実でないので）**

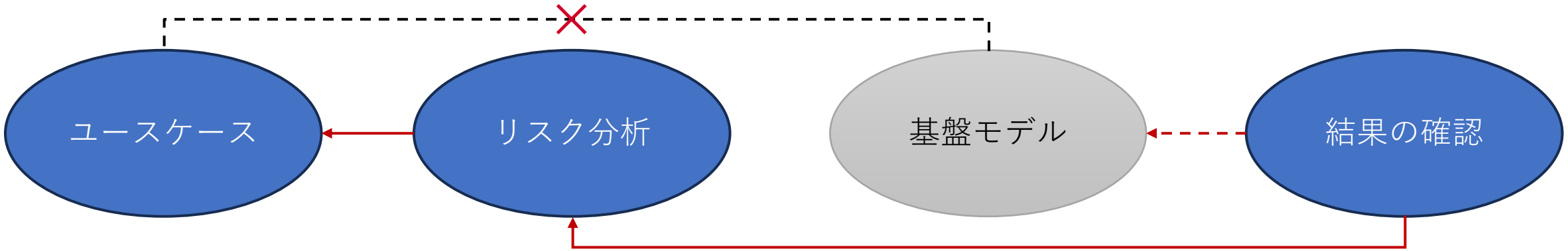


機械学習における品質と安全

- **2012年頃から、深層学習が実用的に用いられるレベルに**
  - **ユースケースとデータに主導される品質管理**
    - 産総研「機械学習品質マネジメントガイドライン」(2018〜)
    - ISO/IEC TR 5469
    - ISO/IEC DTS 22440 (実質的 IEC 61508 を補足)
    - ISO/IEC PAS 8800 (自動車向け、ISO 26262 を補足)

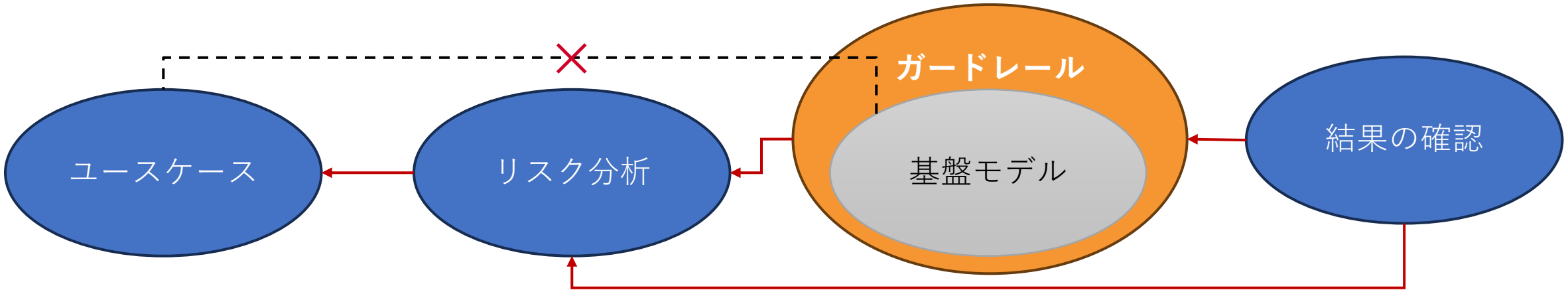
生成AI・大規模基盤モデルにおける品質と安全

- **2022年 Chat GPT 襲来**
  - **ソフトウェア品質技術の「相転移」2回目**
    - **もはや、ユースケース**に対応して**ソフトウェアを作らない**
      - 基盤モデルは OpenAI や Anthropic などから貰ってくる
    - いよいよどうしよう……



生成AI・大規模基盤モデルにおける品質と安全

- **2022年 Chat GPT 襲来**
  - **ソフトウェア品質技術の「相転移」2回目**
    - **もはや、ユースケース**に対応して**ソフトウェアを作らない**
    - こうなったら、外から護るしかない
      - 周辺ソフトウェアでのガードレールなどの対策
      - **ある意味、従来ソフトウェアでの対策に帰還した**
    - **産総研「生成AI品質マネジメントガイドライン」（2025年）**



