第12回 自動車機能安全カンファレンス2024

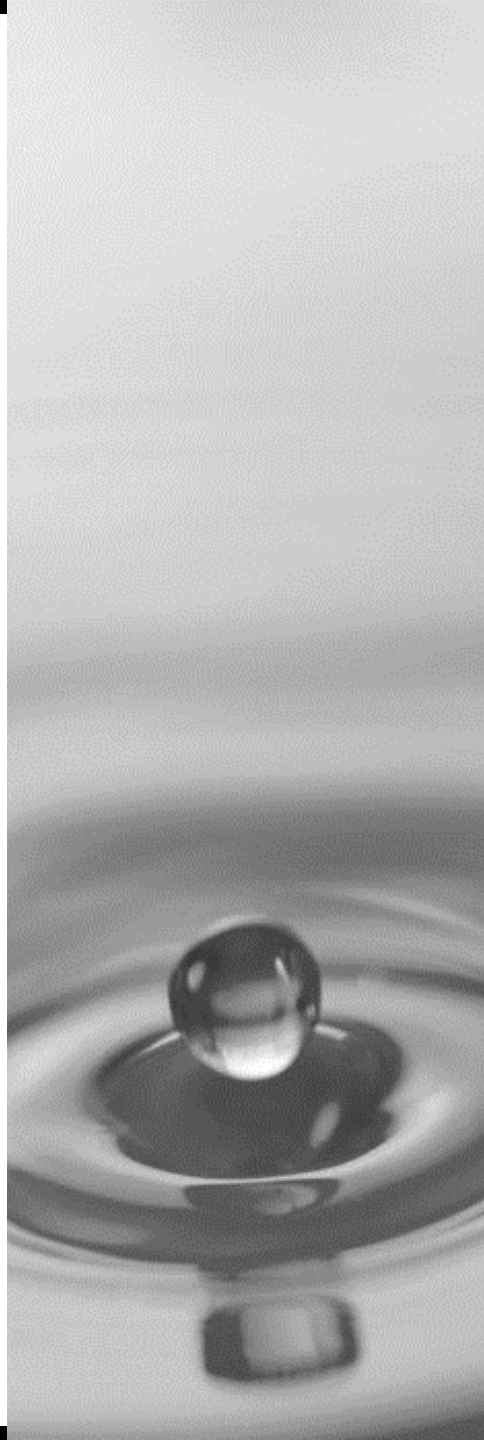
AI法規対応に向けた AIシステムの品質や安全を保証する技術

※AI：Artificial Intelligence（人工知能）

2024年12月5日
株式会社イマジナリー　執行役員
株式会社ヴィッツ　サービス開発部　執行役員
HMCES Project プロジェクトリーダー
森川 聡久

Human-Machine Co-Evolution Systems

IMAGINARY

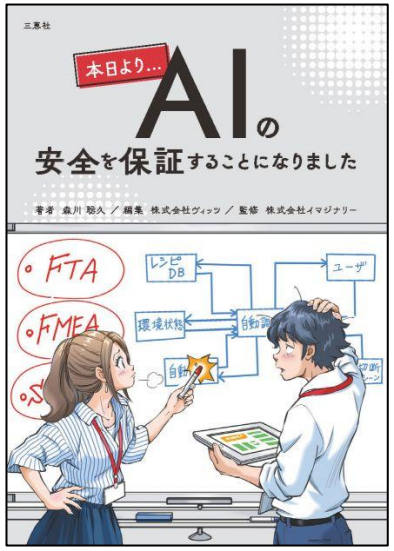


自己紹介

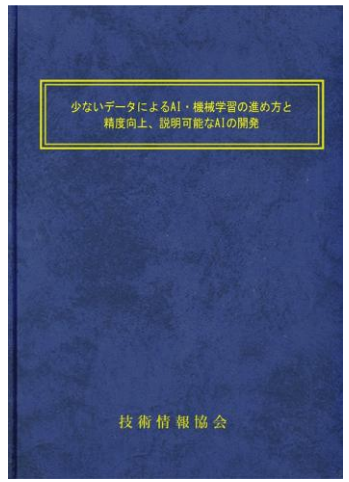
安全性・信頼性の高い組込みシステム開発が得意分野

- 略歴
  - 情報家電の組込みソフト新規開発、車載系ソフトPF開発などを主に経験
  - 2006年より機能安全開発に着手。2010, 2012年に機能安全プロセス認証取得を成功に導く
  - 機能安全/製品安全/AI安全を中心に事業を牽引（国内100社以上の支援実績）
  - AIの機能安全規格適合方法を整理し、国際的に技術提案（テクニカルペーパー公開）
  - ISO/IEC JTC1/SC42 WG3にて、AIの機能安全規格（ISO/IEC TR 5469）策定に貢献
  - 2022年11月：AIシステムの安全保証対策のポイントを、ストーリー仕立てで解説した本を出版
  - **現在、AIシステムの品質&安全論証支援、AI法規対応支援、人とAIの共進化社会に備えた研究開発に従事**

- 外部団体活動：
  - 2007年～2013年：組込みシステム技術に関するサマーワークショップ(SWEST)実行委員（内2008年～2012年 運営委員長）
  - 2011年～現在：システム開発文書品質研究会(ASDoQ)運営委員
  - 2013年～現在：MISRA-C研究会メンバ
  - 2016年～現在：IoT住宅普及に向けた住宅設備機器連携の機能安全に関する国際標準化および普及基盤構築 規格作成WG委員会 オブザーバ
  - 2017年～現在：組込みシステム開発技術研究会(CEST)幹事
  - 2019年～現在：AIプロダクト品質保証コンソーシアム(QA4AI)メンバ
  - 2019年～現在：AI国際標準化 ISO/IEC JTC1/SC42 WG3委員



2022/11/14発売
（Amazonより）



執筆担当：「第6章 第3節
AI モデルの透明化技術と
品質保証」
2024/10/31発売
（技術情報協会より）

2022/7/1発売
（PDFのみ、日本規格協会より）

本日ご紹介する内容

1. 欧州AI法の要求概要
   - 付録１：AI法規や標準化の動向

・各国の法規化拡大
・2024年1月 AI機能安全規格発行（ISO/IEC TR 5469）

2. AIシステムの品質や安全を保証するための
   AIモデル開発の肝

SEAMS
（2017年～）

3. 未来社会に向けた弊社の研究活動
   - 付録２：人間社会とAIの共進化を下支えする基盤技術の研究開発（HMCESプロジェクト）

Human-Machine Co-Evolution Systems
（2022年～）

＜注意事項＞
今回ご紹介する弊社の活動は、まだ標準化や技術確立されていなかったり、合格基準が明確でない新しい領域へのチャレンジでございます。
そのため、弊社の活動内容や考え方が正しいか否かは、現時点では未知でございますので、あくまで１つの参考としてご活用ください。

1. 欧州AI法の要求概要

欧州AI法（EU AI ACT） REGULATION - EU - 2024/1689 - EN - EUR-LEX

- 動向
  - 2021年4月 欧州委員会より提案 ➡ 2023年6月 欧州議会により修正案可決
    ➡ 2024年5月 欧州委員会によりAI法案最終承認(成立) ➡ **2026年 全面施行**

制裁金 最大3500万ユーロ（約56億円）

- 安全上の影響度合いに応じて、AIを４段階に分類
  - 禁止／ハイリスク／限定的リスク／最小リスク
- ハイリスクAI対象（Article 6）
  - (a) 安全コンポーネント/ANNEX Ⅰ 且つ (b)第三者適合性評価対象
    - （ANNEX Ⅰ）欧州調和法制対象：
      - （Section A）機械、玩具、娯楽用具、エレベータ、防御措置、無線機器、圧力機器、ケーブルウェイ設備、個人防御器具、燃焼機器、医療機器、対外診断用医療機器
      - （Section B）民用航空安全、二輪・三輪・四輪車、農林業用車両、海洋機器、鉄道システム、自動車・トレーラー関連
  - （ANNEX Ⅲ）意思決定/健康/権利等に危害あり（生体認証、重要インフラ、教育と職業訓練、雇用・労働者管理・自営業アクセス、必須民間/公的サービス、法執行機関、移民・亡命・国境管理、司法行政と民主的プロセス）
- AI法の役割
  - AIシステムの安全、ガバナンス、市場開発促進
  - CEマーキング等の規制要件、AI認証制度

- 必須要求概要
  - **継続的なリスク管理プロセスと是正措置**
    - **データ品質、文書化、トレーサビリティ、透明性、人間による監視、正確性、堅牢性**
      - 大半がハイリスクAIに関する要求事項
      - **生成AIなども透明化義務が必要**
  - 市場投入前の技術文書
  - 自動ログによる追跡可能性
  - 規制サンドボックス（実験環境で事前確認）
  - 審査
    ＋ EU データベース登録【ハイリスクAI向け】

