AIと社会
― 倫理・リスク・包摂 ―

東京大学　未来ビジョン研究センター
日本ディープラーニング協会
理化学研究所　AIPセンター
江間有沙

科学技術社会論（STS）とは？

- 「怖い人？」ではありません
  - 科学技術研究・研究者の研究（メタ科学）
  - 科学と社会の間で起きている問題
    - 遺伝子組み換え食物,ナノテクノロジー,クローン,原子力

- STS二つの顔
- Science and Technology Studies (学問)
  - 科学技術の社会学、哲学、人類学、歴史学
- Science, Technology and Society（活動）
  - 科学技術政策、科学技術コミュニケーション

AIとは何か /AIの持つ技術的・社会的な課題

「人工知能」の範囲は？

1. 既存のICTの延長
   - **プログラム、ルール**に従う
   - 既存のガイドラインに即した項目、例えば**プライバシー**や**セキュリティ**の問題、技術利用やアクセスにおける格差（**デバイド**）が起きないことや、開発者による**説明責任**や**制御可能性**

2. 「学習」に焦点
   - データを用いて人間の知的活動（認知、思考、推論やそれに基づく意思決定）を部分的に**自律的に行える学習する技術システム**の可能性と課題を扱う
   - **アルゴリズムバイアス**がもたらす社会や雇用、人権等への影響に対する懸念に焦点
   - 一方で、学習の中身がブラックボックス化することにより**公平性**、**透明性**や**アカウンタビリティ**の担保が従来の技術同様に行うことの難しさが課題

3. 「まだ見ぬ技術」
   - **自律、自我、意識**を持つなどの人工知能の可能性を議論
   - 自律型致死兵器（LAWS）など、「まだ見ぬ技術」のリスクについて警鐘