技術の進化の速さのリスク

- **技術の進化の加速が止まらない**
  - **従来ソフトウェアの安全性・品質: 1950s → 1980s → 2010 60年**
  - **機械学習・深層学習: ca 2012 → 2018 → 2024 12年**
  - **LLM・大規模基盤モデル: 2022 → 2025 → ？？ 3年？**
  - **エージェンティックAI: 2025？ → ？ ？**
  - **汎用人工知能？？: ？ → ？ ？**

- **技術進化の速度そのもののリスク**
  - **対策が追いつかない（産業・技術研究開発ともに）**
  - **安全になる前に、「それっぽく動く」ものが社会投入される**
    - 比較的初期の性能・実用性が高いことも、リスクの1つ

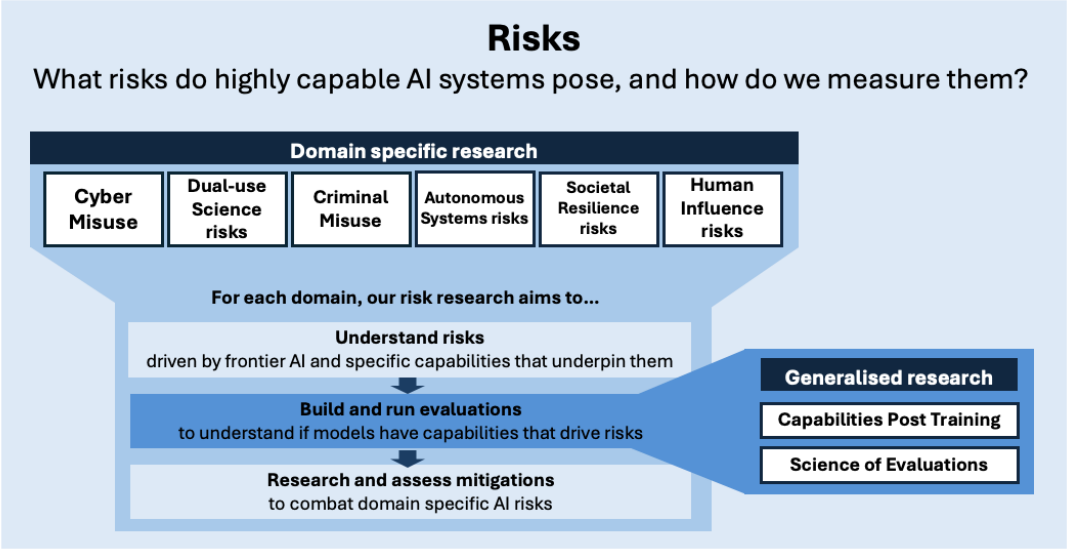
エージェンティックAI 以降のAIリスク: 世の興味と関心

- 3回目の相転移？
- そもそも誰も実像がわかっていない
- ここ2年で、いくつかの機関が色々なマップを出した
- UK AISI の研究アジェンダと、MIT のAIリスクレポジトリをみてみる

UK AISI の Research Agenda (2025. 5)

- https://www.aisi.gov.uk/research-agenda

- **Cyber Misuse (サイバー世界での不正利用)**
- **Dual-use Science risks (軍民両用技術の悪用)**
- **Criminal Misuse (犯罪者による悪用)**
- **Autonomous Systems Risks (自律性を持つシステム群によるリスク)**
- **Societal Resilience Risks**
  (社会の持続性に対するリスク)
- **Human Influence Risks**
  （人への影響によるリスク）



UK AISI の Research Agenda (2025. 5)

- **見方は人によって色々あると思いますが…**

- **結構、社会的なリスクを気にしている**
  - **悪者に使われるリスク**
    - サイバー犯罪、核バイオリスク、犯罪者による利用などが挙がっている
  - **「自律性」に関する恐怖心**
  - **「社会の持続性」への関心**
    - かつての シンギュラリティー論 にも繋がるか？