課題の多いAIモデル開発範囲

・EU AI ActやAI原則等の要求事項・合格基準は具体的ではない
・AIモデル開発（下記赤枠）の品質・安全について、詳細にガイドした標準は比較的少ない

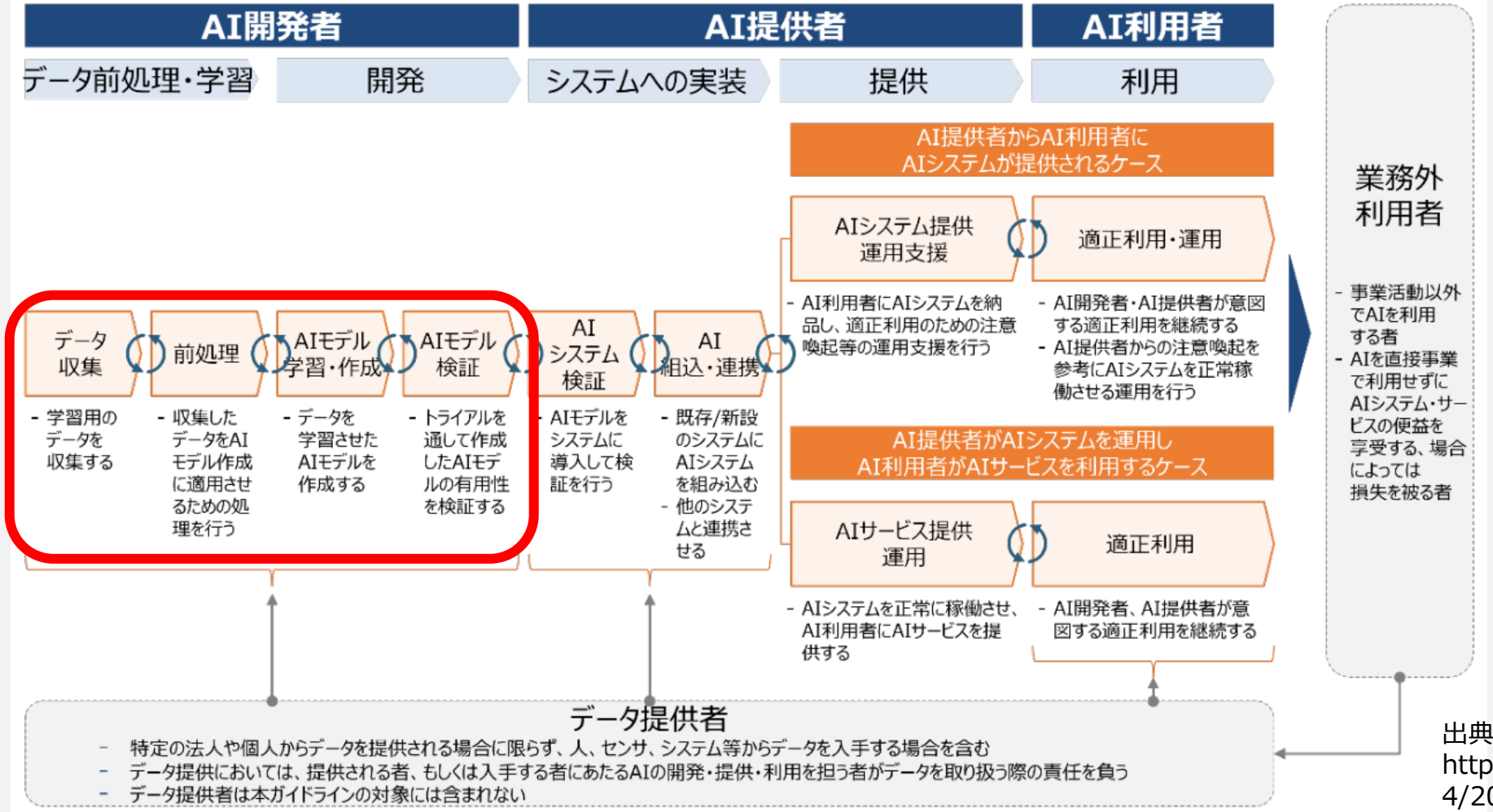


図 3. 一般的な AI 活用の流れにおける主体の対応

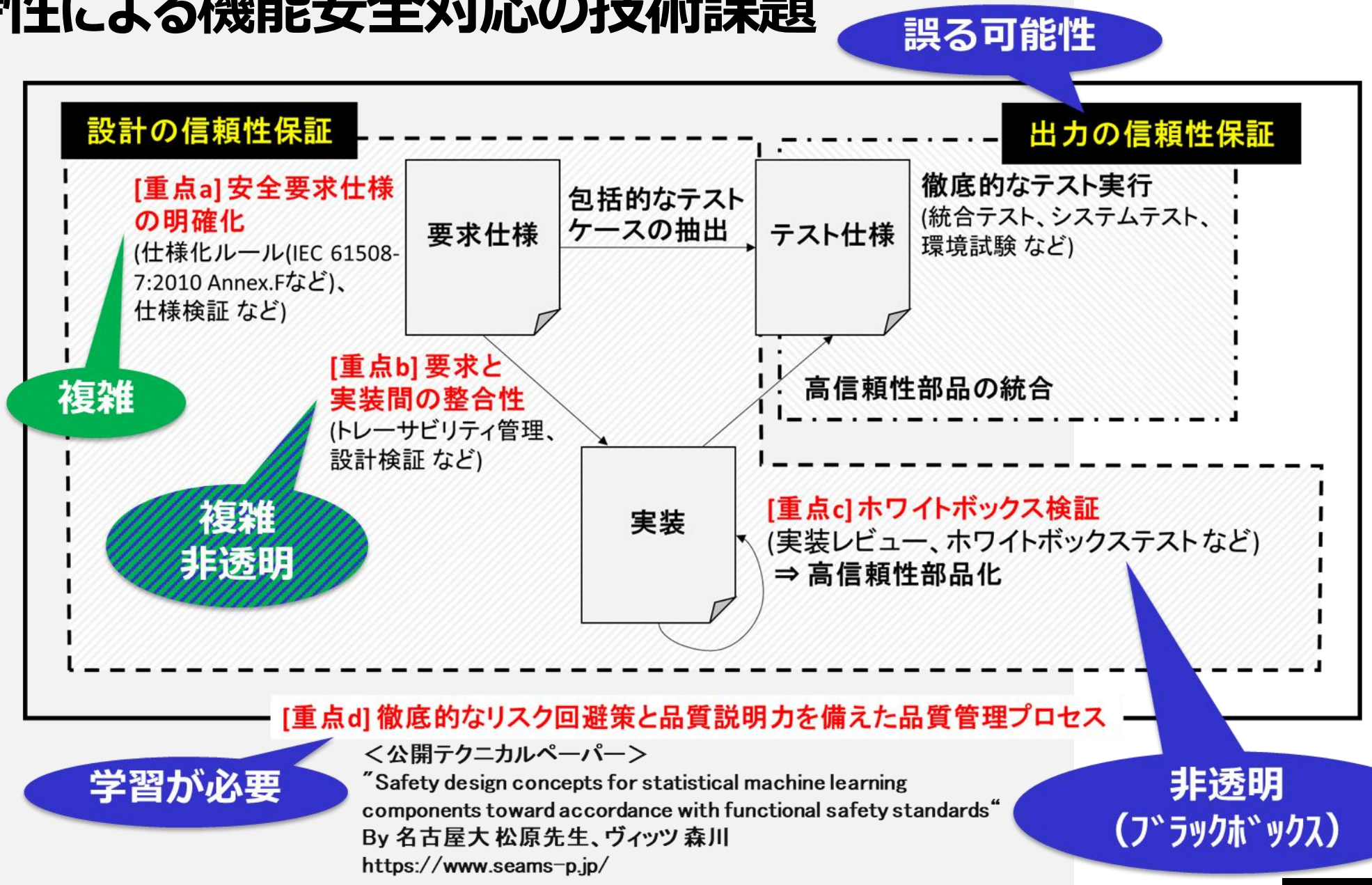
出典：AI 事業者ガイドライン（第 1.0 版）
https://www.meti.go.jp/press/2024/04/20240419004/20240419004-1.pdf

2. AIシステムの品質や安全を保証するためのAIモデル開発の肝

SEAMS

機械学習の特性による機能安全対応の技術課題

- 右図は、機能安全プロセスの「肝」と、機械学習開発のギャップを整理
- 青吹き出し：純粋な機械学習の課題
- 緑吹き出し：複雑なシステム（曖昧な仕様）による課題



＜公開テクニカルペーパー＞
"Safety design concepts for statistical machine learning components toward accordance with functional safety standards"
By 名古屋大 松原先生、ヴィッツ 森川
https://www.seams-p.jp/

