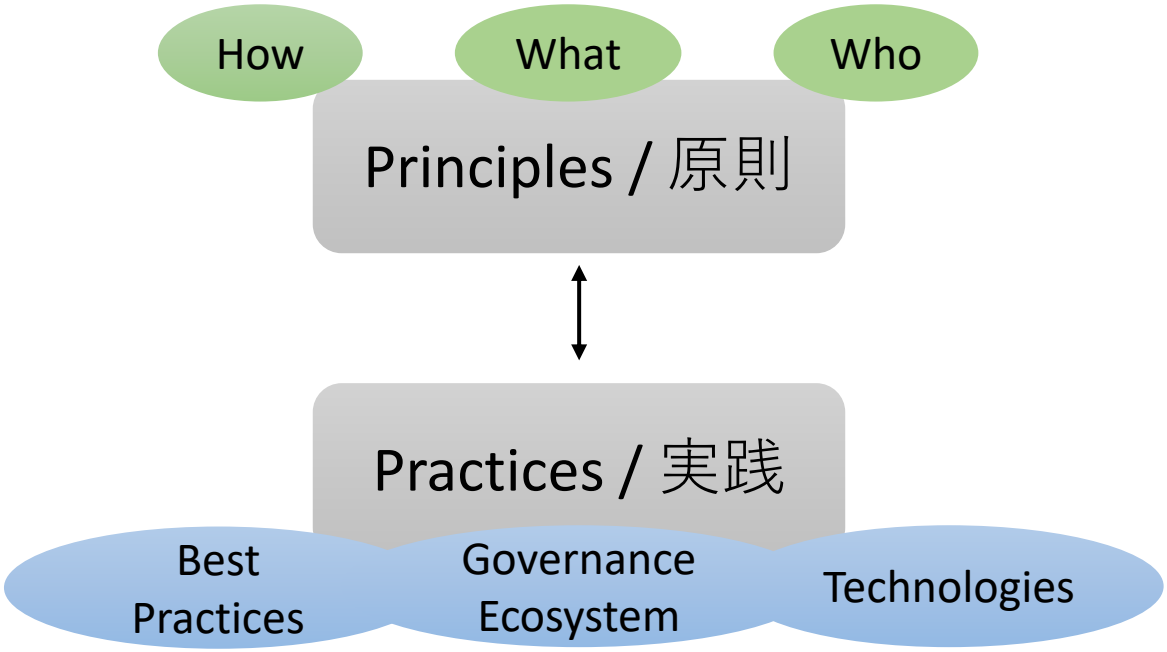
