AIのガバナンス

東京大学国際高等研究所東京カレッジ
理化学研究所　AIPセンター
江間有沙

AIの問題は社会の問題

クイズ：このシステムは合憲？違憲？

アメリカのある州では、過去の犯罪データと照合して、
犯罪者の再犯可能性を予測するシステムが使われている
システムの判断に基づいて、裁判官が判決を下す

システムがアングロサクソン系よりもアフリカ系の人たちのリスクを高評価すると指摘
なぜそうなるか信頼性が検証できないため、法の適正な手続きを受ける被告の権利侵害
システムの利用について州最高裁で争われた

	コケージャン（白人）	アフリカ系アメリカ人（黒人）
再犯率が高いと予測されたが実際には再犯しなかった率	23.5%	44.9%
再犯率が低いと予測されたが実際に再犯した率	47.7%	28.0%

A. 合憲：引き続きこのシステムを使ってもよい
B. 違憲：このシステムは使ってはいけない

出典：https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

再犯リスク予測AI　COMPASの事例

社会的規準「公平」を数学的表現に落とし込む

報道機関

「再犯率が高いと予測されたが、実際には再犯しなかった率」が黒人のほうが多かった。黒人のほうが「再犯する」と思われる偏見のある不公平なシステムである！

「再犯率が低いと予測されたが、実際には再犯しなかった率」が人種に寄らず一致するので公平なシステムである！

開発者

両方を同時に満たす公平性は存在しないので
何が「公平」かは社会的に決めて技術に反映させる必要がある
→様々な人たちの協力が大事になる

AIと差別をめぐる事件簿

- 学習
  - SNSのコメントで、機械が差別的発言をする

    →**SNSユーザ**が悪意を持って差別発言を大量に学習させた
    →システム凍結

- 認識
  - 顔認識システムで、浅黒い肌の女性の認識精度/正答率が低い

    →**学習データ**に「浅黒い肌の女性」データが少なかった
    →警察など公的機関での利用を制限

- 評価
  - 採用判定システムが、女性に関する単語が含まれていると評価を下げる

    →**学習データ**に「女性」データが少なかった
    →そもそも**現実**の技術コミュニティに、女性が少ない
    →システム開発を断念

合憲：システムを利用してよい

ただし！

- システムは参考として使用し、総合的に考えて人間が最終判断をする
- システムの判決に人種の偏りがあると利用者に注意を促す必要がある

- 人間がAIの課題を認識しつつ、利用することが重要となる

人工知能とは

- 1956年に開催されたダートマス会議で、アメリカの計算機科学者であるジョン・マッカーシーが提唱
- 人間の知的活動を定義し、技術的に実現する研究分野
- 大量のデータを学習することで、自ら特徴やパターンを抽出し、初めて見るデータでも分類、識別や予測ができる技術
- 人工知能技術のうち、機械学習は統計学を基礎としているため、過去と未来は変わらないという前提で認識や予測を行う
- つまり、**現在の社会に存在する差別や偏見も再生産してしまう**

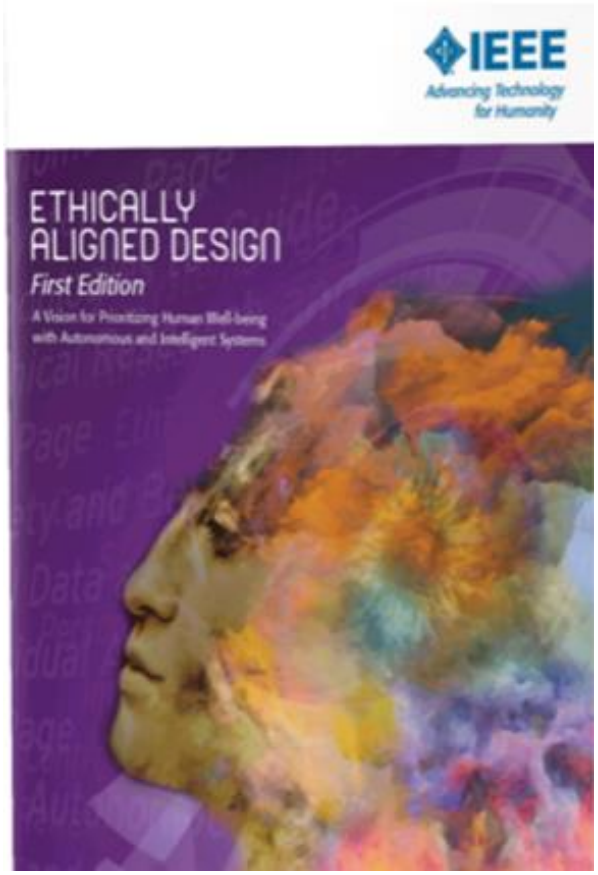
どういう基準でシステムを設計すれば公平？

- 個人の公平性：どのようなクラス（性別・人種・所属等）に所属していても、他の人と同様に扱われる
- 集団の公平性：二つのクラスが同等に扱われる（男性・女性等）
- 結果の平等：各人口統計学的なクラスが同じ結果を得る
- 機会の平等：どのようなクラスであっても、正しく分類される機会が平等である
  - 一方で、不平等な結果をもたらす可能性もある