機能安全適合重点項目に対応するための機械学習プロセス （※弊社提案手法）

重点項目	従来ソフト開発手法	機械学習開発のあるべき姿
【重点a】安全要求仕様の明確化	システム仕様や設計内容を明確に定義（準形式記述なども活用）	特性（苦手ケース）を熟知した上で、システム仕様を明確に定義
【重点b】要求と実装間の整合性	仕様 ⇒ 設計 ⇒ 実装 と段階的に詳細化しながら、トレーサビリティにより一貫性を確保	システム仕様 ⇒ データ要求仕様 ⇒ データセット のトレーサビリティ
設計検証	シミュレーション等による動的検証、 設計レビュー	PoC開発による実現性評価、 DNNの検証観点のレビュー（参考標準：A-SPICE MLE、UL4600、etc）
実装（ホワイトボックス）	ソースコード	学習データセット、ハイパーパラメータ、機械学習アルゴリズム など
実装（ブラックボックス）	バイナリー（機械語）	ネットワーク ＋ 重みづけ
【重点c】ホワイトボックス検証	コードレビュー、 ホワイトボックステスト など	学習データセットの検証（データ選定、データ分布など含む）、 機械学習モデルのバリデーション など
包括的なテストケース抽出・テスト実行	明確に定義された仕様を元に、網羅的にテストケース抽出・テスト実行	同左
【重点d】徹底的なリスク回避策と品質説明力を備えた品質管理プロセス（プロセスの透明化）	ISO9001 ＋ AutomotiveSPICE ＋ 機能安全プロセス	同左 ＋ 機械学習向けに拡張 （AutomotiveSPICE v4 MLE,SUP11、ISO/IEC 5338などのAI標準対応）
危険側故障率の定量評価	ハードウェアメトリクス算出評価 （部品の故障率 ＋ 故障診断率）	危険側になる機械学習の不確かさの定量評価 （IEC 62998-1を応用）

AIシステムの「特性の熟知」が肝

- 特性を熟知することの意義
  - DNNの判断精度改善
  - システムの保証可能条件の根拠として
  - DNN開発をシステマティックにする
    - 従来のソフト開発と同等のプロセスとして扱う
    - 機械学習モデルの透明化につながる
  - システムテストの網羅性
    - エッジ/コーナーケースの把握

- 特性熟知方法
  - ①ベースは各種標準（自動運転の場合の例は下記）に記載の観点（シーン、シナリオ、ODD）を考慮
    - NHTSAガイド：APPROACH FOR DERIVING SCENARIOS FOR SAFETY OF THE INTENDED FUNCTIONALITY
    - ISO 26262-3、ISO 21448、UL4600 、ISO 3450xシリーズ 他
  - ②繰り返しの評価&改善により、可能な限り特性（苦手ケース）を熟知し、製品保証範囲（ODD）を決定する

- 熟知したエキスパートによるジャッジが必要（従来開発同様）

SEAMSプロジェクト　https://www.seams-p.jp/

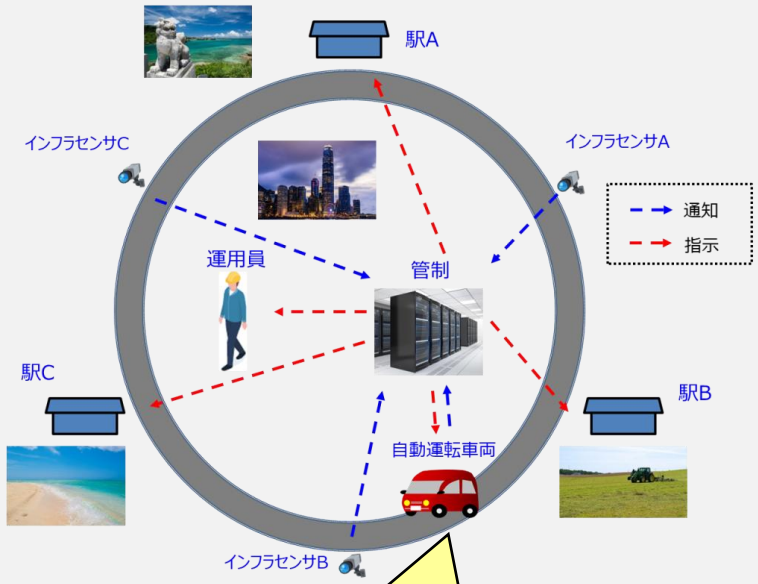
中部経済産業局 平成29年度 戦略的基盤技術高度化支援事業　（2017年10月～2020年3月）

SEAMS　「自律的自動運転の実現を支える人工知能搭載システムの安全性立証技術の研究開発」

パイロットシステムを対象にした具体的な分析・設計・評価の実施

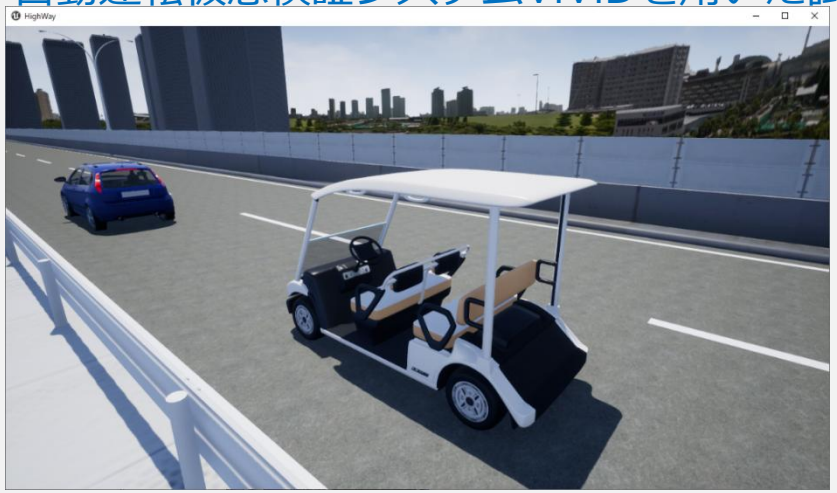
リゾート地における無人輸送サービス

自動運転車両による限定区間内輸送システム

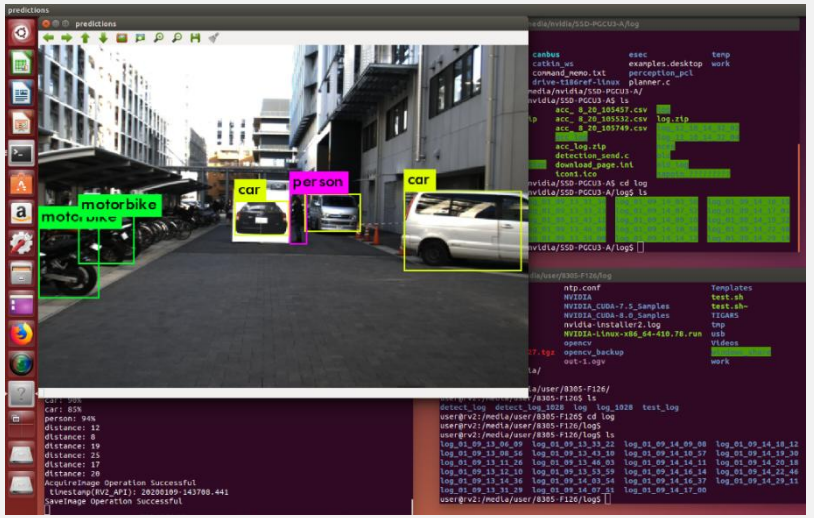


ドライバ無しの自動運転車両
（自動運転レベル4）
最高速度　50km/h

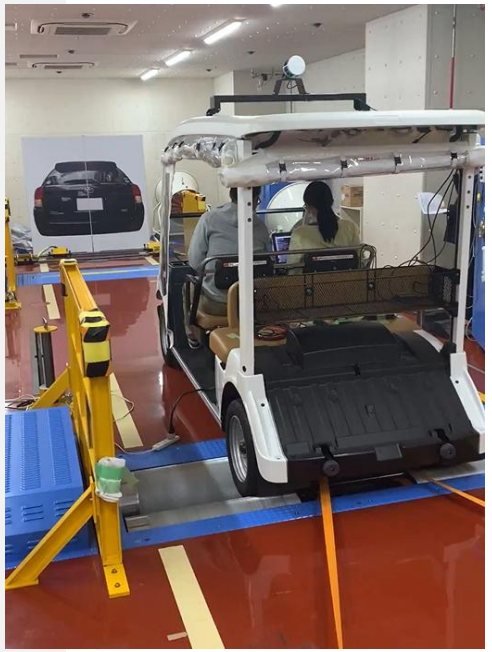
自動運転仮想検証システムViViDを用いた試験



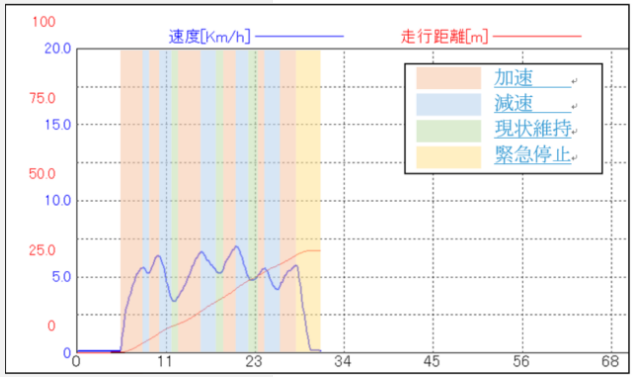
屋外での物体検出性能評価



シャーシダイナモ施設における動作確認および走行試験



車速・走行距離の計測



シミュレータ環境及び自動運転車両（ゴルフカート）を用いた実機環境で、実証実験を実施

SEAMSが提供するもの

SEAMSが提供する主な技術要素

- AIシステムの機能安全設計・評価技術
- AIシステムの不確かさの定量評価手法
- AIシステムの説明性の高い開発・学習プロセス
- 信頼できる機械学習の構築方法、評価方法
- AIの説明可能なモデリング手法
- DNNのトレーサビリティ手法
- 機械学習のデザインパターン、アンチデザインパターン
- 不確実性分析方法
- AIのサイバーセキュリティ対応
- 各種最新AI標準のまとめ
- 自動運転システムの安全性立証技術　他