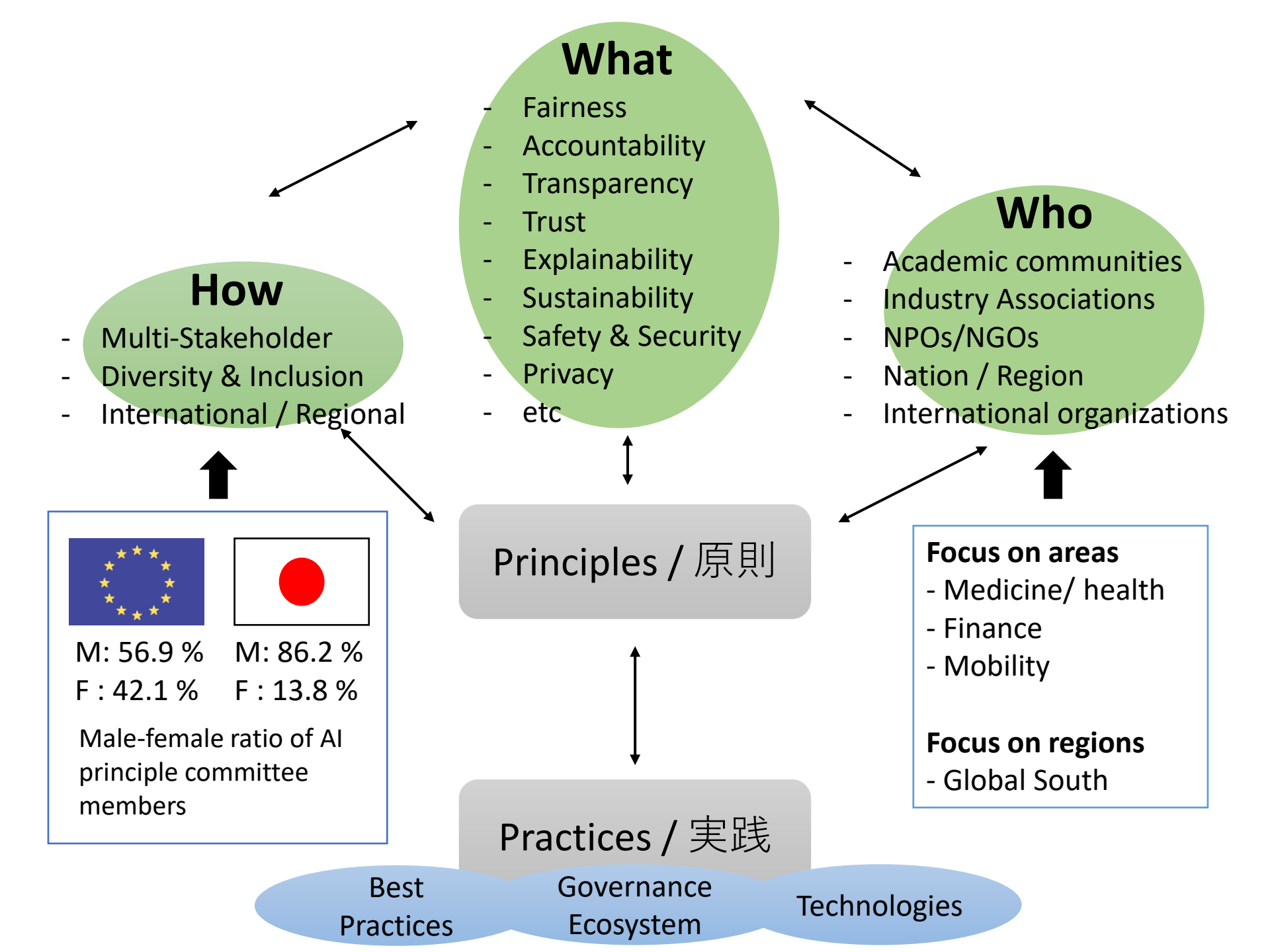