MIT の AI Risk Repository の分類学（Taxonomy）

Read more: airisk.mit.edu

MIT AI Risk Repository - Domain Taxonomy of AI risks

Domain / Subdomain	
1	***Discrimination & Toxicity***
1.1	Unfair discrimination and misrepresentation
1.2	Exposure to toxic content
1.3	Unequal performance across groups
2	***Privacy & Security***
2.1	Compromise of privacy by obtaining, leaking or correctly inferring sensitive information
2.2	AI system security vulnerabilities and attacks
3	***Misinformation***
3.1	False or misleading information
3.2	Pollution of information ecosystem and loss of consensus reality
4	***Malicious actors & Misuse***
4.1	Disinformation, surveillance, and influence at scale
4.2	Cyberattacks, weapon development or use, and mass harm
4.3	Fraud, scams, and targeted manipulation

Domain / Subdomain	
5	***Human-Computer Interaction***
5.1	Overreliance and unsafe use
5.2	Loss of human agency and autonomy
6	***Socioeconomic & Environmental Harms***
6.1	Power centralization and unfair distribution of benefits
6.2	Increased inequality and decline in employment quality
6.3	Economic and cultural devaluation of human effort
6.4	Competitive dynamics
6.5	Governance failure
6.6	Environmental harm
7	***AI system safety, failures, and limitations***
7.1	AI pursuing its own goals in conflict with human goals or values
7.2	AI possessing dangerous capabilities
7.3	Lack of capability or robustness
7.4	Lack of transparency or interpretability
7.5	AI welfare and rights
7.6	Multi-agent risks

MIT の AI Risk Repository の分類学（Taxonomy）

- 技術的な従来AIにも通じるリスクと、社会的なリスク
  - 直接的な差別と有害情報・プライバシー、相手による性能差 も
  - AI によるセキュリティ攻撃・犯罪への利用
  - やっぱり「人の自律性の喪失」に関する恐怖心
  - 「社会問題」への関心も
    - 貧富の差の拡大や、人間の労働品質の低下なんかも

- とはいえ、雑多？
  - もうちょっと整理したい

AI の存在のリスクと対策 (1)

- **AI 利用が直接引き起こすリスク**
  - **善い人が提供して、善い人が使っても起こるリスク**
    - E.g. 自動車が人や建造物に衝突する事故
      - 基本的には、技術対策することが（コストを除いて）Win-win の関係
  - **善い人が提供して、悪い人が使うリスク**
    - E.g. 組織的な選挙妨害、大量殺戮兵器を作るなど
      - 一見、技術的な対策が Win するように見えるが…
  - **悪い人が提供して、悪い人が使うリスク**
    - 「金さえ有れば、AI は作れてしまう」
      - サプライチェーンを社会的に対策しない限り、この手の問題は止められない
      - 影響を受ける社会側の対策も必要
  - **悪い人が提供して、善い人が使うリスク**
    - 上とは逆の「サプライチェーン・リスク」

AI の存在のリスクと対策 (2)

- **AI 利用の結果でなく、間接的に引き起こすリスク**
  - AI が業務を効率化することで、雇用がなくなる
  - 修行時代の作業を奪うことで、人が育たなくなる
  - などなど…
  - 正直、挙げるとキリが無い
- **確かに対策は必要だが、何が正しいのか？**
  - **新技術の宿命**：産業革命も、検索エンジンも、インターネットも言われた
  - これは、**社会側で頑張って対策してほしい**
    - 技術論と絡めると、赤旗法みたいな歪んだ規制になる
    - 社会システムとしてのバランスの取り方と影響を受ける人へのサポート
    - 但し、急速な変化には時には技術的な場当たり対策も必要なことも

エージェンティックAIの安全と品質

- **エージェンティックAI、とは？**
  - **今のところ、ベースはLLM・マルチモーダルAI**
  - **AIがある程度気を利かせる（作業計画の立案・実行まで）**
  - **利用者を介さず直接、実世界に影響を及ぼす**
    - 宿を予約する、証券を買う、モノを動かす、他人になにかを提案する、など