実開発を踏まえて
開発者が楽になるもの
を開発者目線で整備

AIシステムの
安全基準適合支援ツール
（旧称：SEAMSガイドライン）

- 手順書
- テンプレート
- チェックリスト
- 開発成果物作成例
- 技術カタログ
- 検証ツール　など

弊社事例：物体検出モジュール設計書の作成例

AI開発向けに技術的な情報を拡張した開発文書が必要

現実的な開発スタイルを踏襲し、以下標準のAI関連要求を満たす開発事例を作成（開発テンプレート化）

・UL 4600 8.5章　・EU AI Act TitleIII-Chapter2　・ISO/TR 4804（SaFAD）　・ISO 34502

AI標準への適合戦略

ODD観点を元にした網羅的な仕様定義

「基準AIモデル」の定義

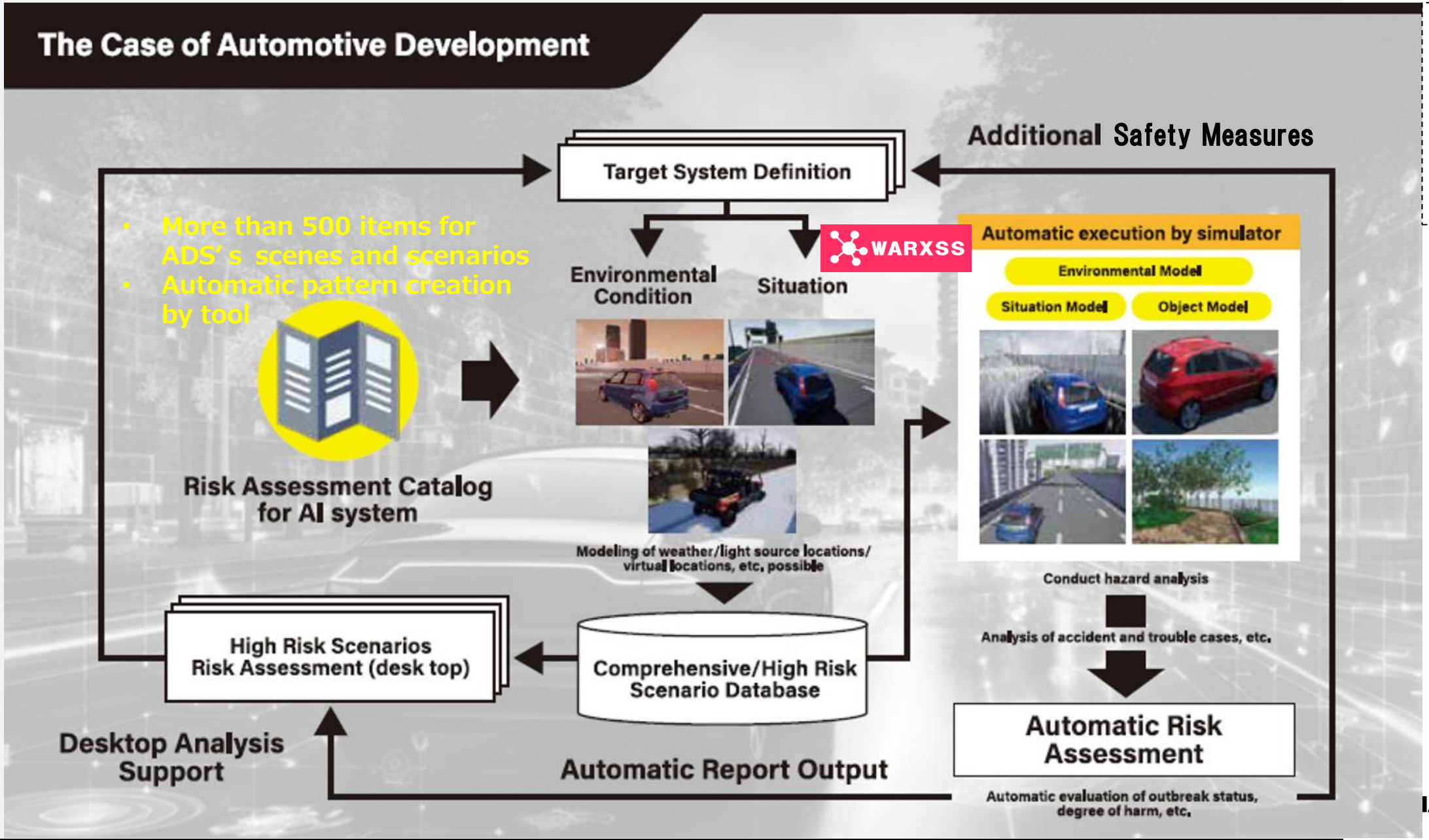
「基準AIモデル」に対する差分開発プロセスを定義

アノテーション/ラベリング仕様定義を追加

「基準AIモデル」に対する差分開発の品質論証

ODDに対応したデータ選定の具体的手法

理詰めによる網羅検証 × 仮想環境による検証自動化



※WARXSS®は、3D空間を用いたMaaSのリスク検証ツールです。MaaSに登場する移動体(モビリティ/その他車両/人)の動きを3Dの仮想空間の中で再現し、あらゆるシチュエーションにおけるモビリティサービスのリスクを視覚的に確認できます。また、LiDARセンサの検知範囲を表示する機能や、設置カメラ視点の表示にも対応しています。

３．未来社会に向けた弊社の研究活動

未来社会を創る『共進化』とは

人・機械(AI)・社会が『共に進化』

①AIシステムの継続的な高度化【機械(AI)の進化】
②個々の人間の価値観・行動の変化【人間の進化】
③社会基盤・生活様式など、社会システムの
継続的な変革【人間社会/環境の進化】

＜未来社会を創る共進化コア技術＞
- 過去：車、パソコン、インターネット、スマホ、etc
- 今後：**人工知能（AI）**

Society 5.0, Industry 5.0 を“促進”

- Well-being向上
- 生産性向上
- まだ見ぬ未来社会を創造

潜在リスク!!

「機械の潜在能力を持続的に向上させる 共進化(Co-evolution)ガイドラインの研究開発」

Human-Machine Co-Evolution Systems

HMCES Project　https://www.hmces-p.jp/

令和４年度 中小企業庁 成長型中小企業等研究開発支援事業 採択事業（2022年9月～2025年3月）

安全安心な
人-機械協調の
システム開発要求

[Theme-1]

Human-Machine Co-Evolution Systems GUIDELINES

人-機械協調システムの安全安心を
担保するための要求整理

Contribution

標準化活動
（AIシステム、協調安全など）

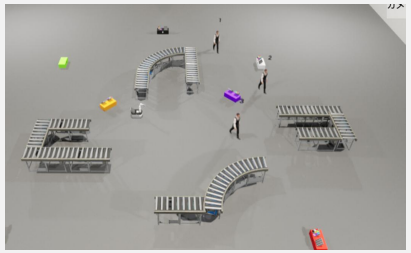
Feedback

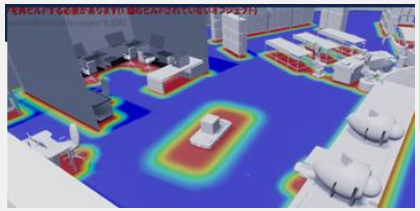
ガイドライン/法整備

安全安心な
システム運用要求
/社会ルール要求

[Theme-3]

Advanced Pilot APPLICATION





手･足･目･耳･口の
統合制御ロボット開発

[Theme-2]