AIのガバナンス

AIサービス管理のキーワード

- 原則から実践へ

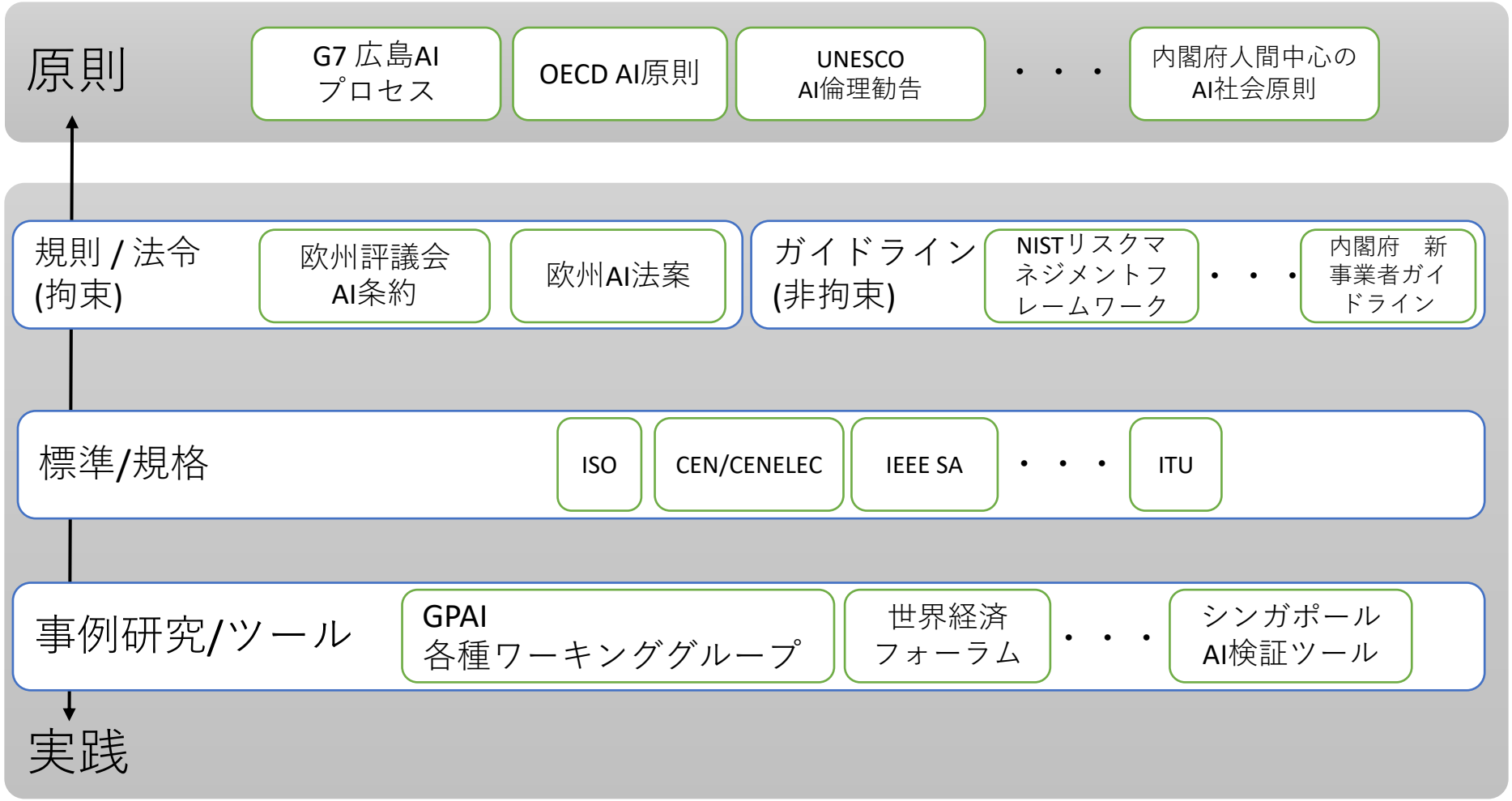


- ガバナンス
  - 社会情勢や技術の進展に伴い、議論されるべき内容や目的設定は**更新し続ける必要**があり、そのためには政府、企業、大学、研究機関、一般の人々等、**多様なステークホルダーが協働**してルール、制度、標準化、行動規範等のガバナンスについて**問題を設定し、影響を評価し、意思決定を行う**とともに**実装できる体制**が整っていることが必要（内閣府『人間中心のAI社会原則』[2019]より）

AIガバナンスの議論

	国内	国際
2016		**G7香川・高松情報通信大臣会合**：AI研究開発原則ガイドライン案提案 **PAI**：8つの信条（Tenets)
2017	**総務省**：AI開発ガイドライン	**FLI**：アシロマAI原則
2018		
2019	**総務省**：AI利活用ガイドライン **内閣府**：人間中心のAI社会原則	**OECD**：AIに関するOECD原則 **G20茨城つくば貿易・デジタル経済大臣会合**：G20AI原則の合意
2020		**GPAI**：4つのWGで議論開始
2021	**経産省**：AI原則実践のためのガバナンス・ガイドライン	**UNESCO**：AIの倫理勧告 **EU**：AI規制法案の発表
2022		**CoE（欧州評議会）**：AI条約のゼロドラフト公開
2023	**総務省・経産省：**統合ガイドライン検討開始	**G7広島サミット**：広島AIプロセス **英国**：AI安全性サミット
2024	**内閣府**：AI事業者ガイドライン	**CoE（欧州評議会）**：AI条約締結 **欧州連合**：AI規制法案採択

下線はハードロー


原則
G7 広島AI
プロセス
OECD AI原則
UNESCO
AI倫理勧告
・・・
内閣府人間中心の
AI社会原則
規則 / 法令
(拘束)
欧州評議会
AI条約
欧州AI法案
ガイドライン
(非拘束)
NISTリスクマ
ネジメントフ
レームワーク
・・・
内閣府　新
事業者ガイ
ドライン
標準/規格
ISO
CEN/CENELEC
IEEE SA
・・・
ITU
事例研究/ツール
GPAI
各種ワーキンググループ
世界経済
フォーラム
・・・
シンガポール
AI検証ツール
実践

議論のポイント①
AIとリスク

1. AIがもたらすリスク
   - ① AI（機械学習）固有の特徴により生じるリスク
     - 制御不可能性や透明性の低下等
   - ② IT技術に共通するリスク
     - プライバシー、セキュリティ、データの偏り、悪用、誤回答、秘密漏えい、著作権や肖像権を侵害するリスク等

2. AI推進の障害となる問題
   - ① 現行制度がAIの設計・開発・運用・利用等の障害となり、AIサービスの適用関係が不明瞭でイノベーションが阻害される
     - 医師法などの業法、著作権法、個人情報保護法等

https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2023/09/policy_recommendation_tg_20230915j.pdf