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

参考）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

結果：今後もシステムを利用してよい（合憲）

- 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こと

- 様々な判断の場面に使われる技術
  - 機械を賢く使うためにも、機械の構造や限界を認識しておくこと

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**

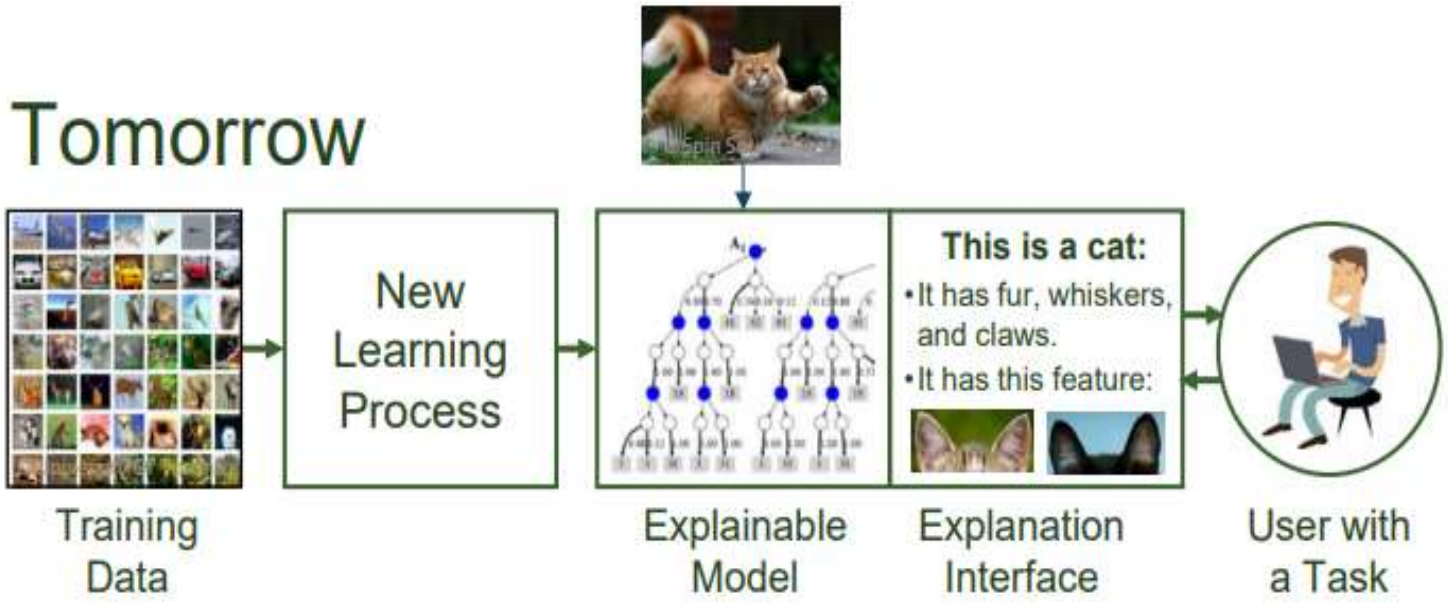
どういう基準でシステムを設計すれば公平？

- 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

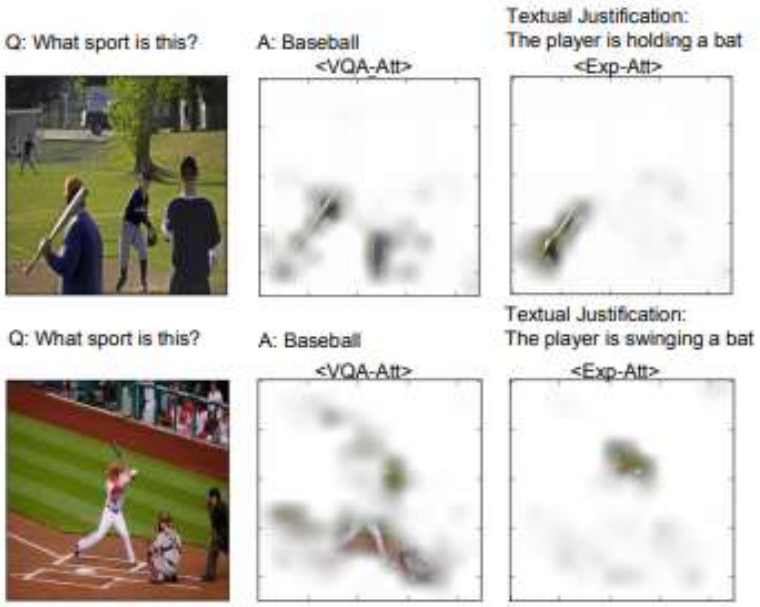
あなたが所属するコミュニティに「たりない」視点はありますか？

説明可能AI(XAI)と信頼

- 機械学習を予測や判別を行うときに、対象とする画像や音声などのデータのどの部分に重点を置いてみているのかの根拠を示してくれる
  - なぜそのような判断を下したのか、なぜほかの可能性が排除されたのか、どういう情報に重きを置いて判断をしているのか
  - どういう情報が足りていないと正答率が下がるだろうかといった仮説をたてて、再学習させることもできる



https://www.vertex-pts.com/single-post/darpa-xai-project



https://arxiv.org/pdf/1612.04757v1.pdf

(a) Husky classified as wolf (b) Explanation

https://arxiv.org/pdf/1602.04938.pdf

- 技術者がシステムを修正していく場合や、専門家が原因や理由を把握するには有効
- 一方、説明があるから「納得・安心」できるのではなく、何か問題があっても事後的に誰かが責任を取ってくれる、対策を取ってくれるから「納得・安心」できる場合もある

情報技術の特徴と様々な懸念

- 情報技術の特徴
  - 技術革新のスピードが速い
  - 気付かないうちに普及する
  - 「最適解」を計算：誰にとっての最適解？

  - 生命にかかわるリスク
    - 保険制度などの対応、責任の所在
  - 尊厳にかかわるリスク
    - プライバシー概念の解釈、「誘導」されること
  - 金銭にかかわるリスク
    - 法制度による対応
  - 自律性にかかわるリスク
    - 雇用、人間の役割から人類の存続など広範囲な考え

AIガバナンス
原則から実践へ

人工知能学会　倫理指針

Japan Society for AI *Ethical Guidelines*

1. 人類への貢献　Contribution to humanity
2. 法規制の順守　Abidance of laws and regulations
3. 他者のプライバシーの尊重　Respect for the privacy of others
4. 公平性　Fairness
5. 安全性　Security
6. 誠実な振る舞い　Act with integrity
7. 社会に対する責任　Accountability and social responsibility
8. 社会との対話と自己研鑽　Communication with society and self-development
9. 人工知能への倫理順守の要請　Abidance of ethics guidelines by AI
   - 人工知能が社会の構成員またはそれに準じるものとなるためには、上に定めた人工知能学会員と同等に倫理指針を遵守できなければならない。
     AI must abide by the policies described above in the same manner as the members of JSAI, in order to become a member or a quasi-member of society.

http://ai-elsi.org/archives/info/20170503

内閣府「人間中心のAI社会原則」

- 基本理念
  - (1) 人間の尊厳が尊重される社会
  - (2) 多様な背景を持つ人々が多様な幸せを追求できる社会
  - (3) 持続性のある社会

- 人間中心のAI社会原則
  1. 人間中心の原則
  2. 教育・リテラシーの原則
  3. プライバシー確保の原則
  4. セキュリティ確保の原則
  5. 公正競争確保の原則
  6. 公平性、説明責任及び透明性の原則
  7. イノベーションの原則

https://www8.cao.go.jp/cstp/aigensoku.pdf

国際企業におけるAI倫理原則

- **Global Company**
  - **Sony** Group AI Ethics Guidelines (2018.9)
  - **Fujitsu** Group AI Commitment (2019. 2)
  - **NEC** Group AI and Human Rights Principles (2019.4)
  - **NTT Data** AI Guidelines (2019.5)