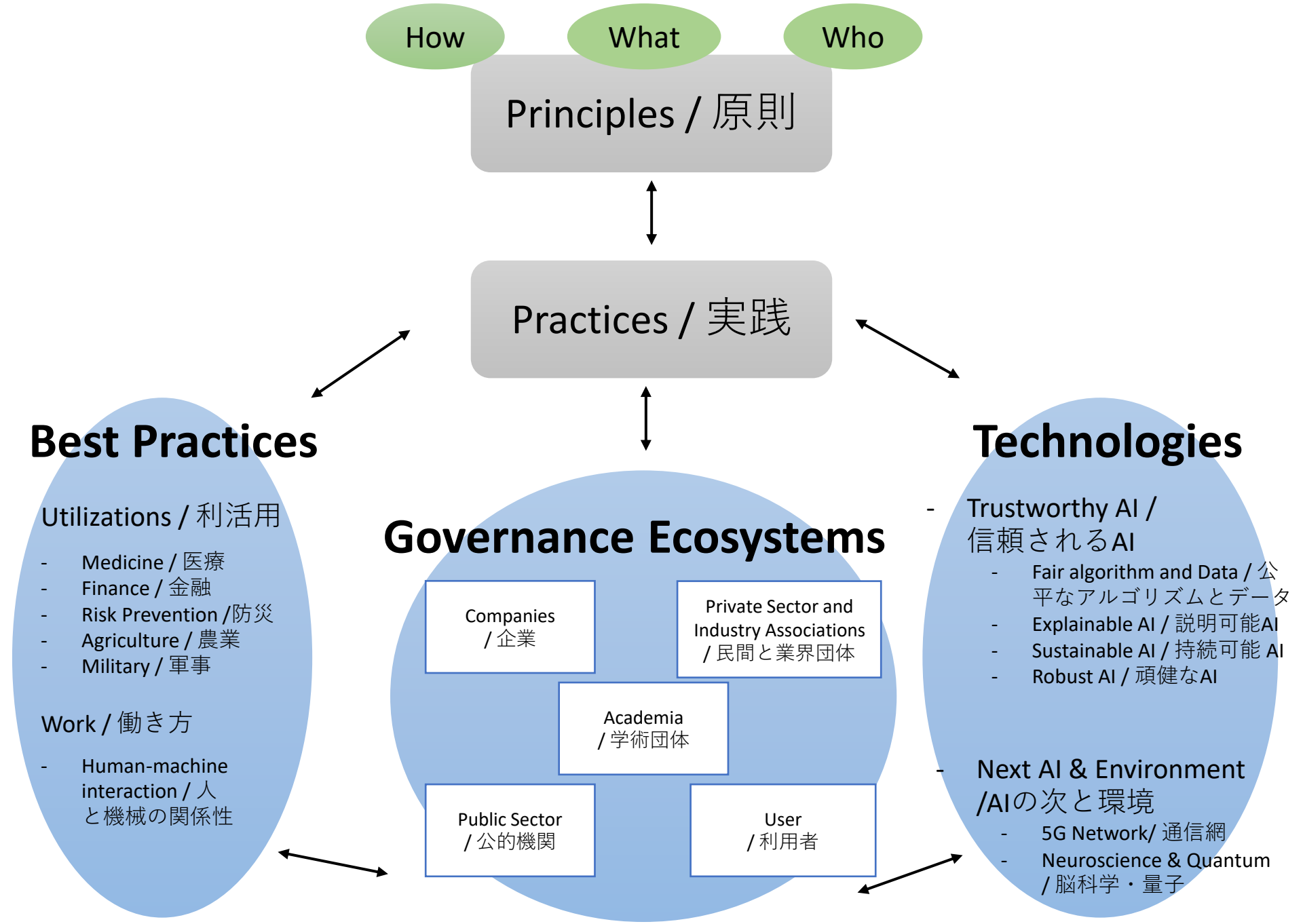
AIガバナンス
原則から実践へ