議論のポイント②
相互運用性（Interoperability）

- G7広島サミットのコミュニケ等でAIガバナンスの「枠組み間の相互運用性」がキーワードに

- 相互運用性とは？
  1. 国際標準
     - ISO, IEEE, NIST, CEN-CENELECなどの標準化
  2. 「枠組み間の相互運用性」
     - リスク特定、評価、対応等の「枠組み」の相互運用性（透明性）を確保することで各国・地域、組織のAIに対する規律が並存、協調できるようにする

https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2023/09/policy_recommendation_tg_20230915j.pdf

様々な規律の在り方

		国家によるエンフォース（強制的な執行）の有無	
		国家がエンフォースする	国家はエンフォースしない
規律を形成する主体	国家	法令・一部ガイドライン	ガイドライン・行政指導
	産業界・個別企業	法令を背景とする標準化（強制規格）	業界ガイドライン・社内ポリシー・業界標準
	それ以外（市民団体・学術団体等）	慣習法	市場・投資・モラル・規範・学会基準・慣習・評判

https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2023/09/policy_recommendation_tg_20230915j.pdf

Ecosystem of AI governance / AIガバナンスのエコシステム

Companies / 企業

- Corporate vision / ビジョン
- AI Principles / AI原則
- Risk Assessment / リスク評価
- Risk Control / リスクコントロール
- Employee/ Employer Education / 教育

Private Sector and Industry Association / 民間と業界団体

- Guidelines / ガイドライン
- Standards / 標準
- Audit / 監査
- Insurance / 保険
- Fact Check / ファクトチェック

Academia / 学術団体

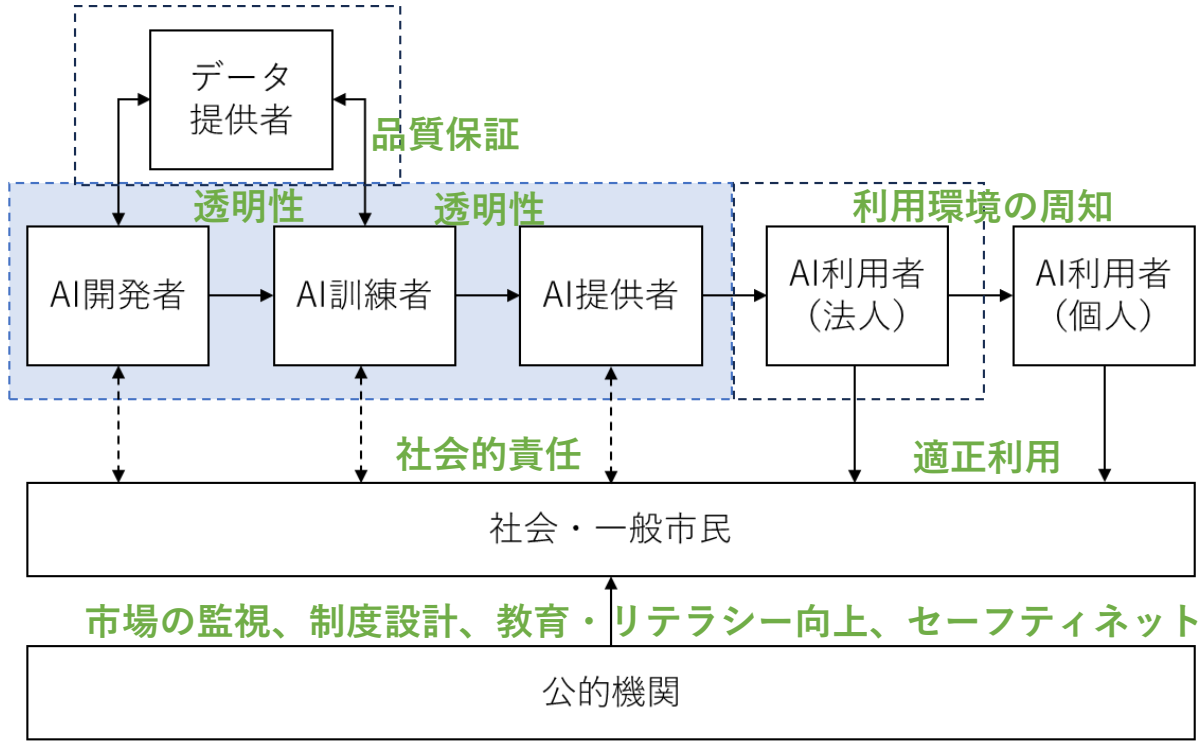
Public Sector / 公的機関

- Hard Law, Soft Law / 法や規制
- Third-Party Incident Committee / 事故調査制度
- Whistleblower System / 内部告発制度

Users / 利用者

- Education / 教育
  - Engineer / エンジニア
  - Experts / 専門家
  - Policy makers / 政策関係者
  - General publics / 一般

議論のポイント③
AIシステム関係者と責任



AI開発者・AI学習者・AI提供者は同一組織の場合もあれば、別々の組織のあ場合もる。データ提供者とAI開発者が同一組織の場合もある。自社製品を利用する場合は、AI開発者/AI提供者と利用者（法人）が同一組織となる場合もありうる。