- **Start-ups**
  - **ABEJA** launched 3rd party committee “Ethical approach to AI” (2019. 7)
    - “このたび、EAAを発足させたのは「AIと倫理」は**企業の経営戦略**に直結する問題だと認識しているからです。”

日本の企業が掲げるAI原則

発表年	企業名
2018	ソニーグループ
2019	NTTデータ、沖電気工業、クウジット、J.Score、日本電気、野村総合研究所、富士通、三菱総合研究所、リクルート
2020	日本ユニシス（BIPROGY）、日立製作所、富士フイルム
2021	KDDI、スタジアム、三菱電機
2022	コニカミノルタグループ、ABEJA、ソフトバンク、Zホールディングスグループ、パナソニック、東芝

AIネットワーク社会推進会議　報告書2022などを参考に作成
https://www.soumu.go.jp/main_content/000826564.pdf

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and Data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability

April 2019

BEIJING AI PRINCIPLES

人工智能北京共识

R& D Principles

- Do Good
- For Humanity
- Be Responsible
- Control Risks
- Be ethical
  - Making the system as fair as possible, reducing possible discrimination and biases …
- Be Diverse and Inclusive
- Open and Share

Use

- Use wisely and properly
- Informed-consent
- Education

Government

- Optimizing Employment
- Harmony and cooperation
- Adaption and moderation
- Subdivision and implementation
- Long-term planning

https://www.baai.ac.cn/blog/beijing-ai-principles

（表１）

	指針等の数	人間中心	人間の尊厳	多様性包摂	持続可能な社会	国際協力	適正な利用	教育リテラシー	人間の判断の介在制御可能性	適正な学習（学習データの質）	AI間の連携	安全性	セキュリティ	プライバシー	公平性	透明性説明可能性	アカウンタビリティ	堅牢性	責任	追跡可能性	モニタリング監査	ガバナンス	その他
米国	6	1	2			1	1		1	1		3	3	1	4	5	4	1	2	2	3		
カナダ	1							1								1							1
英国	6		2		1			1		2		1	1	2	5	5	5	1	2		2		1
イタリア	1									1				1		1							
オランダ	2		1				1		1	2				1	2	2	1		1	1			
スウェーデン	2		1	1			1	1					1	1	1	2							
デンマーク	2		2	2			1		1			1	1	1	2	2	1	1			1		
ドイツ	3	1	2	2	3				1	2		2	2	2	2	3	2	3	1	1	1		
ノルウェー	2		2		1		2		1			1		1	2	1	1	1		1	1		
フィンランド	2		2	1			1		1			1		1	1	2	1		2	2	1		
フランス	1								1						1	1	1				1		
インド	1		1	1								1	1	1	1	1	1		1				
韓国	1		1	1		1						1		1	1	1	1	1					
シンガポール	2		1												2	2	1						
中国	3		3	2	2	2	2	2	3	1		3	2	2	3	2	2	2		2	3	3	
オーストラリア	1	1	1									1	1	1	1	1	1		1		1		
EU	2		2	1	2	1	2		1			2	1	2	2	1	2	2	2	1	1		
国際機関	2		2	2	2	1	2	1	1		1	2	1	2	1	2	1		1	1	1		
合計	40	3	25	13	11	6	13	6	12	9	1	19	14	20	31	35	25	12	13	11	16	3	2

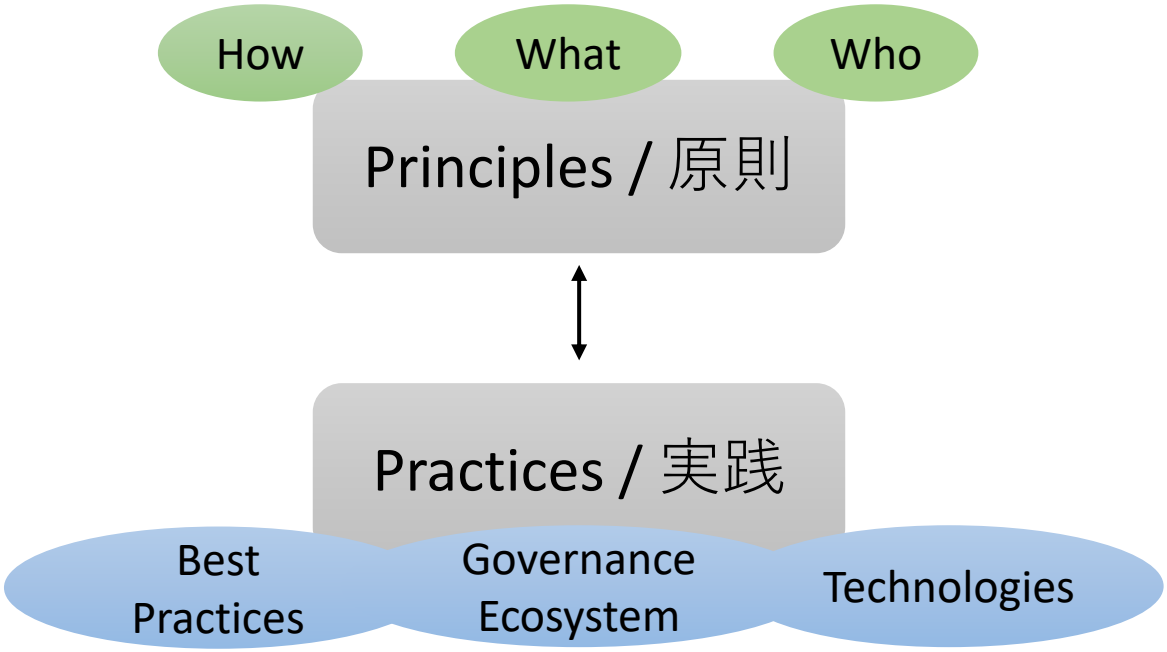
（注）記載があるものはピンク色に着色。また、各国等の指針等の数に対し、半数を超える指針等に記載のある場合、赤字で表示。