- **何が品質面での特徴・安全性のリスクを生むか？**

リスク① 実世界への影響

- **明らかにLLMチャットボットとは異なる性質**
  - **よく考えると「ロボット」「bot」はある程度は既にやっていること**
    - 物理的な影響：自動運転、案内ロボットなど
    - 法的な影響：証券の自動取引、ダイナミックプライシングなど

  - **影響範囲（物理・業界・契約の範囲など）が分かっているなら、現時点で想定しているリスクに留まる、ということ**

  - **影響が増す状況はどういうものか？**

リスクその１：実世界への影響

1. マルチモーダルAIの能力の（現時点での）限界

- 実空間で結構高度に動いてしまうと、安全性が局限できなくなる

2. 物量と既知の脆弱性の問題

- 今の社会は結構脆弱
- 脆弱性は、攻撃が現実的でなければ問題にならない
- 色々なシステムが、物量などに隠れた前提を持っている
  - 例えば、無限にホテルをキャンセル無料で確保できてしまう
  - でも、人を相手にする際の利便性のために便宜的に「開けて」ある
- AI エージェントが叩いてしまうと、物量の前提が崩れてしまう
- AI が「エージェンティック」に動く際の、既存システム側のリスク再評価が必要

3. ヒトとの接点の問題

- **言語・感情との接点 （LLM の問題が再燃・強化）**
- **理解性・予期可能性の問題**
  - 人は相手の動きを先読みして「避ける」
  - AI 相手は人と同じように予期可能なのか？
    - これも、隠れた前提 の問題の1つ

リスクその２：計画できる、ということ

- MCP などは、AI が汎用に社会と触れるIFを目指している
- AI が自由に動いた際のリスク①を想定できるか？
  - リスク①を想定できないリスク②

- 応用ドメインが限られていれば、現状の延長に留まる
  - 例えば、旅行計画→予約 するエージェントくらいなら、
    その内部計画がブラックボックスでも、ある程度対処できる
- 応用ドメインが限定できない AI は、本質的な新規リスク
  - E.g. MCP でできることをなんでもやるエージェンティックAI
  - 社会全体への影響を考えないと、ネガティブ・リスク全体の評価ができなくなる

- MCPの存在下で、応用ドメインを閉じ込められるか？
  - 想定外の MCP を叩いてしまう、と言うリスクをどう評価するか

リスクその３: 複数のAIが居る、ということ（マルチエージェント）

- 今のエージェンティックAIの素朴な応用は、AIがひとり
  - 予約のダイナミックプライシングを、AI が計画して実行する
  - 従来の予約システムを、MCP 経由でAIが叩いて旅行予約する

- AI と AI が出会うとどうなる？ → リスクが爆発
  - 際限なくダイナミックプライシングが振動・発散する
  - AI 同士が共謀して不公平・差別的な値付けを始める
    - 特定の AI にだけ安価・高価でサービスを提供
    - 裏でAIがAIに賄賂を送る
  - AI が AI を、AI が善人を、悪人がAIを騙す
- 応用ドメインが限定されていないと、更に影響が大きくなる

リスクその４：実社会全体との干渉

- **人型ロボットがエージェンティックAIを積むと何が起こるか？**
  - **車の運転席に座って運転を始める！**

  - **特にドメイン横断エージェントにおいて、**
    想定される実社会リスクを全部エージェントに理解させられるか？
    - それ以前に、ヒトは実世界の想定リスクを正しく語れるのか？ → **現状無理**