→ 責任の向きを示す。例えばAI開発者はAI学訓練とデータ提供者に対して責任を負う（表3）。

⇢ 企業の社会的責任や、社会からのフィードバック等の間接的な責任の向きを示す。

■事業者間（青枠内）

- **事業者間**で契約や約束を取り交わし、リスク対応や責任を取る
- **公的機関**は契約の公平性等を監視する

■事業者ー消費者間

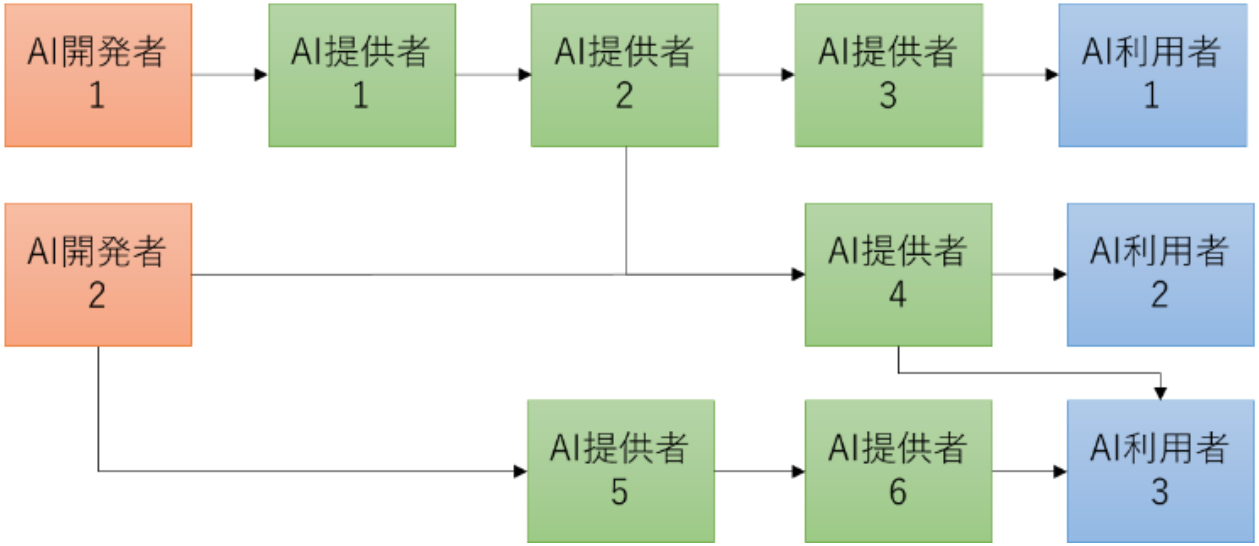
- **AI提供者**が事前・事後対応を行う
- **利用者**も適正利用する
- **公的機関**も事件・事故時の原因究明や被害救済の仕組等を形成する

https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2023/09/policy_recommendation_tg_20230915j.pdf

組織や国をまたぐAI開発・提供

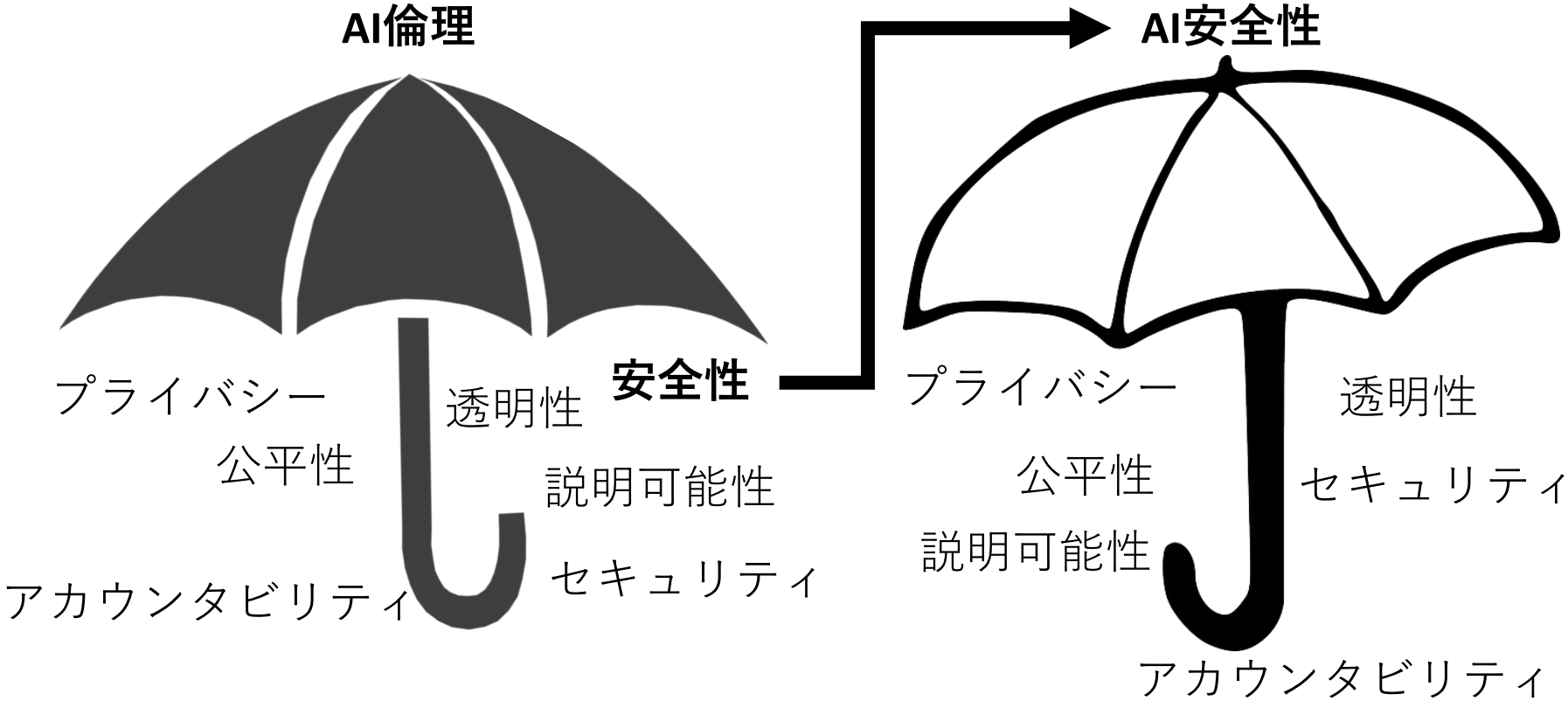
(1) AI開発から提供まで同一組織が担う例

(2) AI開発から提供までが異なる組織で担われている複雑な例



https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2023/09/policy_recommendation_tg_20230915j.pdf