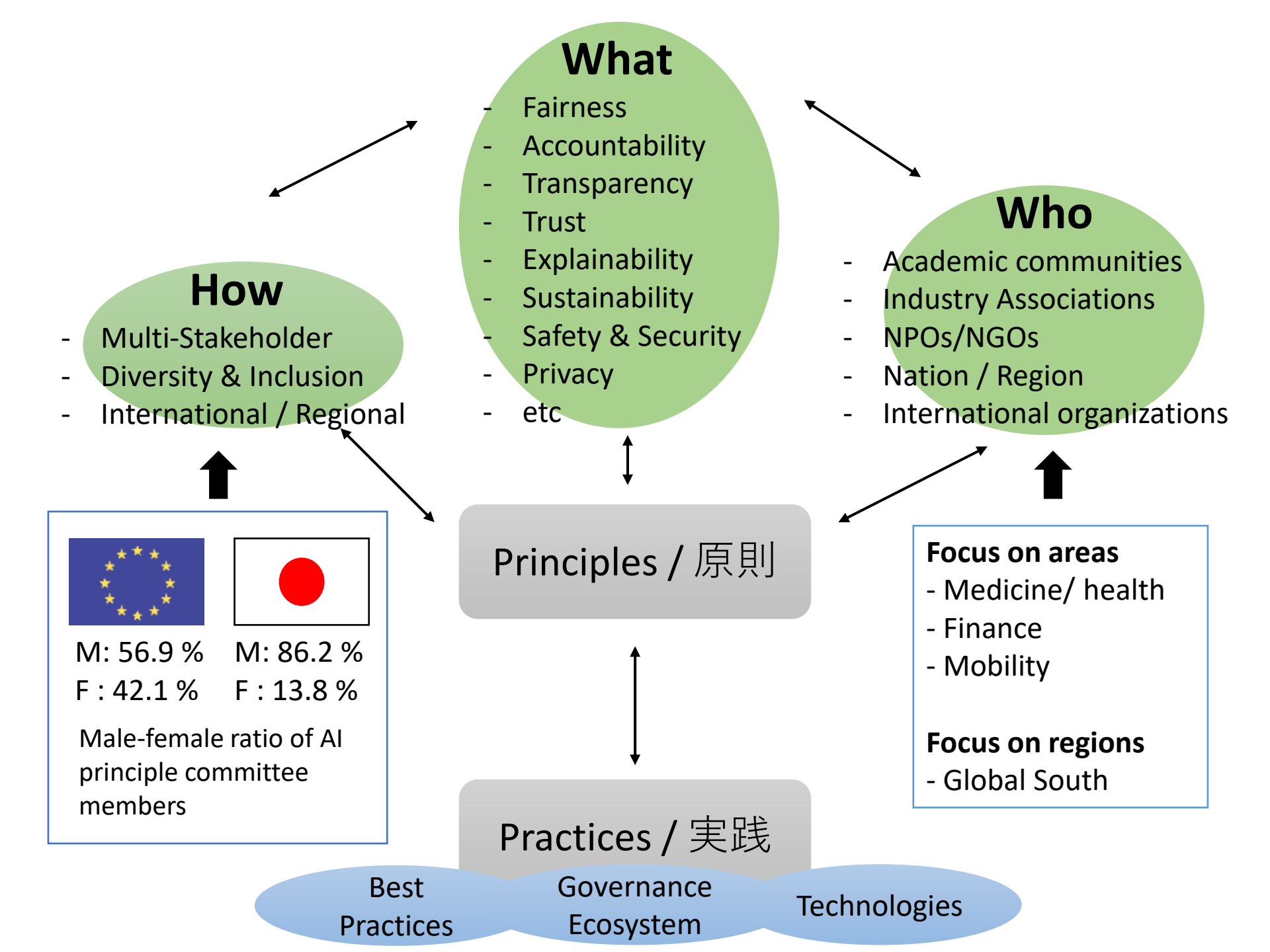
https://www.soumu.go.jp/main_content/000826564.pdf