Human-Machine Co-Evolution Systems PLATFORM

基盤技術
の
社会実装

人-機械の
共生社会システム

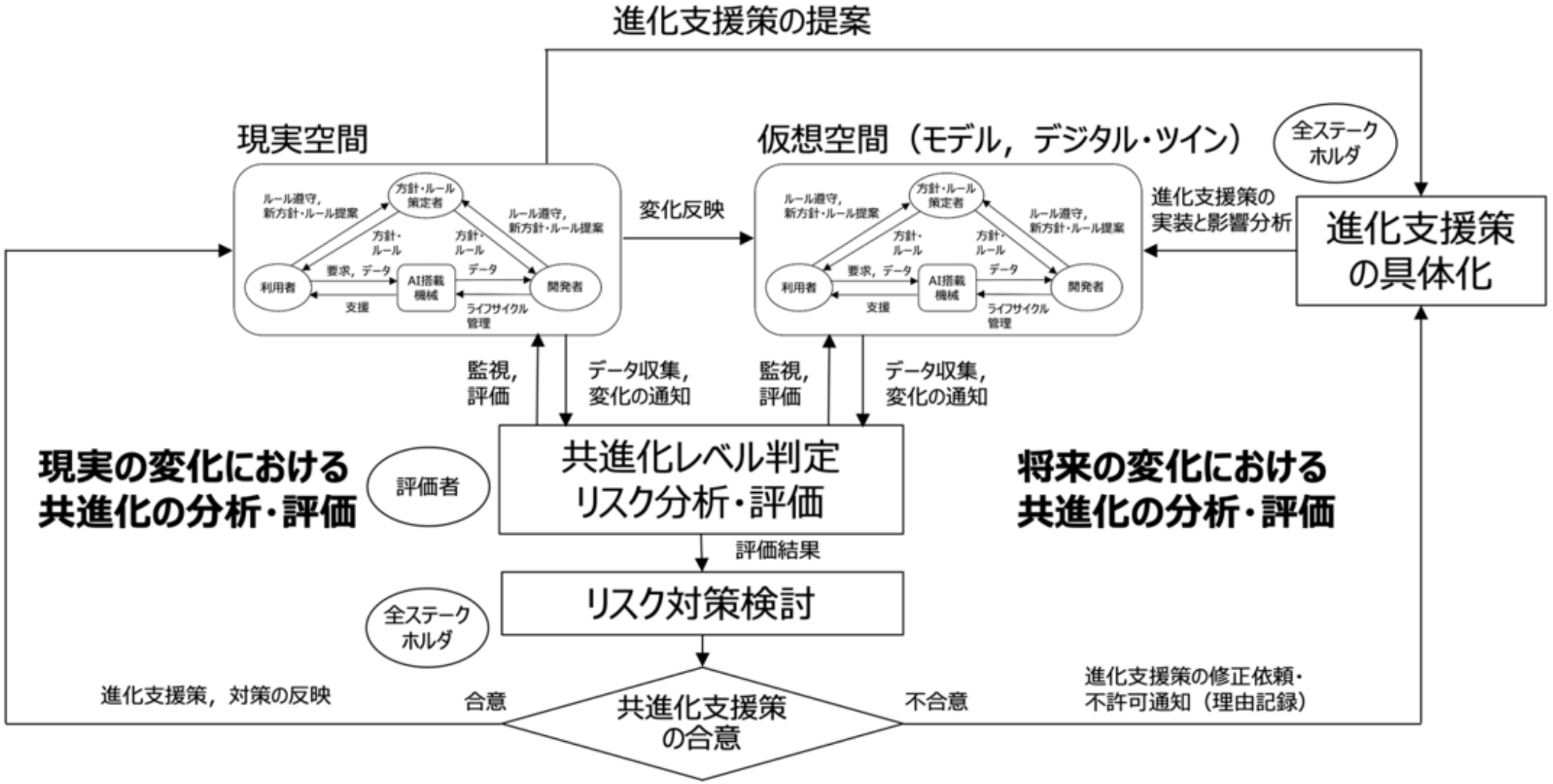
Human-Machine Co-Evolution Systems PLATFORM

人･機械(AI)･社会の『共進化』による未来社会の創造
➡ Society5.0／Industry5.0 基盤技術が必要

＜応用例＞
･インテリジェント協働ロボット（工場内）
･Housekeeper robot（家庭内）
･Secretary robot（オフィス内）
･etc

【研究実績】共進化フレームワーク（共進化ガイドブックに記載）

共進化フレームワークを提案し、各工程（ISO/IEC 22989ベース）× 各ステークホルダー の実施事項を整理



研究実施体制

Human-Machine Co-Evolution Systems

※アドバイザー/オブザーバーとして
今後も研究にご協力いただける
企業様募集中！

中小企業庁

公益財団法人
中部科学技術センター
（管理法人）

研究実施機関

株式会社イマジナリー(PL)
（機能安全）

国立大学法人 名古屋大学(SL)
（安全・セキュリティ）

UNIVERSITY

株式会社ヴィッツ（システム設計）
株式会社アトリエ（安全工学）
合同会社Gomes Company（人工知能実装）

国立研究開発法人 産業技術総合研究所
（協調安全）

川上・川下ネットワークを
利用したコンソーシアム活動
産官学連携による推進

アドバイザー

三菱電機 情報技術総合研究所
産業機械知見による助言

株式会社アイシン
人ー機械のコミュニケーションの助言

スズキ株式会社
人-機械の協調制御の助言

一般財団法人 日本自動車研究所
安全・サイバーセキュリティに関しての助言

コベルコ建機株式会社
建設機械知見による助言

株式会社明電舎
自動運転などの安全性評価システム開発者視点での助言

株式会社UL Japan
品質・安全の専門家視点での助言

DNV ビジネス・アシュアランス・ジャパン株式会社
認証機関という安全の専門家の視点での助言

オブザーバー

（未登録協力者）

株式会社日立製作所
自動制御システムの研究従事者の視点での助言

ISO/IEC JTC1/SC42 （AI国際規格策定）
AIシステムの標準化活動の視点での助言

独立行政法人中小企業基盤整備機構

中部経済産業局 地域経済部イノベーション推進課

まとめ

当社グループの保有する安心安全を担保する技術を結集して我々は未来社会に貢献してまいります

IMAGINARY × 株式会社 ヴィッツ

SEAMS

Human-Machine Co-Evolution Systems

WARXSS

＜主な技術支援＞
- エンジニアリング
- コンサルティング
- ツール/マテリアル販売
- 研修講師
- 評価、監査
- システム構築

IMAGINARY

THANK YOU!

株式会社イマジナリー
執行役員
森川 聡久

+81 50 5211 5282
morikawa@imaginary-inc.jp
www.imaginary-inc.jp

付録1：
AI法規や標準化の動向

主なAI原則・法規制関連情報 ～SEAMSガイドラインが対応するものを抜粋～

国/地域	AI原則や法規制関連情報	URL
欧州	EU AI Act ※2024年5月21日　欧州委員会により最終承認	https://artificialintelligenceact.eu/
欧州	消費者を保護し、イノベーションを促進するための、 新しい製造物責任指令（COM(2022) 495）とAI責任指令	https://ec.europa.eu/commission/presscorner/detail/en/ip_22_5807
米国	NIST AI RMF（AI Risk Management Framework）	https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development
米国	アメリカ大統領令　政府におけるAIの使用に関する原則	https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/
国際	OECD AI原則	https://oecd.ai/en/ai-principles
UAE	ドバイ AI原則と倫理	https://www.digitaldubai.ae/initiatives/ai-principles-ethics
マルタ	TOWARDS TRUSTWORTHY AI MALTA‘S ETHICAL AI FRAMEWORK	https://malta.ai/wp-content/uploads/2019/10/Malta_Towards_Ethical_and_Trustworthy_AI_vFINAL.pdf
中国	北京AI原則	https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/
英国	Understanding artificial intelligence ethics and safety	https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety
英国	Algorithmic Transparency Recording Standard	https://www.gov.uk/government/publications/algorithmic-transparency-template
シンガポール	MODEL ARTIFICIAL INTELLIGENCE GOVERNANCE FRAMEWORK SECOND EDITION	https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf
カナダ	Directive on Automated Decision-Making	https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

主な標準化活動・ツール等 ～SEAMSガイドラインが対応するものを抜粋～

国/地域	標準化活動・ツール等	URL
国際	ISO/IEC JTC1/SC42 Artificial intelligence	https://www.iso.org/committee/6794475.html
米国	IEEE 70xxシリーズ（自律的で知的なシステムにおける倫理の実践）	https://ethicsinaction.ieee.org/p7000/
英国	BSI PAS 188xシリーズ（自動運転の安全に関する規格群）	https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development
欧州	ETSI ISG SAI（欧州電気通信標準化機構　人工知能の保護に関する業界仕様グループ）による標準化検討	https://www.etsi.org/technologies/securing-artificial-intelligence
欧州	ALTAI（THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE）	https://dm2ue6l6q7ly2.cloudfront.net/wp-content/uploads/2020/07/24110834/ALTAI_final_14072020_CS_accessible2_jsd5pdf.pdf
国際	GPAI（Global Partnership on AI）「Responsible Development, Use and Governance of AI Working Group Report」	https://gpai.ai/projects/responsible-ai/gpai-responsible-ai-wg-report-november-2020.pdf
国際	OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS	https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm
日本	AI 事業者ガイドライン【総務省＆経済産業省】	https://www.meti.go.jp/press/2024/04/20240419004/20240419004-1.pdf
日本	ＡＩ利活用ガイドライン ～ＡＩ利活用のためのプラクティカルリファレンス～【総務省】	https://www.soumu.go.jp/main_content/000637097.pdf
日本	AI・データの利用に関する契約ガイドライン －AI編－【経済産業省】	https://www.meti.go.jp/press/2019/12/20191209001/20191209001-3.pdf
ドイツ	German Standardization Roadmap Artificial Intelligence	https://www.dke.de/en/areas-of-work/core-safety/standardization-roadmap-ai
シンガポール	ISAGO（Companion to the Model AI Governance Framework －Implementation and Self-Assessment Guide for Organizations）	https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgisago.pdf
シンガポール	AI GOVERNANCE TESTING FRAMEWORK & TOOLKIT	https://file.go.gov.sg/aiverify-primer.pdf
フランス	Be involved in writing voluntary standards for artificial intelligence	https://www.afnor.org/en/news/shaping-european-ai-leadership/