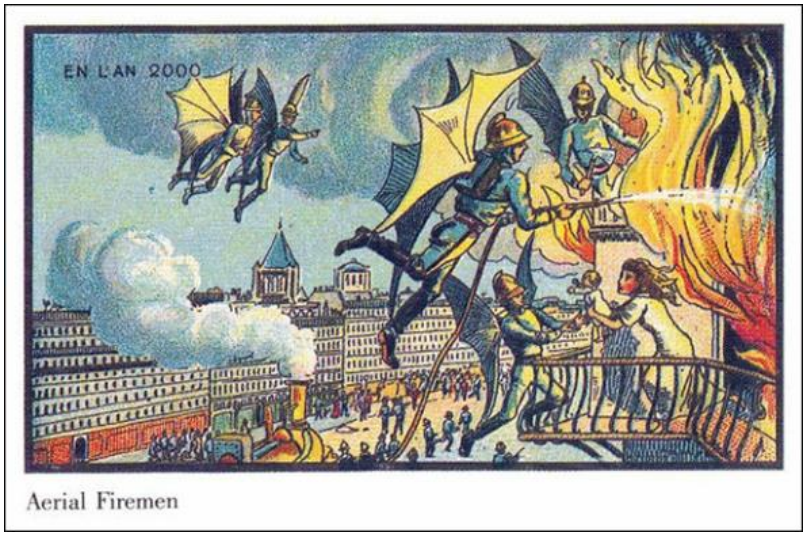
英国AIサミットの功罪


AI倫理
プライバシー
公平性
アカウンタビリティ
透明性
説明可能性
セキュリティ
安全性
AI安全性
プライバシー
公平性
説明可能性
透明性
セキュリティ
アカウンタビリティ

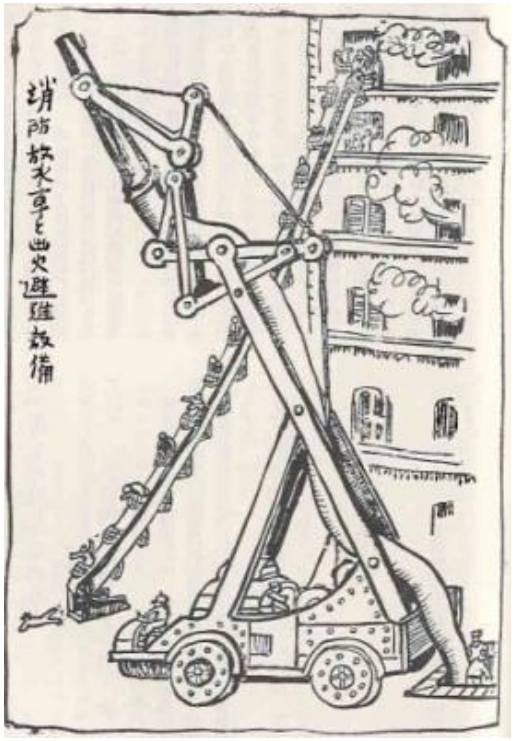
働き方・生き方の多様化

科学技術と社会：未来を考える

- 科学技術の社会面
- 社会の中の科学技術
  - どういう社会になっていくか
  - どういう未来に私たちは住みたいのか

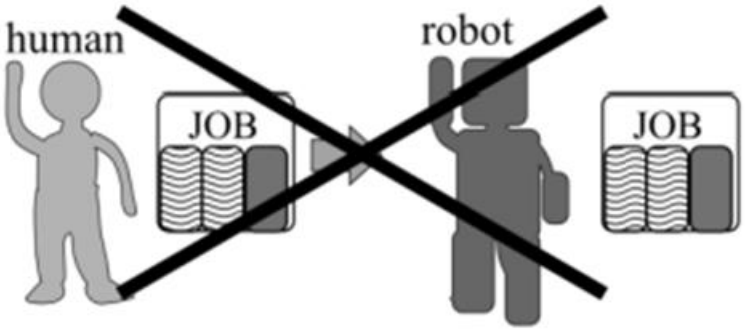


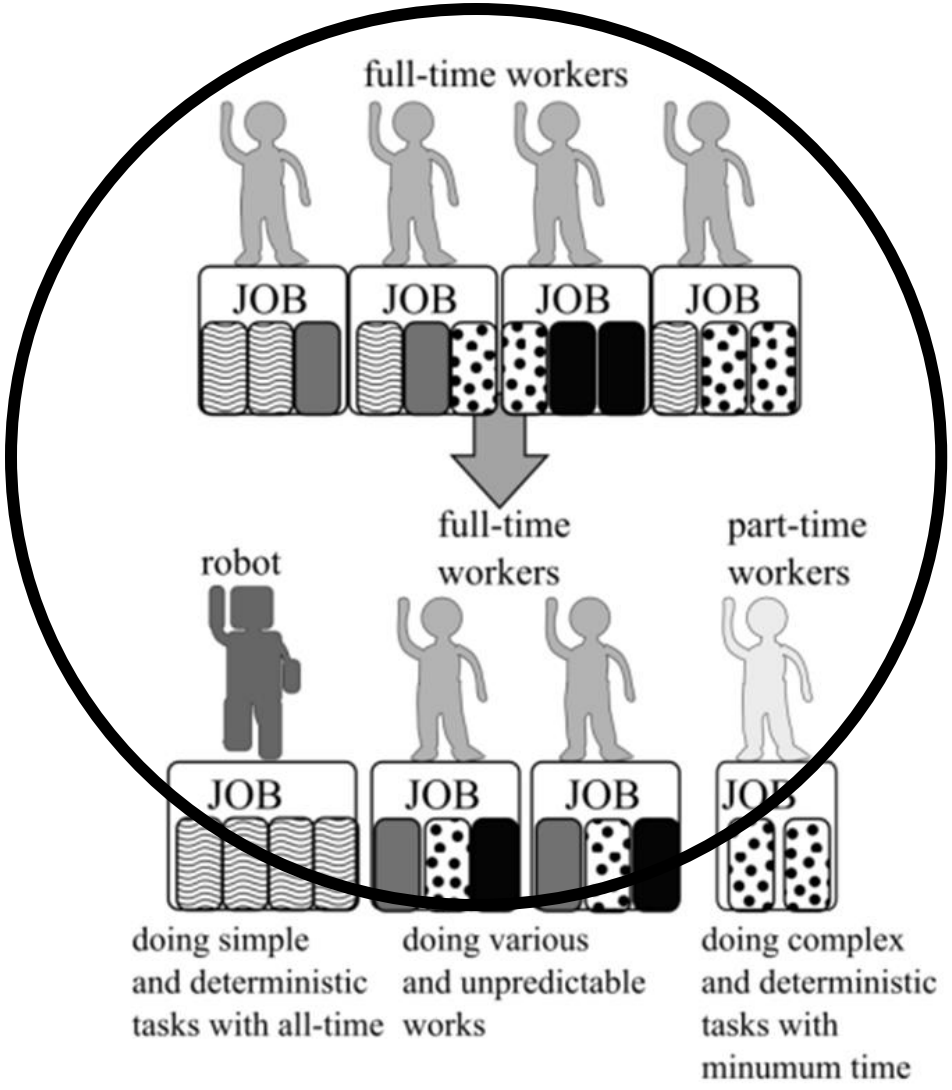
Jean-Marc Cote
“In the Year 2000”
～1900年パリ万博展示ポストカード～