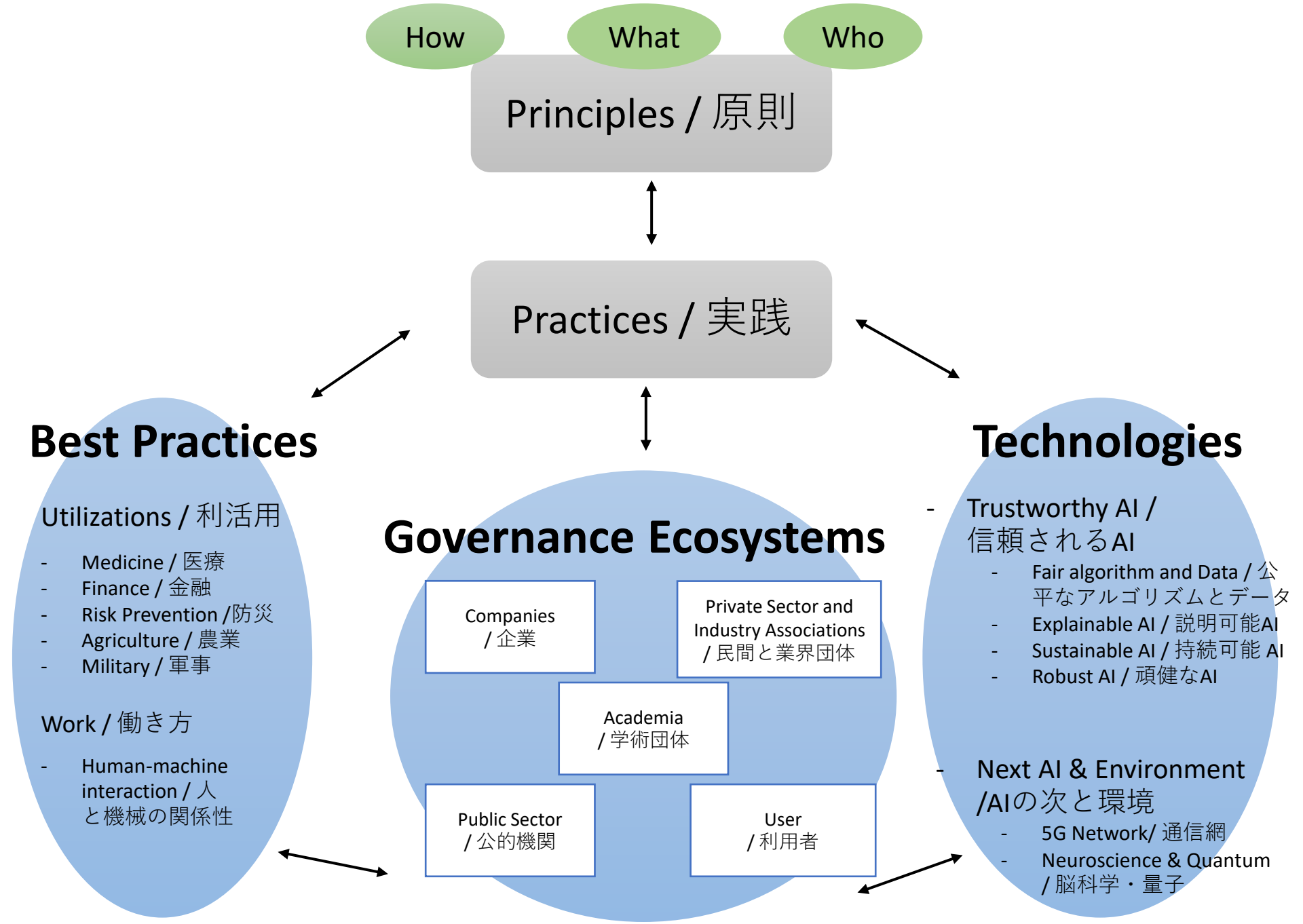
人工知能の原則で重要とされる価値の論点整理

米ハーバード大学	中国科学院
International Human Rights（国際的な人権）	Humanity（人間性/人道）
Promotion of Human Values（人間の価値の促進）	Collaboration（協調）
Professional Responsibility（専門家の責任）	Share（共有）
Human Control of Technology（技術の人間管理）	Fairness（公平性）
Fairness and Non-discrimination（公平性と非差別）	Transparency（透明性）
Transparency and Explainability（透明性と説明可能性）	Privacy（プライバシー）
Safety and Security（安全性とセキュリティ）	Security（セキュリティ）
Accountability（説明責任/答責性）	Safety（安全性）
Privacy（プライバシー）	Accountability（説明責任/答責性）
ー	AGI（汎用人工知能）

米ハーバード大学バークマンクラインセンターが作成した報告書（http://wilkins.law.harvard.edu/misc/PrincipledAI_FinalGraphic.jpg）と中国科学院の研究者らが作成した報告書（https://arxiv.org/abs/1812.04814）から作成