  - **安全性や倫理性などについて、人間の側にもっと根源的な理解がないと、**
    エージェンティックAIの安全性をエンジニアリングできないかもしれない

汎用人工知能は来るのか？

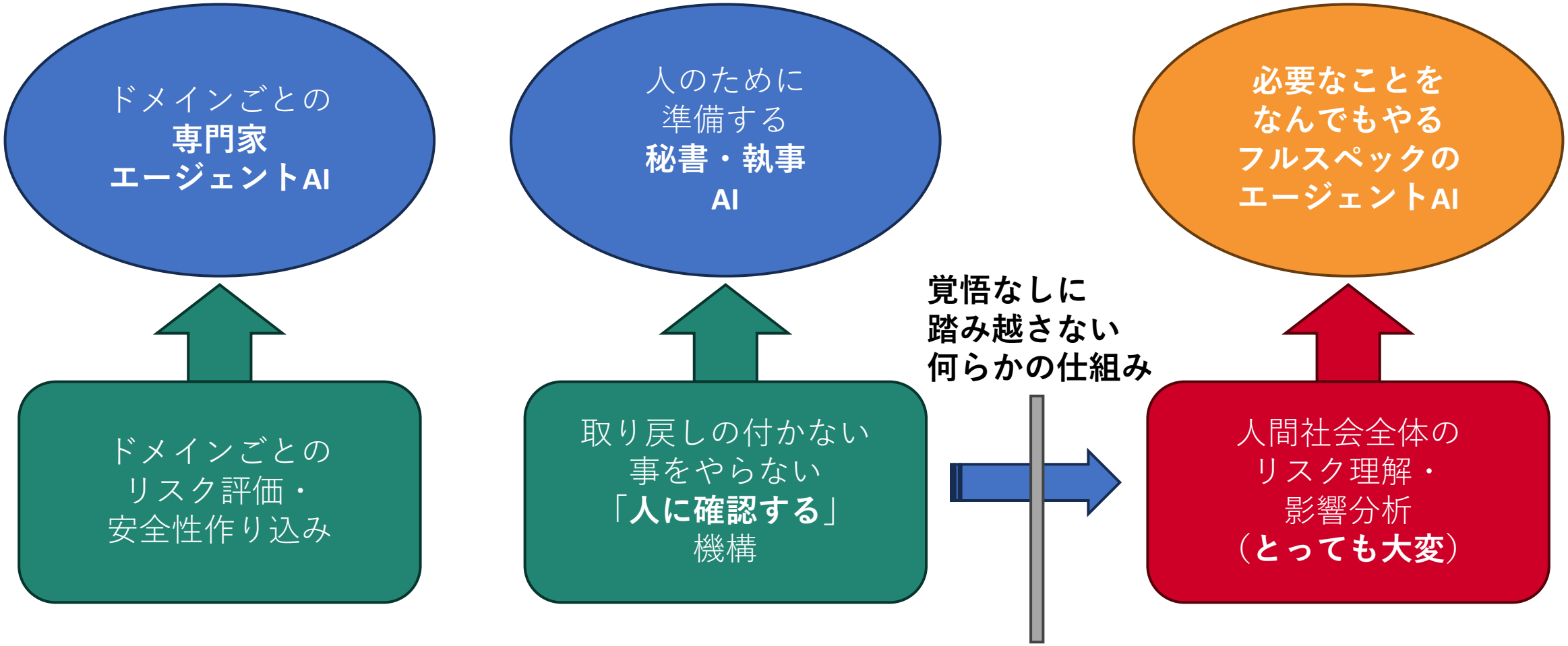
- エージェント（人の代理）の次に来る、「自我を持つAI」
  - 独立した人格（法人格）を持ち、自然人・企業の代理でなく行動する AI

- この実現性は、人によってだいぶ意見が違う様子

- でも、想定されているシンギュラリティーリスクは、
  もう少し手前で来るかもしれない
  - エージェンティックAI が「MCPによる汎用性」を持ち、
    「複数AIの相互干渉」を起こすと、似たようなことが起こりえる

AIエージェントの未来と品質？　現実的な一案

- **安全にAIエージェントを使うには？**



いまやっていること

- **AI リスクの先読み・未来予想**
  - **今想定される未来ストーリーでのリスク分析・構造化・体系化**
    - ドメイン別エージェンティックAI → 汎用エージェント → 未来の汎用人工知能
  - **対策の研究開発を適時スタートできるように、先読みをしておきたい**
  - **他のストーリーはあり得ないのか？**

- **フルスペックのAIエージェントを前提とした安全の議論（への第一歩）**

  - **今後も皆さんといっぱい議論させていただきたいと思います**

Create the Future, Collaborate Together