三宅雪嶺主催『日本及び日本人』
百年後の日本特集（1920）

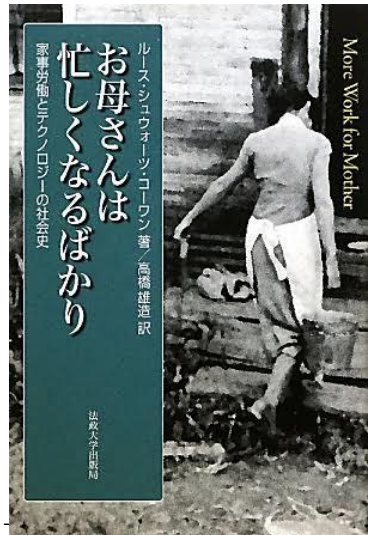
Robots/AI replace tasks, not jobs





科学技術と社会の相互作用

- 道具は適切に作動するように他の道具と結ばれてシステムを形成している。一部の道具が作動しなければ逆に労働は増すことがある
- 例えば19世紀の工業化と20世紀の家庭家電によって洗濯など個々のタスクは省力化された。しかしそのため洗濯の量や頻度が増え、召使や業者がやっていた仕事が主婦の仕事となり、洗濯に付随する他のタスク、例えば洗濯物を干す、たたむ等はいずれも自動化されていなかったため、結果的に「お母さんたち」の忙しさは変わらない、あるいは余計な仕事が増えた



ルース・シュウォーツ・コーワン「お母さんの仕事は忙しくなるばかり」

よく聞かれること

- 専門性が必要とされる職場における人手不足
- ベテランがいない、新人をすぐ現場にださないといけない
- （生成）AIをスーパーバイザーとして使えるようにしたい
- そのために今までのマニュアルやベテランの人の受け答えなどの事例集/データを学習させる

- そこにおけるベネフィットは何か？
- 逆にリスクや課題とは何か？

人と（生成）AIの関係性/リスクをどう整理するか

- どんな軸で考える？

	正しいか正しくないかの判断が容易にできる	正しいか正しくないかの判断が容易にできない
間違ったときのリスクが小さい	→△注意しながら使っていっていい領域 ルーティンワーク等（議事録の作成・要約等）を任せ、人間がチェックをしていくことが重要	→〇使っていっていい領域 chatGPTを使ったブレストなど
間違ったときのリスクが大きい	→△注意しながら使っていっていい領域？ 人間が責任を取ることができる枠組みか？	→×使っていくべきではない 人間でも判断ができず、リスクが大きい

  - 適切性（真理性）の基準があるかどうか、判断ができるかどうか
  - 適切性が重要かどうか、参考とする/そのまま利用するのか

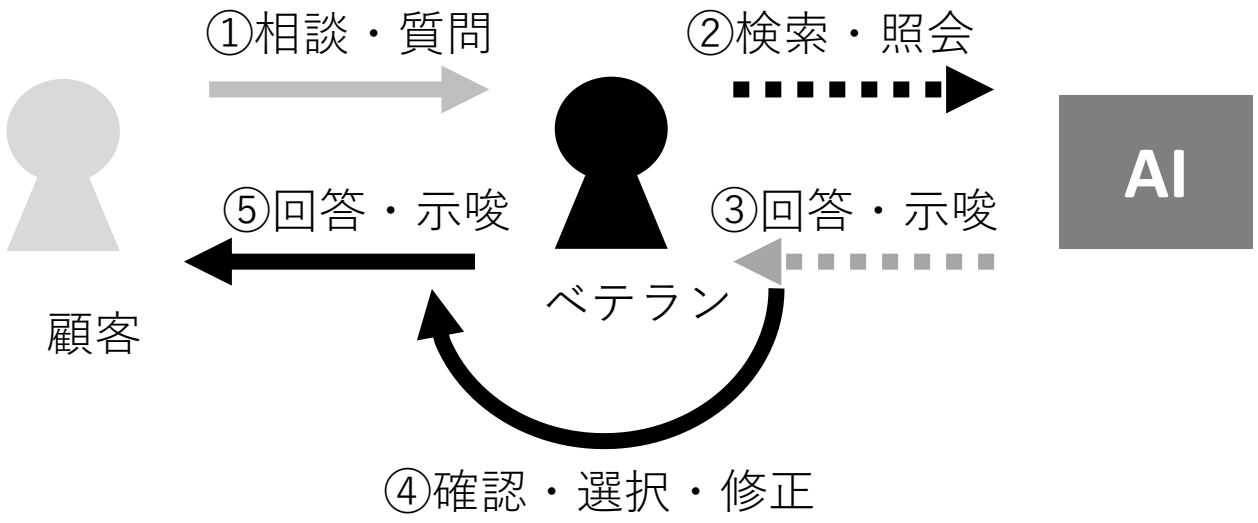
- AIの過小評価/過大評価と利用者のトレーニング

AIを専門家が使う

- 前提となる問い
  - 事例が十分あるのか、言語化されているのか、ルーティン（定型）かどうか、単純な作業か（データ）
  - 良い基準があるのか（モデル・アルゴリズム）
  - 真偽性を判断ができるものか否か、あるいはそれが重要となるか否か（評価）
  - 人の関与が必要なものではないか（場面・文脈）