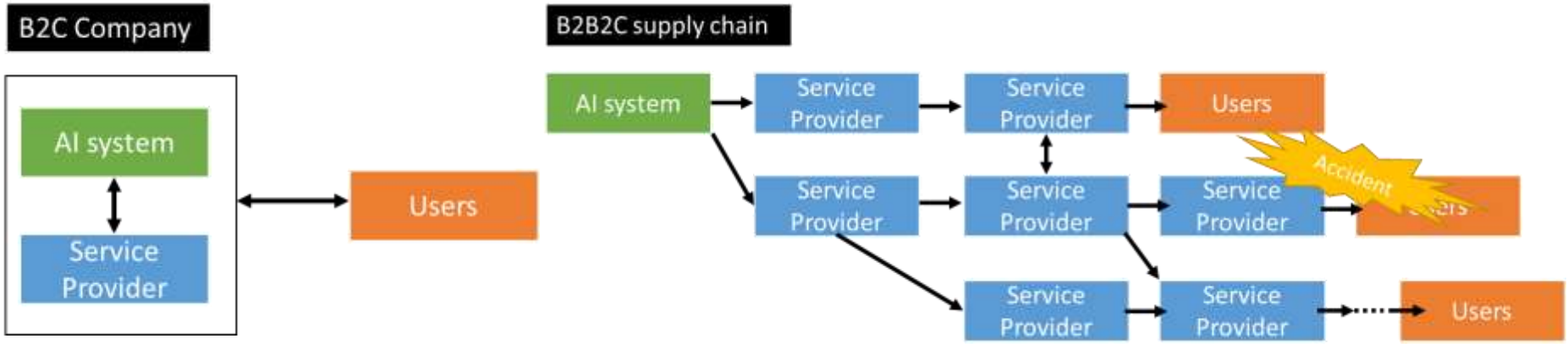

How
What
Who
Principles / 原則
Practices / 実践
Best Practices
Governance Ecosystem
Technologies


What
- Fairness
- Accountability
- Transparency
- Trust
- Explainability
- Sustainability
- Safety & Security
- Privacy
- etc
How
- Multi-Stakeholder
- Diversity & Inclusion
- International / Regional
Who
- Academic communities
- Industry Associations
- NPOs/NGOs
- Nation / Region
- International organizations
M: 56.9 %
F : 42.1 %
M: 86.2 %
F : 13.8 %
Male-female ratio of AI principle committee members
Principles / 原則
Focus on areas
- Medicine/ health
- Finance
- Mobility
Focus on regions
- Global South
Practices / 実践
Best Practices
Governance Ecosystem
Technologies


How
What
Who
Principles / 原則
Practices / 実践
Best Practices
Utilizations / 利活用
- Medicine / 医療
- Finance / 金融
- Risk Prevention /防災
- Agriculture / 農業
- Military / 軍事
Work / 働き方
- Human-machine interaction / 人と機械の関係性
Governance Ecosystems
Companies / 企業
Private Sector and Industry Associations / 民間と業界団体
Academia / 学術団体
Public Sector / 公的機関
User / 利用者
Technologies
- Trustworthy AI / 信頼されるAI
  - Fair algorithm and Data / 公平なアルゴリズムとデータ
  - Explainable AI / 説明可能AI
  - Sustainable AI / 持続可能 AI
  - Robust AI / 頑健なAI
- Next AI & Environment /AIの次と環境
  - 5G Network/ 通信網
  - Neuroscience & Quantum / 脳科学・量子

日本の産業構造の特徴

- 日本はB2Cではなく B2B企業が多く、AIサービス/システム開発において、データ取得・AIモデル開発・サービス提供が異なるアクターで行われる傾向があるため、事故時の対応などの議論が急務である



Ex. 1) Ecosystem of AI governance /
例1) 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

研究会：AIガバナンスとその評価

座長	江間有沙（東京大学 未来ビジョン研究センター 特任講師）
検討課題	多様なアクターによる管理・評価の体制の在り方を「ガバナンス」と定義し、どのようなガバナンスの形がありうるのか調査し、提言されるAIの構築の一助とする。

第４回　AIサービスに係る保険

開催日時	2020年10月26日（月）
内容	話題提供/ディスカッション
テーマ	AIを含むサービスに係る保険
話題提供者	高岡浩（損害保険ジャパン株式会社） 永野智也（東京海上日動火災保険株式会社）

第３回　AIシステムにおける監査・保証

開催日時	2020年9月25日（金）
内容	話題提供/ディスカッション
テーマ	AIシステムにおける監査・保証
話題提供者	阿子島隆（Japan Digital Design株式会社） 長谷友香（有限責任監査法人トーマツ）

配布資料はこちら（JDLA会員限定）

開催レポート

第２回　AI倫理ガイドライン

開催日時	2020年8月25日（火）
内容	話題提供/ディスカッション
テーマ	AI倫理ガイドライン
話題提供者	松本敬史（有限責任監査法人トーマツ/研究会副座長）