主なAIセキュリティ関連技術等 ～SEAMSガイドラインが対応するものを抜粋～

国/地域	AIセキュリティ関連標準化活動	URL
国際	ISO/IEC 27090　— Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems	https://www.iso27001security.com/html/27090.html
欧州	ETSI ISG SAI（欧州電気通信標準化機構　人工知能の保護に関する業界仕様グループ）による標準化検討	https://www.etsi.org/technologies/securing-artificial-intelligence
欧州	ENISA「Artificial Intelligence Cybersecurity Challenges」	https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges
米国	NIST IR 8269「A Taxonomy and Terminology of Adversarial Machine Learning」	https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf
米国	Microsoft「Threat Modeling AI/ML Systems and Dependencies」	https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml
日本	機械学習システムセキュリティガイドライン　（機械学習工学研究会 MLSE）	https://github.com/mlse-jssst/security-guideline
日本	日本銀行金融研究所「機械学習システムのセキュリティに関する研究動向と課題(2018.8)」	https://www.imes.boj.or.jp/research/papers/japanese/kk38-1-6.pdf
日本	日本ネットワークセキュリティ協会「パネルディスカッション「AIセキュリティ」その脅威と対策を考える（JNSA NSF 2020）」	https://www.jnsa.org/seminar/nsf/2020/data/A2_IoTSecurityWG.pdf

EU AI ACT ハイリスクAIへの主な開発要件（1/3）

※2021年ドラフト提案時の要求ベース

- **リスクアセスメント&リスク低減プロセス**（第9条）＋**その検証**（第9条） HARA

- **品質基準を満たすトレーニング、検証、およびテストのデータセット**に基づいた開発（第10条）

  データセット

- AI システムの準拠を評価するために必要なすべての情報を提供する（付属書IVに規定の要素を含むこと）（第11条）

- ハイリスクAI システムの稼働中に**イベント（「ログ」）の自動記録を可能にする機能**を備えて設計および開発されるものとします。（第12条）

  機能

  開発プロセス

- ユーザーがシステムの出力を解釈して適切に使用できるように、その操作が十分に**透過的であることを保証する方法で設計および開発**されるものとします。このタイトルの第3 章に記載されているユーザーとプロバイダーの関連する義務の遵守を達成するために、適切な種類と程度の透明性が確保されるものとします。（第13条）

EU AI ACT ハイリスクAIへの主な開発要件（2/3）
※2021年ドラフト提案時の要求ベース

- **ユーザーがシステム出力を解釈して適切に使用できるようにするための情報**（第13条 第3項） 機能
  - (a) プロバイダーの身元と連絡先の詳細、および該当する場合はその正式な代表者の詳細。
  - (b) 以下を含む、ハイリスク AI システムの性能の特性、能力、および限界
    - (i) その意図された**目的**。 HARA/TARA 開発プロセス AI性能
    - (ii) 第 15 条で言及されている、高リスク AI システムが**テストおよび検証**され、予想される**精度、堅牢性、およびサイバーセキュリティ**のレベル、およびその予想されるレベルに影響を与える可能性のある既知および予測可能な状況正確性、堅牢性、サイバーセキュリティ。
    - (iii) 意図された目的に従って、または合理的に予測可能な誤用の条件下でのハイリスク AI システムの**使用に関連する既知または予見可能な状況**。これは、健康と安全または基本的権利へのリスクにつながる可能性があります。 HARA
    - (iv) システムの使用が意図されている個人または個人のグループに関するそのパフォーマンス。
    - (v) 適切な場合、入力データの仕様、または AI システムの意図された目的を考慮した、**使用されるトレーニング、検証、およびテストデータセット**に関するその他の関連情報。 データセット
  - (c) 最初の適合性評価の時点でプロバイダーによって事前に決定された、リスクの高い AI システムおよびそのパフォーマンスへの変更。
  - (d) ユーザーによる AI システムの出力の解釈を容易にするために導入された技術的手段を含む、第 14 条で言及される**人間による監視手段**。 HARA
  - (e) リスクの高い AI システムの予想寿命と、ソフトウェアの更新に関するものを含む、その **AI システムの適切な機能を確保するために必要なメンテナンスとケアの措置**。 HARA

EU AI ACT ハイリスクAIへの主な開発要件（3/3）

※2021年ドラフト提案時の要求ベース

- 意図された目的に照らして適切なレベルの精度、堅牢性、およびサイバーセキュリティを達成し、**ライフサイクル全体で一貫して機能するように設計および開発**されるものとします。（第15条 1）

開発プロセス

- **精度**のレベルと関連する精度の指標は、付属の使用説明書で宣言するものとします。（第15条 2）

AI性能

- 特に自然人や他のシステムとの相互作用により、システムまたはシステムが動作する**環境内で発生する可能性のあるエラー、障害、または矛盾に関して回復力**があるものとします。（第15条 3）

HARA/TARA ➡ FS/CS

- 権限のない第三者が**システムの脆弱性を悪用**して使用またはパフォーマンスを変更しようとする試みに関して、回復力があるものとします。（第15条 4）

- 市場に投入された後、またはサービスに投入された後も学習を続けるリスクの高いAIシステムは、将来の運用の入力として使用される出力（「フィードバック ループ」）によって**バイアスがかかる可能性のある出力が適切に処理されるよう**に開発する必要があります。（第15条　3）

HARA ➡ FS

信頼できるAIの評価リスト（ALTAI）

- THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI)
- 2020年に欧州委員会より提案された**信頼できるAIの自己評価用のチェックリスト**
- 信頼できるAIの7つの要件として、以下を挙げている。
  - 人的機関と監視
  - 技術的な堅牢性と安全性
    - セキュリティ
    - 安全性
    - 精度
    - 信頼性、フォールバック計画、および再現性
  - プライバシーとデータガバナンス
  - 透明性
    - トレーサビリティ
    - 説明可能性
    - AIシステムの制限に関するオープンなコミュニケーション
  - 多様性、無差別および公平性
  - 社会的および環境的幸福
  - 説明責任

AI事業者ガイドライン（総務省・経済産業省）

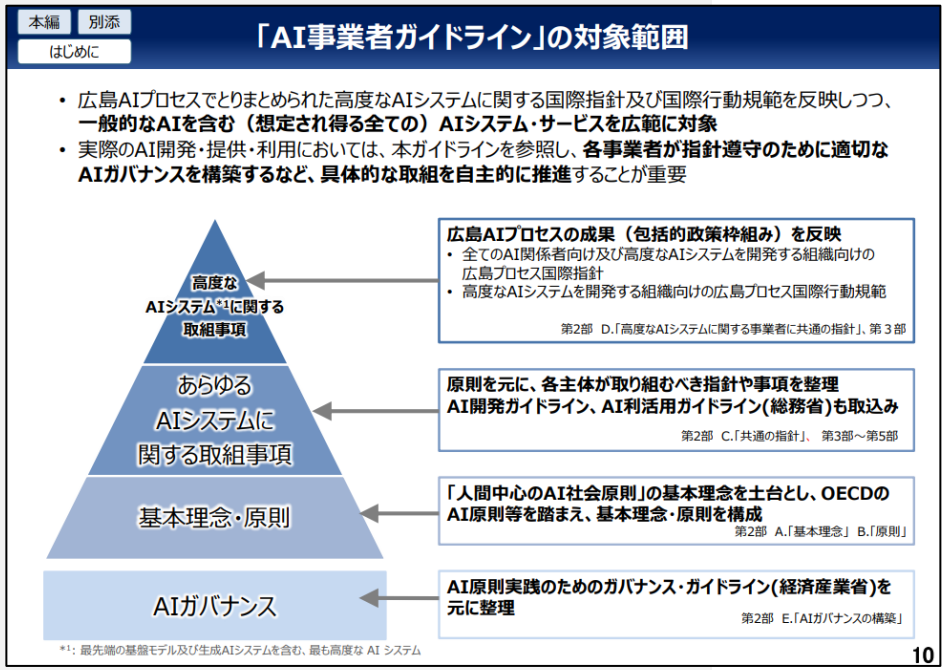
➡ 2024年1月：総務省、経産省：「AI事業者ガイドライン」案発表

https://www8.cao.go.jp/cstp/ai/ai_senryaku/7kai/13gaidorain.pdf

➡ 2024年4月：「AI事業者ガイドライン」発行

他国のAI原則（2020年～）でも重視されている項目
国際見渡して、基本的に大きな違いは無さそう

- 10原則
  - ①人間中心
  - ②安全性
  - ③公平性
  - ④プライバシー保護
  - ⑤セキュリティ確保
  - ⑥透明性
  - ⑦説明責任
  - ⑧教育・リテラシー
  - ⑨公正競争確保
  - ⑩イノベーション