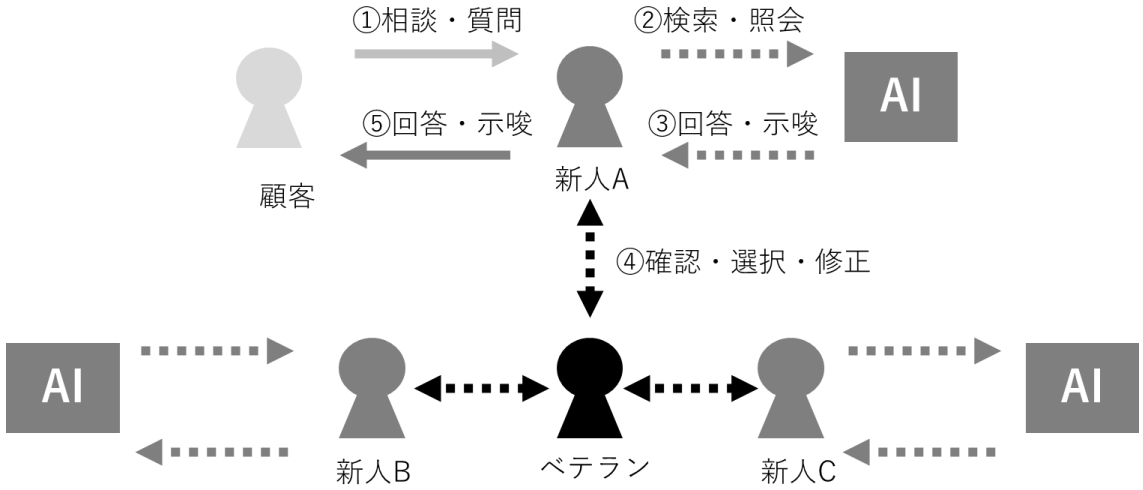
アシスタントAI

- その仕事内容に精通しておりAIの出力の間違いを適切に評価できる人や場合のみ使う
- あくまで最終判断は人、問題あったときの責任主体も人



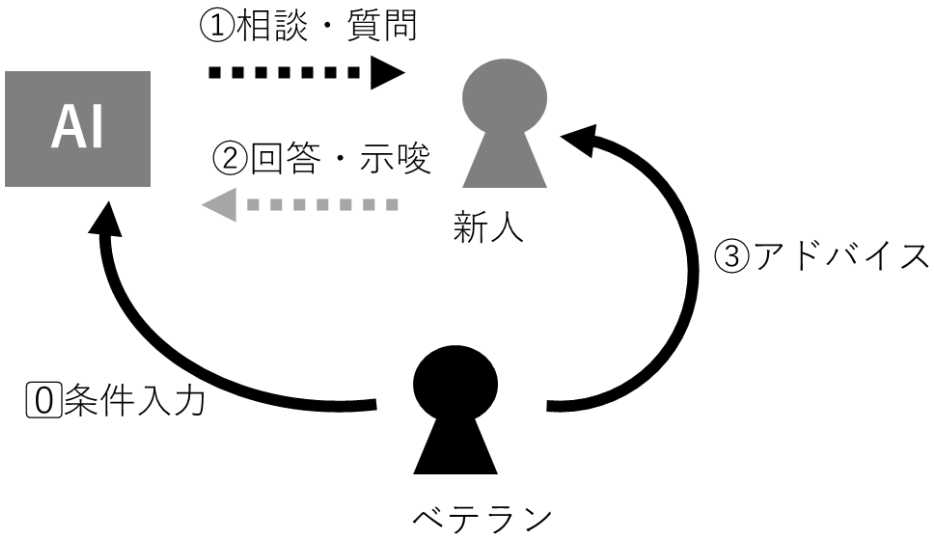
アドバイザーAI

- 経験は浅いがその仕事に精通できるようになるためのアドバイスをもらう
- 労働負荷軽減や人手不足解消といった観点から、このような使い方を求めるニーズはある
  - ただしAIの出力間違いを適切に評価できない可能性があるため、ベテランの監視がないと逆に問題が出てくることもある。
  - リスクや影響が少ない場面で使うことが推奨される。
  - 利用していることを周知することで異議申し立てができるようにする



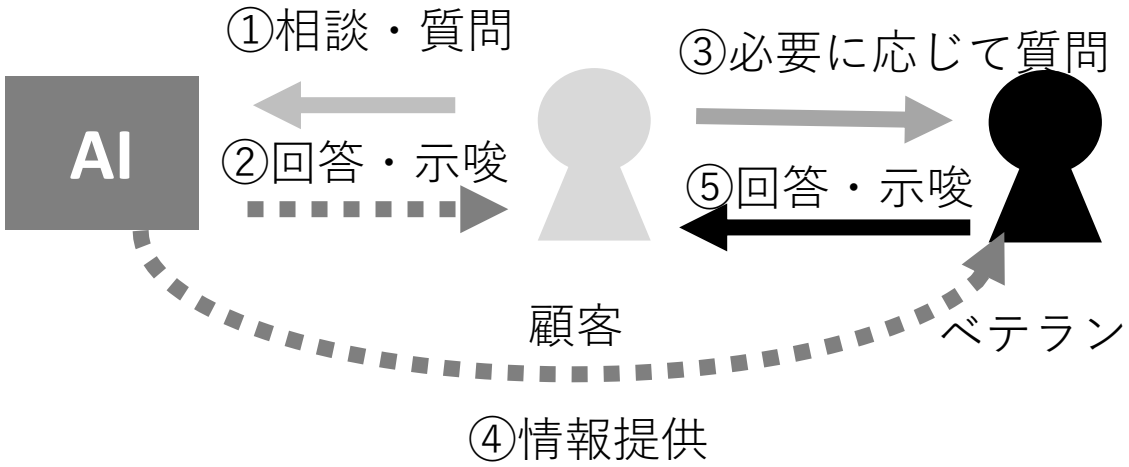
AIを訓練ツールとして使う

- トレーニングAI
  - 専門家としては経験の少ない人が、AIをクライアントや患者等に見立てて、カウンセリングやコンサルなどの練習に使う
  - ２４時間いつでも練習ができる
  - うまくいかなかった場合、人間のスーパーバイザーに相談をすることもできる
  - 間違ったとしても実質的な被害はない