配布資料はこちら（JDLA会員限定）

開催レポート

第１回　AIガバナンスをめぐる国内外の動向

開催日時	2020年7月31日（月）
内容	本研究会について/話題提供/ディスカッション
テーマ	AIガバナンスをめぐる国内外の動向
話題提供者	江間有沙（東京大学 未来ビジョン研究センター/研究会座長）

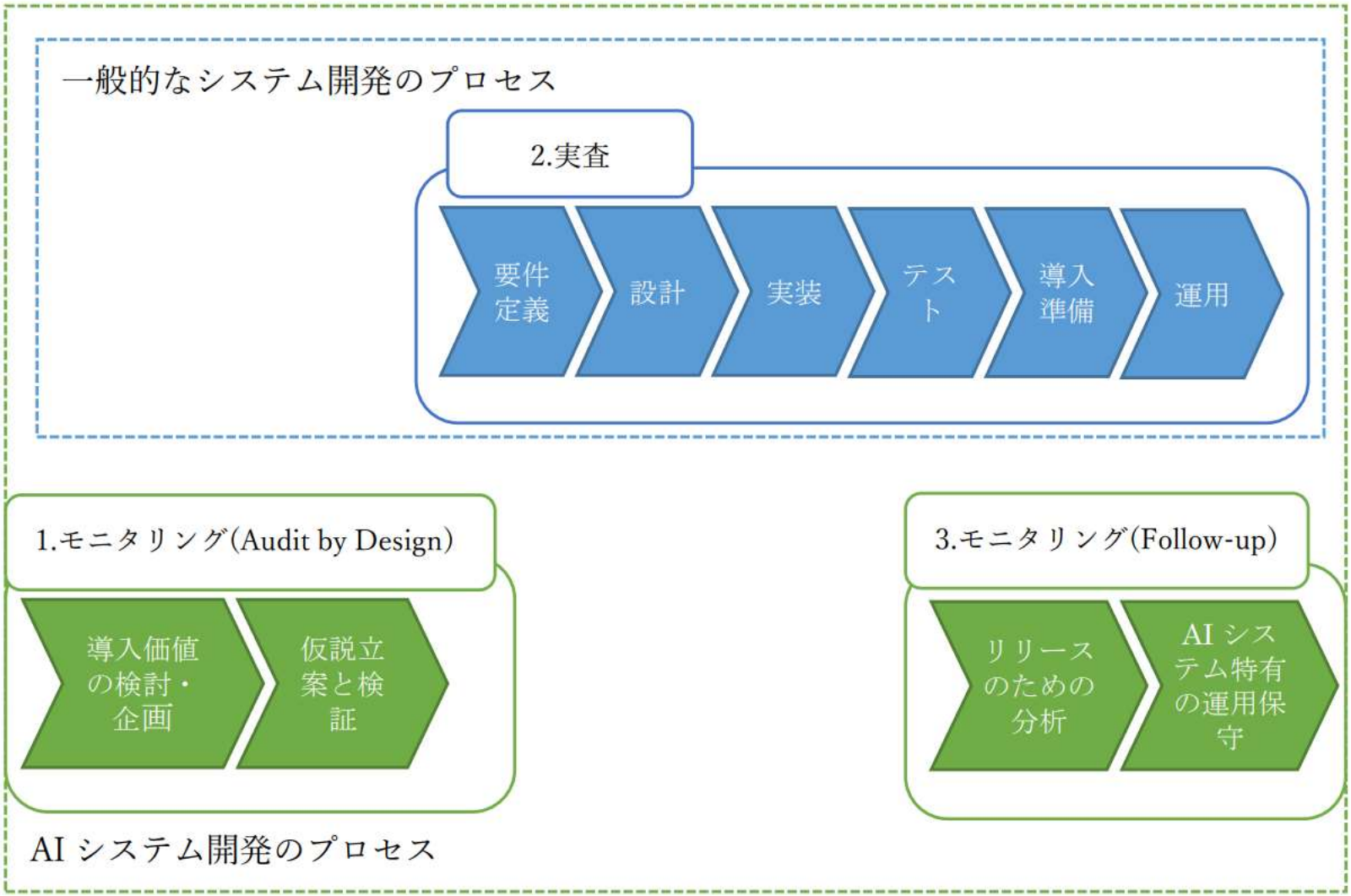
配布資料はこちら（JDLA会員限定）

開催レポート

https://www.jdla.org/about/studygroup/sg01/

AIと監査の会報告書より抜粋

図 1 AI 監査システムの流れ



Companies
/ 企業

Corporate vision and values /
ビジョンや実現したい価値

↔ AI Principles

https://ifi.u-tokyo.ac.jp/projects/ai-service-and-risk-coordination/

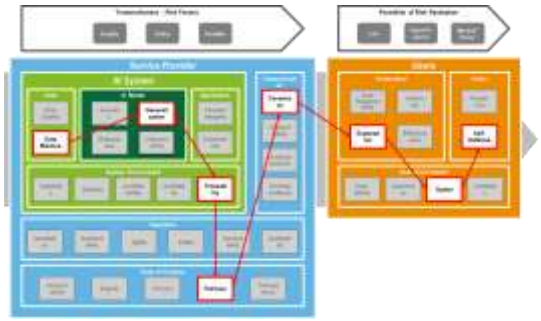
↔ - Facial Recognition
- Building Entrance