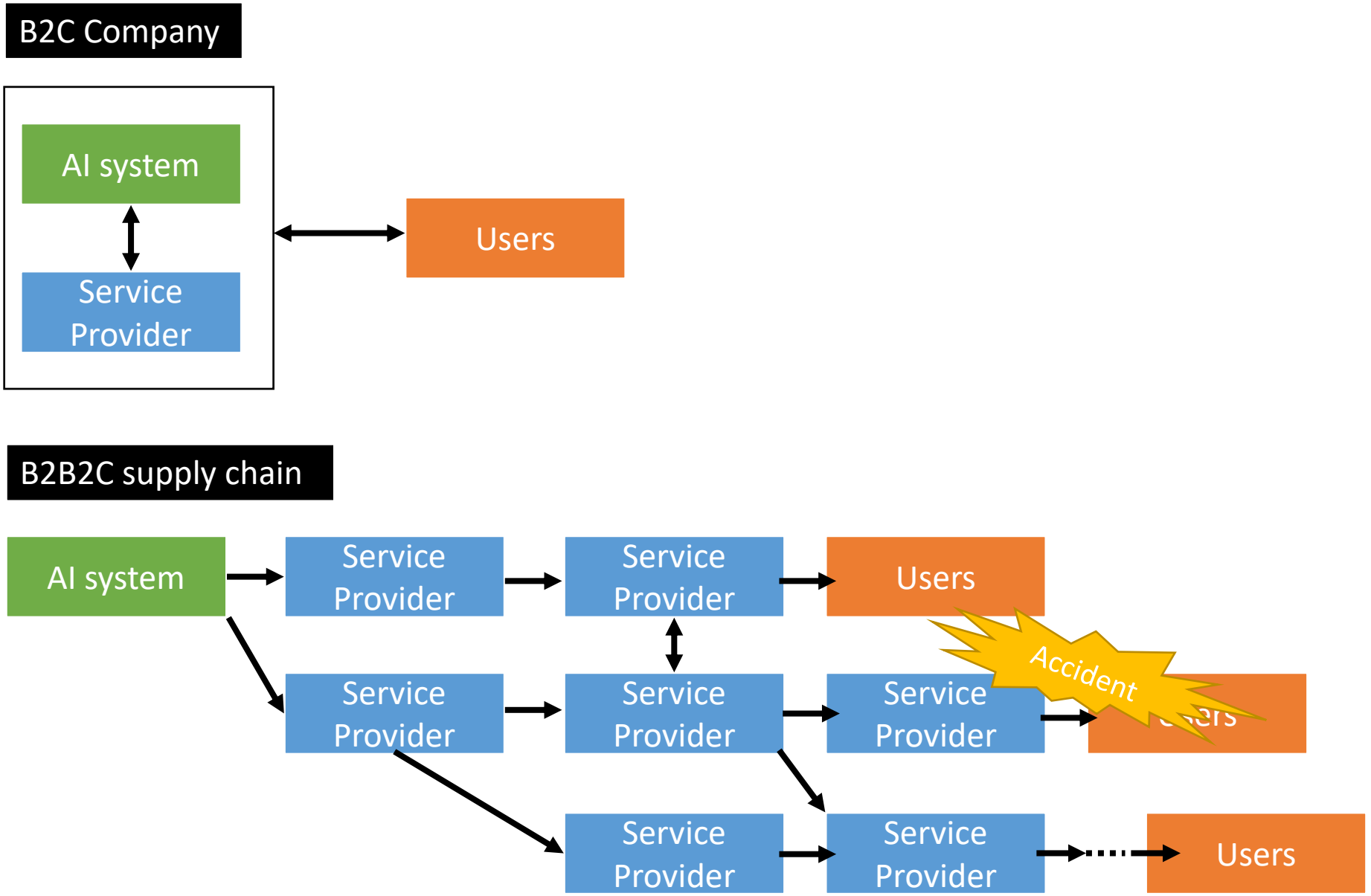
東京大学未来ビジョン研究センター准教授
理化学研究所AIPセンター客員研究員

江間　有沙


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


B2C Company
AI system
Service Provider
Users
B2B2C supply chain
AI system
Service Provider
Service Provider
Users
Service Provider
Service Provider
Service Provider
Accident
Service Provider
Service Provider
Users

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

Companies
/ 企業

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

↔ AI Principles

Consider AI Service Requirements &
Technologies / AIサービス要件と技術把握

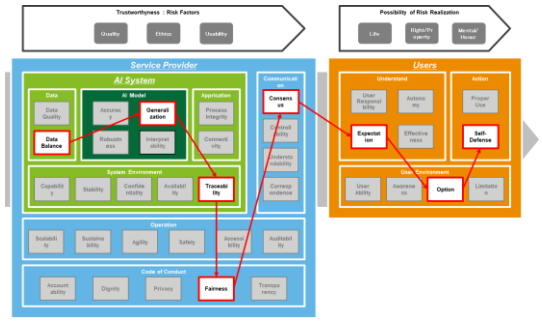
↔
- Facial Recognition
- Building Entrance