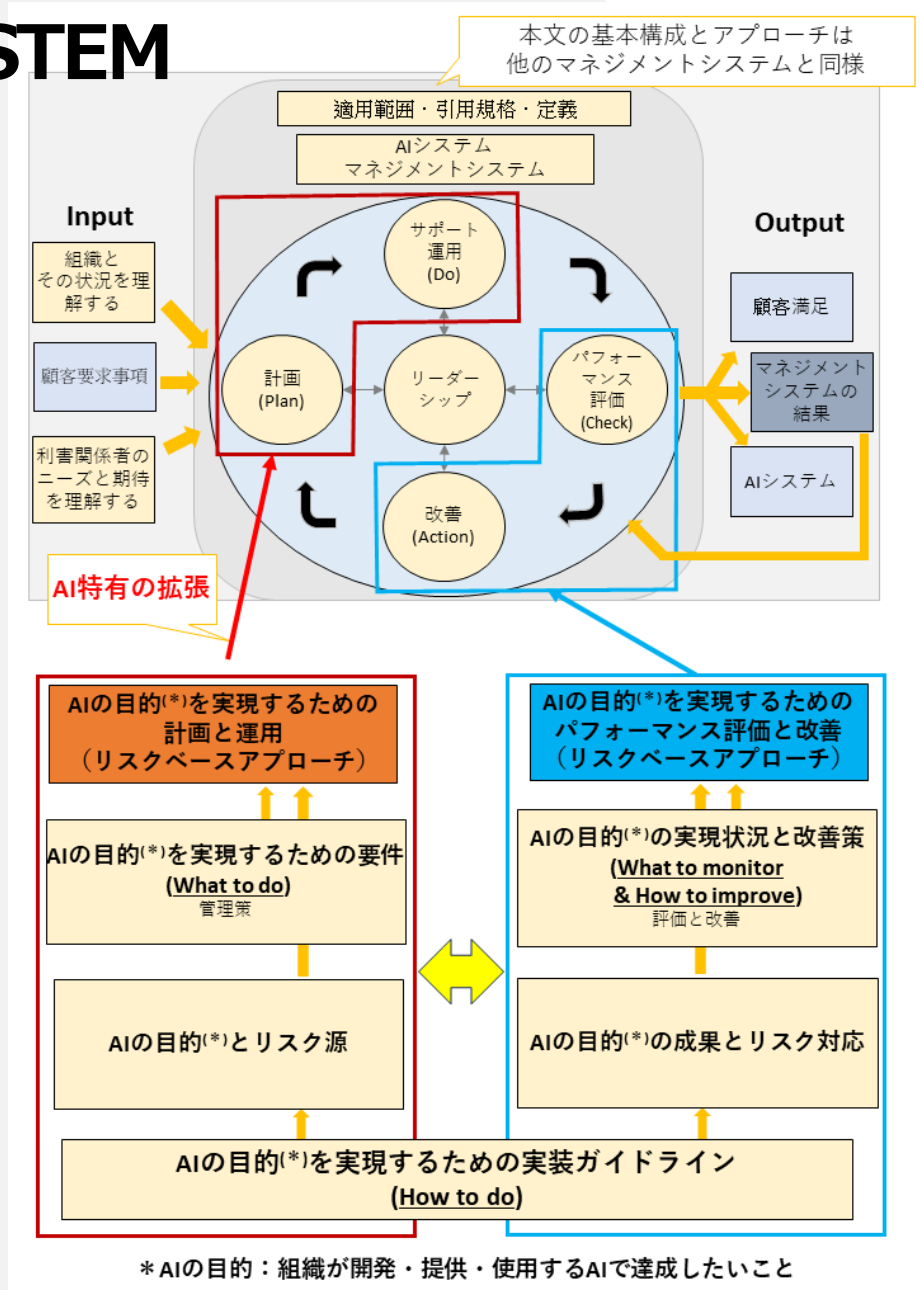
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240119_3.pdf

ISO/IEC 42001 AI MANAGEMENT SYSTEM

- 2023年12月発行
  - https://www.iso.org/standard/81230.html

- 経済産業省プレスリリース
  - https://www.meti.go.jp/press/2023/01/20240115001/20240115001.html
  - 右図は本サイトより引用

- 右図の通り、基本的には**「リスクベースアプローチ」**
  - 従来対応と同様



自動運転安全規格 ANSI/UL 4600
8.5 MACHINE LEARNING AND "AI" TECHNIQUES

- 8.5章より MANDATORY ＆ REQUIRED関連抜粋
  - 種類、技術
  - 構造
    - ネットワークタイプ、レイヤー数、パイパーパラメータなど
  - 役割定義
  - アルゴリズムの設計と実装の完全性
  - アルゴリズムによる判断の妥当性
  - 性能メトリック、評価、トレーサビリティ
    - ROCカーブ、偽陽性率、偽陰性率、精度など
  - V&V手法の妥当性
  - ツール、技術の使用方法の妥当性
  - データの種類と量が運行設計領域（ODD）に適合するか
  - 学習データの収集方法と管理方法
    - データ来歴、データ収集機器、データの前処理、データストレージ・管理ツール
  - データの整合性保証
  - データ収集の問題検出のためのテスト・分析
  - データ変動に対する許容可能なロバスト性
  - 導入後の変更対応（重み調整、再学習、構成変更）
    - 変更内容の説明、再検証
  - 学習を補足するためのフィールドデータ

NIST AI RMF PLAYBOOK：MEASURE 2.9

https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook

- **AIモデルは説明され、検証され、文書化され、AIシステムの出力は、MAP機能で特定されるように、その文脈の中で解釈され、責任ある使用とガバナンスを通知する。**
  - **説明可能なモデル、事後説明、監査ログ**を作成するためにシステムを開発する。
  - 可能または利用可能な場合は、従来の一般化線形モデルやペナルティ付き一般化線形モデル、決定木、最近傍モデルやプロトタイプベースのアプローチ、ルールベースのモデル、一般化加法モデル、説明可能なブースティング・マシン、ニューラル加法モデルなど、**本質的に説明可能なアプローチ**を利用する。
  - 説明の正確さ、明確さ、わかりやすさについて、AI関係者、エンドユーザー、影響を受ける可能性のある個人またはグループからフィードバックを得るために、**展開前に説明方法とその結果の説明をテスト**する。
  - モデルの種類（畳み込みニューラルネットワーク、強化学習、決定木、ランダムフォレストなど）、データの特徴、学習アルゴリズム、提案されている用途、決定閾値、学習データ、評価データ、倫理的配慮など、**AIモデルの詳細を文書化**する。
  - 人口統計学的グループおよび展開状況に関連するその他のセグメントにわたる**パフォーマンスとエラーの測定基準を設定し、文書化し、報告**する。
  - **視覚化、モデル抽出、特徴の重要性など、さまざまな方法を用いてシステムを説明**する。説明では複雑なシステムを正確に要約できないことがあるため、忠実性、一貫性、堅牢性、解釈可能性などの特性に従って説明をテストする。
  - 忠実度（ローカルおよびグローバル）、曖昧さ、解釈可能性、対話性、一貫性、攻撃/操作への耐性などの特性に従って、**システム説明の特性を評価**する。
  - **システム説明の品質**をエンドユーザーや他のグループとテストする。
  - 説明プロセスをゲーム化するなど、外部からの操作に対する脆弱性を回避するため、**モデル開発プロセスの安全性**を確保する。
  - 生産データに応じて調整されるモデルを含め、**時間の経過に伴うモデルの変化をテスト**する。
  - データステートメントやモデルカードなどの**透明性ツールを使用して、説明や検証情報を文書化**する。