Create Risk Scenarios
/ リスクシナリオ作成

↔ - Data bias & wrong feedback
- Influence of the noise

Draw Risk Chains
/ リスク要因の関係性作成

↔



-Usage Policy
- Education
- Data modification

↔ Consider Risk Control
/ リスク管理実行

https://ifi.u-tokyo.ac.jp/projects/ai-service-and-risk-coordination/

重要なリスクシナリオに対するコントロールのサマリー

- 各リスクチェーンの検討結果を集約 -

Step5

実現すべき価値・目的		リスクNo.	リスクシナリオ	不確実性	環境変化	利用者起因	RC	コントロールのサマリー		
								AIシステム	サービスプロバイダ	ユーザー
1	人材採用レベルの維持・向上	R001	適切な評価	○			●	個別モデルの開発環境 個別モデルの学習データ 個別モデルの開発	個別モデルの目標精度 個別モデルの開発体制 個別モデルの性能監視	正確なフィードバック
		R002	予測性能の維持	○			●	予測性能の確保 利用ログの記録	予測性能の検証 代替運用の取り決め	代替運用の実施
		R003	ノイズによる影響	○			●	敵対的事例の学習 モデルの頑健性 判断根拠の出力	判断根拠の分かりやすさ 開発チームへの連携 判断根拠の検証	判断根拠の検討
		R004	虚偽の申込	○			●	予測性能の確保 判断根拠の出力	過去の不正事例との検証 利用部門への注意喚起	人間による最終判断
		R005	過度なAI依存	○		○	●	予測精度の確保 判断根拠の出力	AIモデル性能の開示 判断根拠の分かりやすさ	予測性能・リスクの認識 人間による最終判断
		R006	誤ったフィードバック	○		○	●	学習データの検証 学習時の情報の保管	判断精度や異常値の検証 教師ラベルのデータ修正	正確なフィードバック 人事システムからの反映
		R007	人材トレンドの変化	○	○		●	データ分布変化の認識 汎化性能の見直し	データ分布／正解率の検証 再学習	人材トレンド変化の認識
		R008	新たな職種	○	○		●	個別モデルの開発環境 個別モデルの学習データ 個別モデルの開発	個別モデルの開発体制 個別モデルの検証	新たな職種要件の検討
2	採用活動に係るコストの削減	R009	コスト超過						コスト管理	
3	海外グループを含めたサービス提供	R010	地域の会社への対応	○	○		●	個別モデルの開発環境 個別モデルの学習データ 個別モデルの開発	個別モデルの目標精度 個別モデルの開発体制 個別モデルの性能監視	
		R011	不十分な開発スピード					再学習環境の準備 モデル開発者の確保	PJ体制の確保	
4	企業の社会的責任(公平性のある採用活動)	R012	判断根拠情報の不正販売	○		○	●	判断根拠の出力 利用ログの記録	管理責任の合意形成 不正利用の監視 外部弁護士との連携	不正利用の管理責任 ペナルティの理解 担当のローテーション
		R013	公平性	○		○	●	データの偏り モデルの汎化性	公平性ポリシーの検討 ネガティブな判断の開示	AIの判断傾向の理解 公平な最終判断
		R014	予測結果の目的外利用			○		データ保護	アクセス管理 目的内利用	データの取扱
		R015	風評被害			○		データ保護	職業倫理の教育	職業倫理の教育
		R016	プライバシー保護			○		データ保護	法令順守の教育	法令順守の教育 データの取扱

https://ifi.u-tokyo.ac.jp/projects/ai-service-and-risk-coordination/

責任ある研究・イノベーション
Responsible Research and Innovation (RRI)

応答性と変化への適応性
変化する状況、知識や展望に対応して思考や行動様式、組織構造を包括的に変更できること

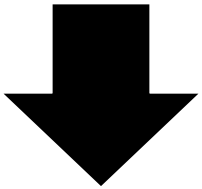
先見性と省察性
研究やイノベーションがどのように未来を形成するかをよりよく理解するための前提、価値観や目的を熟考し、影響を想定すること

多様性と包摂性
研究・イノベーションの実践、普及と意思決定において、科学技術の発展の早い段階から多様な人を巻き込むこと

公開性と透明性
人々が情報を精査し対話できるように、方法、結果や影響についてバランスよく伝達すること

https://www.rri-tools.eu/

- コリングリッジのジレンマ
  - 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
  - 一度普及してしまった技術は制御するのが難しい



- 「社会実験」ではなく「実験社会」
- 私たち一人一人の役割と責任とは？
- どのような生き方・働き方を目指す？