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

↔

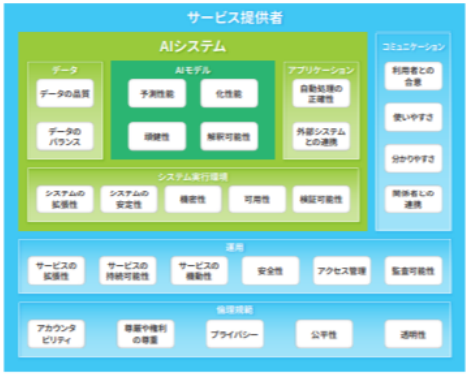
Consider Risk Control
/ リスク管理実行

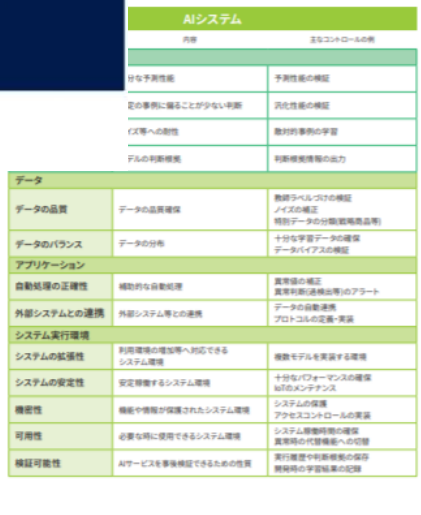
リスクアセスメント&コントロール (Case2：採用AI)
- リスクの複雑性評価 → リスクコントロールを検討 -

実現すべき価値・目的（リスクの影響先）		リスクNo.	リスクシナリオ	技術的難易度	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	○			●	システム環境の確保 個別モデルの開発	目標精度の設定・見直し 再学習の指示	正確なフィードバック
		R002	予測性能の維持	○			●	十分な正解率の確保 検証可能性の確保	予測性能の検証 再学習	代替運用
		R003	ノイズによる影響	○			●	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 判断根拠の検証	判断根拠の検討
		R004	虚偽の申込	○			●	十分な正解率の確保 判断根拠の出力	過去の異常事例の検証 利用部門との連携	最終選考プロセス
		R005	過度なAI依存	○		○	●	判断根拠の出力	判断精度の開示 判断根拠の分かりやすさ	予測性能・リスクの認識 最終判断プロセス
		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・異常値の検証 再学習	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	分布／正解率の監視 再学習	人材トレンド変化の認識
		R008	新たな職種	○	○		●	開発環境の準備 学習データ確保 モデルの開発	開発体制の整備 開発したモデルの検証	要求仕様の定義
2	採用活動に係るコストの削減	R009	コスト超過					適切な価格設定	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	○	○		●	システム環境の確保 個別モデルの開発	個別の目標精度の設定 モデルの性能監視 開発体制の確保	
		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	○		○	●	操作ログの記録	不正利用の監視 外部弁護士との連携	内部牽制
		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

ケース事例も公開中！

Case01.採用AI (2021/07)
Case02.無人コンビニ(2021/07)
Case03.送電線の外観検査ドローン (2021/07)
Case04.不良品検知AI (2021/07)
Case05.道案内ロボット(2021/07)
Case06.再犯可能性の検証AI (2021/07)