付録２：
人間社会とAIの共進化を下支えする基盤技術の研究開発（HMCESプロジェクト）

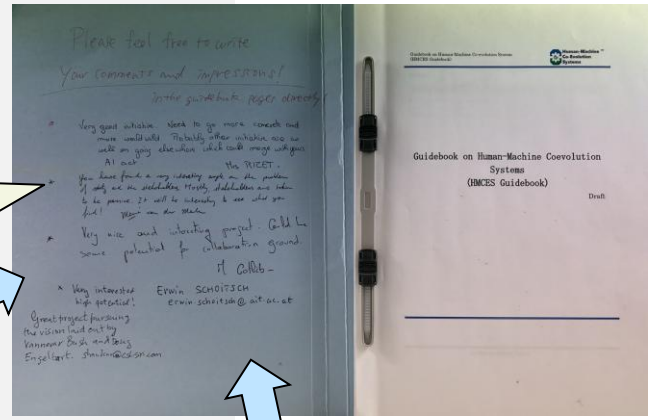
Human-Machine Co-Evolution Systems

【研究実績】共進化ガイドブックのドラフト作成＆コメント獲得

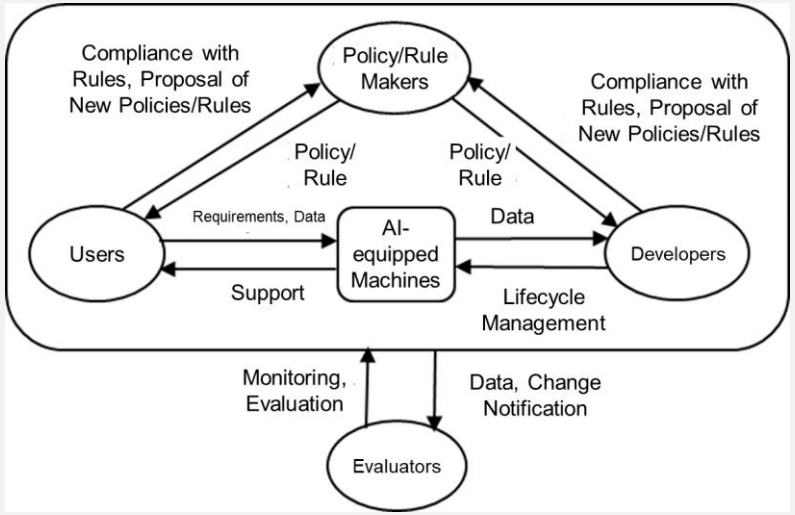
- 共進化システムの定義、共進化システムを構成する**ステークホルダの定義**を整理
- 共進化システムライフサイクルの定義を、**ISO/IEC 22989 のAIシステムライフサイクル**をベースに整理
- 共進化システムライフサイクルの各フェーズ毎に、**人-機械の共進化に必要な要件を整理**
- 共進化ガイドブックを国際学会（SafeComp2023, 2024）のプレゼン＆ポスターブースにて紹介し、国際的な専門家らのフィードバックコメントを獲得

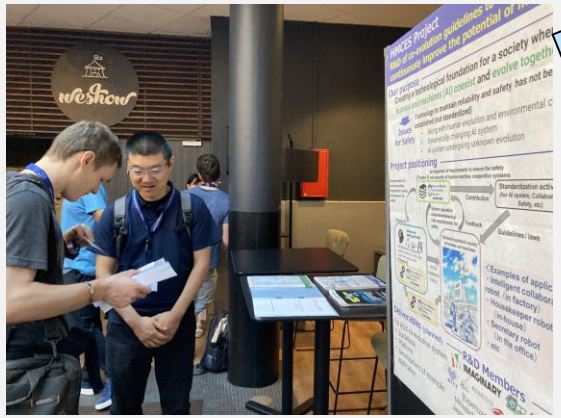
■研究活動ならびにガイドブックを紹介し、専門家から意見を獲得

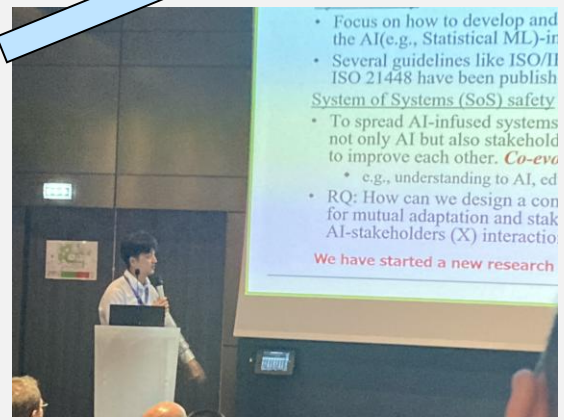
＜主なコメント整理＞
・本活動の必要性などポジティブな評価
・類似研究活動とのコラボ/融合希望
・技術課題：ステークホルダ毎（異なる立場）の説明責任方法、PFの機能への期待

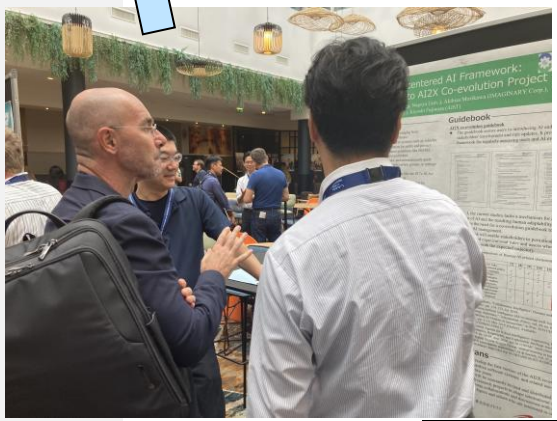


■共進化システムとステークホルダの定義







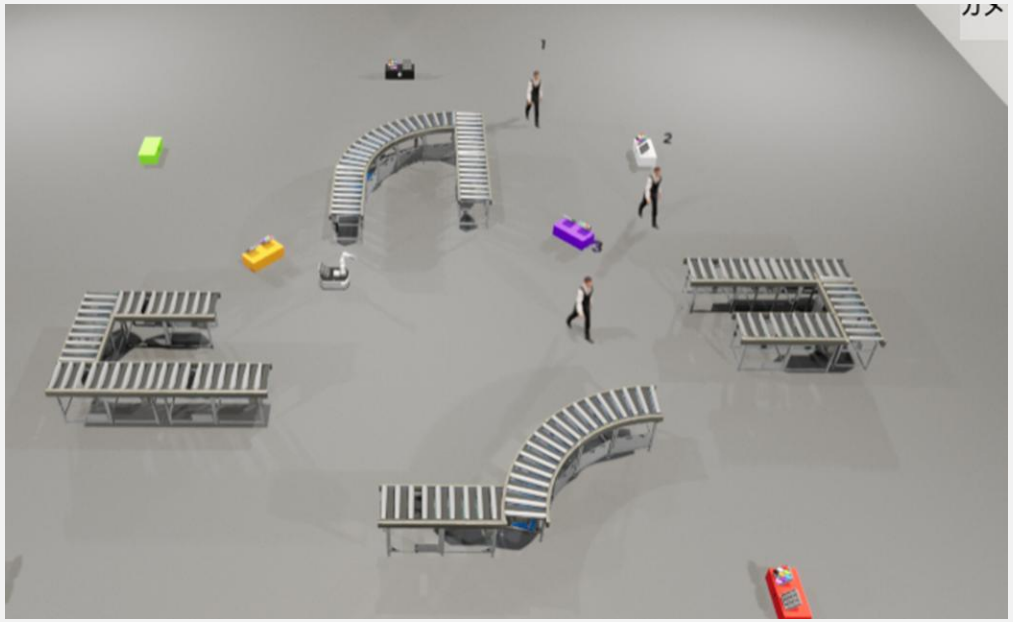


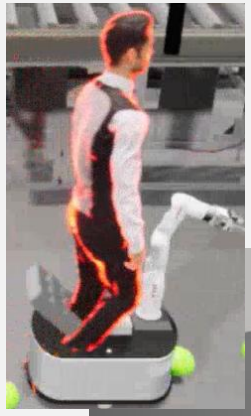
【研究実績】共進化システムの検証手法開発の取り組み

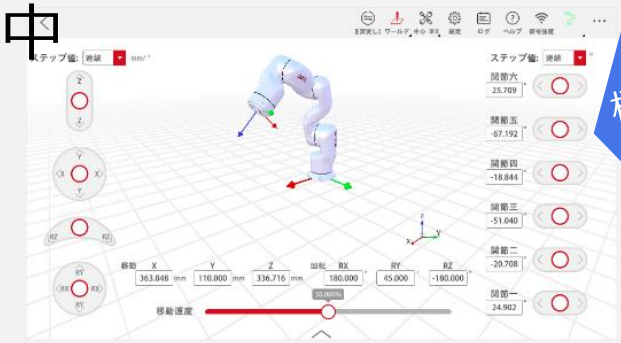
- 2023年9月 欧州の安全の知見者と技術検討実施
- 変化し続けるシステム、未知のシステムに対し、安全性を検証する手法について、具体的なアプリケーションを題材に検討
- 獲得技術
  - 共進化システムを分析・評価するための**モデリング、分析・評価方法**のコツ
  - 共進化システムの安全性を**形式検証**するための方法
- 今後、共進化ガイドブックにフィードバック予定

【研究実績】共進化パイロットシステム開発ならびに評価

- 共存・協働ロボット開発
  - システム構成：手（アームロボット）、足（自律搬送ロボット）、目（カメラ）、耳＆口（AIコンシェルジュアプリ）
- 人とロボットが、共通空間にて、分担しながら業務を実施
  - 拠点での作業　＋　拠点間の移動
- 人のスキル向上、ロボットの性能向上等のお互いの変化に応じ、共に振る舞いが進化
- 仮想空間上に同様システムを構築し、柔軟な動作検証
- 現在、共進化システム検証に必要な機能拡張中







極力忠実に仮想化