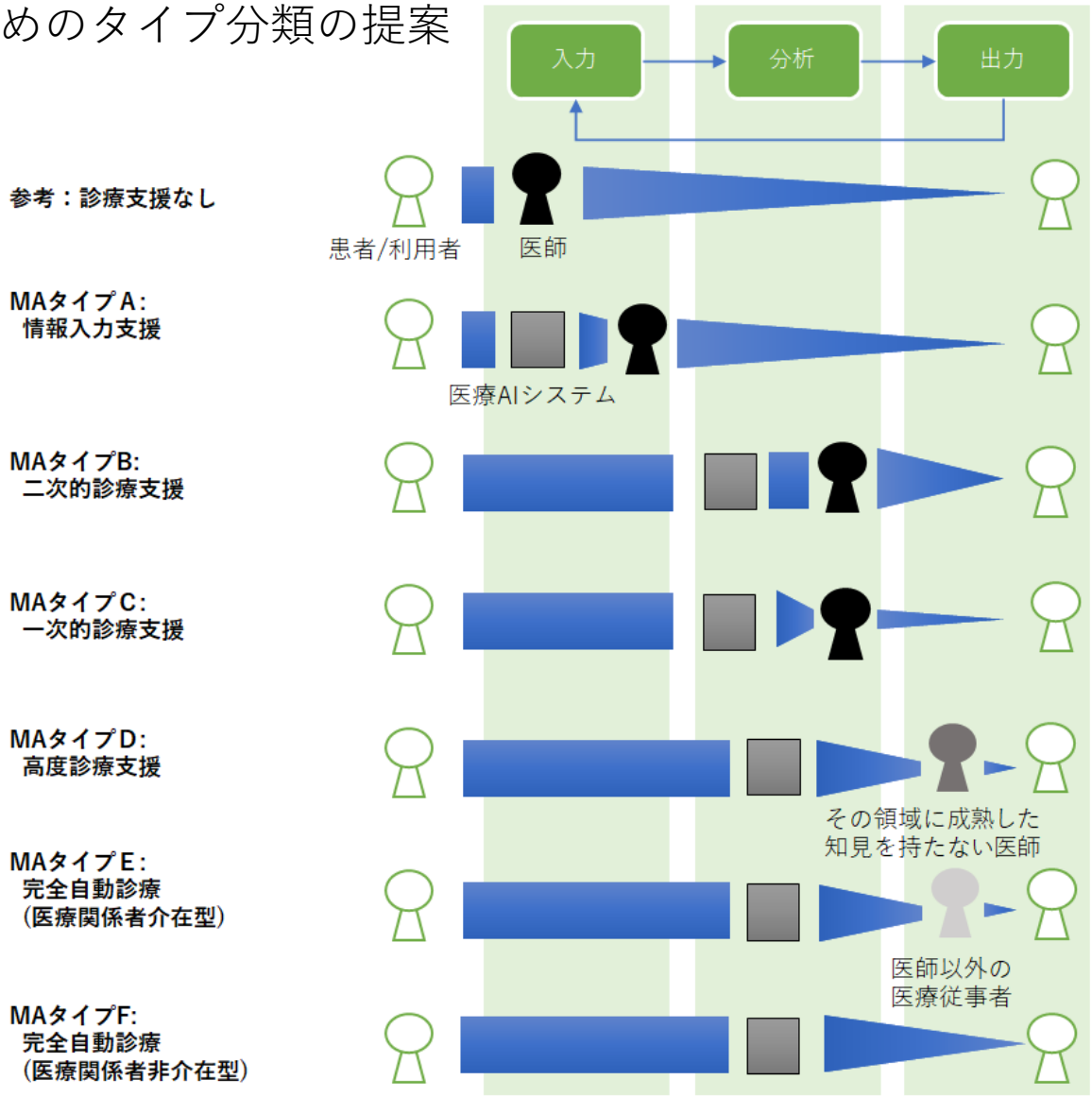
AIを専門家として使う

- インフォメーションAI
  - ある仕事に知識のない素人が専門家に相談する前に、概要や相場観を知るために使う
    - 専門家に相談する前の情報入手など
    - AIのほうが質問がしやすかったりする場合もある
    - リスクが低いものに関しては安価に相談が可能だが、問題があった場合の問い合わせ窓口などは必要
    - 一方で、嘘を出力するAIに惑わされた人たちによって専門家の仕事を増やしてしまう可能性もある
    - あくまで参考程度という表示がわかるデザインが必要



- プロフェッショナルAI
  - AIの精度が非常に高い、あるいはリスクが低い場合、あいが自動的に判断ができる。あるいは専門家の個別知識に限界がある場合や平準化が必要な場合に役立つ可能性もある
    - 責任の所在はどこにあるのか、真偽性の判断は誰がするのか
    - 社会としてリスクを受け入れられるか
  - 囲碁・将棋等の領域ではいかに使うことができるかが焦点になってもいる

医療 AI ソフトウェアシステムの開発と実装を推進するためのタイプ分類の提案



https://ifi.u-tokyo.ac.jp/news/5456/

https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2020/02/policy_recommendation_tg_20200213.pdf

リスクが高い領域とは？

- センシティブなデータを扱っている
- 人々の権利を侵害する可能性のある領域
- 人の意思決定に近い
  - アシスタントではなくアドバイザーあるいはプロフェッショナル的な使い方
- 影響を受ける人の規模・範囲が広く可能性が高い

コリングリッジのジレンマ（1980）

情報の問題
　技術が社会で使われる前にその影響力を予測することは難しい
力の問題
　一度普及してしまった技術は制御するのが難しい

- 「永遠のβ版」に対しては社会実験ではなく、「実験社会」に住んでいる自覚
  - アジャイル・ガバナンス
- 開発・利活用・問題発生時に影響の想定が求められる
  - 近年の、情報技術に対するELSI（倫理的・法的・社会的課題）への要請
  - ニーズからだけではなく、「どういう社会に住みたいのか」というビジョンも重要