AIシステム

構成要素	内容	主なコントロールの例
予測性能	予測性能の検証	
汎性能	汎性性能の検証	
頑健性	敵対的事例の学習	
解釈可能性	判断根拠情報の出力	
データ		
データの品質	データの品質確保	教師ラベルづけの検証 ノイズの補正 特別データの分類(範囲商品等)
データのバランス	データの分布	十分な学習データの確保 データバイアスの検証
アプリケーション		
自動処理の正確性	補助的な自動処理	異常値の補正 異常判断(通検出等)のアラート
外部システムとの連携	外部システム等との連携	データの自動連携 プロトコルの定義・実装
システム実行環境		
システムの拡張性	利用環境の増加等へ対応できるシステム環境	機能モデルを実装する環境
システムの安定性	安定稼働するシステム環境	十分なパフォーマンスの確保 IoTのメンテナンス
機密性	機能や情報が保護されたシステム環境	システムの保護 アクセスコントロールの実装
可用性	必要な時に使用できるシステム環境	システム稼働時間の確保 異常時の代替機能への切替
検証可能性	AIサービスを事後検証できるための性質	実行履歴や判断根拠の保存 開発時の学習結果の記録

サービス提供者

構成要素	内容	主なコントロールの例
倫理規範		
アカウンタビリティ	説明責任、利用者の保護	責任関係の明確化 適切な期待値の検討
尊厳や権利の尊重	利用者側の優先権の検討	人間の介在度合の定義
公平性	サービス全体での公平性	公平性を確保する項目の定義
プライバシー	個人情報の適切な管理	プライバシーポリシーの遵守
透明性	AIサービスに係る必要な情報の開示	ステークホルダーごとに開示すべき情報の整理
運用		
サービスの拡張性	利用環境の増加等へ対応できるサービス体制	サービス提供に十分な体制 必要な支援へのサポート
サービスの持続可能性	持続的なサービスの維持	サービス要件の見直し 再学習・継続学習
サービスの機動性	適時の対応	アジャイル開発の実施
安全性	サービス全体での安全性の確保	システム外の安全対策 利用者の不正利用への対処
アクセス管理	適切なシステムの利用権限・範囲の管理	適切なアクセス権限の設定 不正な利用者の制限
監査可能性	第三者を含む必要な監視	AIシステムのモニタリング トラブル発生時の検証 不正利用の監視
コミュニケーション		
利用者との合意	利用者との認識合わせ	利用者責任の合意 AIシステムの変更への合意
使いやすさ	AIサービスの使いやすさ	使いやすいUI
分かりやすさ	AIサービスが出力する情報の分かりやすさ	判断根拠の分かりやすい表現
関係者との連携	利用者に対するリモートでのサポート	利用者のリモート支援 専門家との連携

ユーザ

構成要素	内容	主なコントロールの例
理解		
利用者の責任	利用者責任の理解	利用者責任の合意(最終判断等) AIシステムの変更への合意
利用者の自律性	自律的に判断すべき内容の理解	AIの判断の訂正
サービスの期待値	AIサービスに対する期待精度の理解	期待値(予測精度等)の理解 ネガティブな判断傾向の理解
影響の把握	利用者への影響の把握	AIサービスに係るリスクの理解 ワークスタイルの変化の理解
利用環境		
利用者の能力	サービス利用において必要な知識・スキルの習得	必要なリテラシーの確保 利用者の教育
AIの存在の認識	AIの存在の認識	AIの存在の明示 利用者への通知
制御可能性	利用者側での制御	マニュアル手順の整備 AI判断の訂正機能
利用者への制限	不正利用等の防止に向けた利用者の制限	利用者への技術的な制限 法的な制約
行動		
適正利用	正しいサービス利用	適切な目的に沿うサービス利用 教師ラベルの正しい設定
自己防御	利用者自身の保護	不利益が発生した際のクレーム 環境変化に係る事前連携

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

責任ある研究・イノベーションとは？

- 2000年代前半から欧米で議論され始める
- 欧州連合のHorizon 2020（2014-2020）でも言及

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

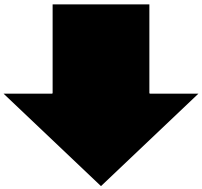
先見性と省察性
研究やイノベーションがどのように未来を形成するかをよりよく理解するための前提、価値観や目的を熟考し、影響を想定すること

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（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



- 設計の段階から影響の想定が求められる
- 「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？