7th Grand Canvas: AI品質の未来を共に描く
～AI品質マネジメントネットワーキングシンポジウム～

エージェント経済圏の勃興
- AIセーフティとAI品質に関する課題 -

2026年2月3日

日本電気株式会社
上席主席研究員
森永　聡
mori-chin@nec.com

森永　聡（もりなが　さとし）　博士（工学）

日本電気株式会社　研究開発部門　上席主席研究員
産業技術総合研究所　ＮＥＣ-産総研AI連携研究室　副連携研究室長
神戸大学　数理・データサイエンスセンター　客員教授
日本人工知能学会　理事
電子情報通信学会　AI相互運用分科会　委員長
自律調整SCMコンソーシアム　理事長

略歴

平成6年4月：　ＮＥＣ入社
　耐故障システムの研究開発に従事。
平成12年1月～平成20年3月：金融（監督）庁出向/兼務
　国際課課長補佐、銀行一課特別研究員として、銀行のリスク規制制度の設計と実施を担当。
平成27年4月～　ＮＥＣデータサイエンス研究所　主席研究員
　機械学習・データマイニング・自動推論研究チームのリーダーとして、データ分析原理の理論的研究、当該原理に基づき効率よく分析を行うエンジンの開発、そのビジネス応用を推進。
平成28年1月～　産総研　NEC-産総研人工知能連携研究室　副連携研究室長　兼務
令和3年10月～　自律調整SCMコンソーシアム　理事長　兼務

主な業績：活動企画/立案から、基礎/応用研究遂行、事業化/社会実装までを実現

耐故障システムの冗長性固定信頼度最大化（2000年頃）、WEB上の評判分析サービス（2005年頃）
数理統計に基づく銀行のオペリスク管理規制（2008年頃）、ホワイトボックス予測（2015年頃）
自動データサイエンス(2020年頃）、シミュ・機械学習融合（2023年頃）、自動交渉SCM（2025年頃）

エージェント経済圏の勃興
- AIセーフティとAI品質に関する課題 -

生成AI技術の進展による、業務用エージェントの実用化

- Phase 1：単なるツールやデータ、デバイスが、部下・同僚・代理人へ
- Phase 2：社内でのエージェント間の協調・連携
- Phase 3：会社の境界を越えたエージェント間の調整・交渉

エージェント経済圏（X-as-an-Agent Economy）の形成・進展へ→数百兆円@2030

エージェントの、エージェントによる、エージェントのための、ビジネス

- エージェント（の動作）自体の取引（業務委託、派遣、エージェント売買等）
- エージェントによる経済活動（売買、投資、融資、保険、広告、、、、）
- そのためのサービス（実行基盤／能力検定／研修／保証／監視／隔離／捕獲／駆除／破壊等）

NECの取り組み

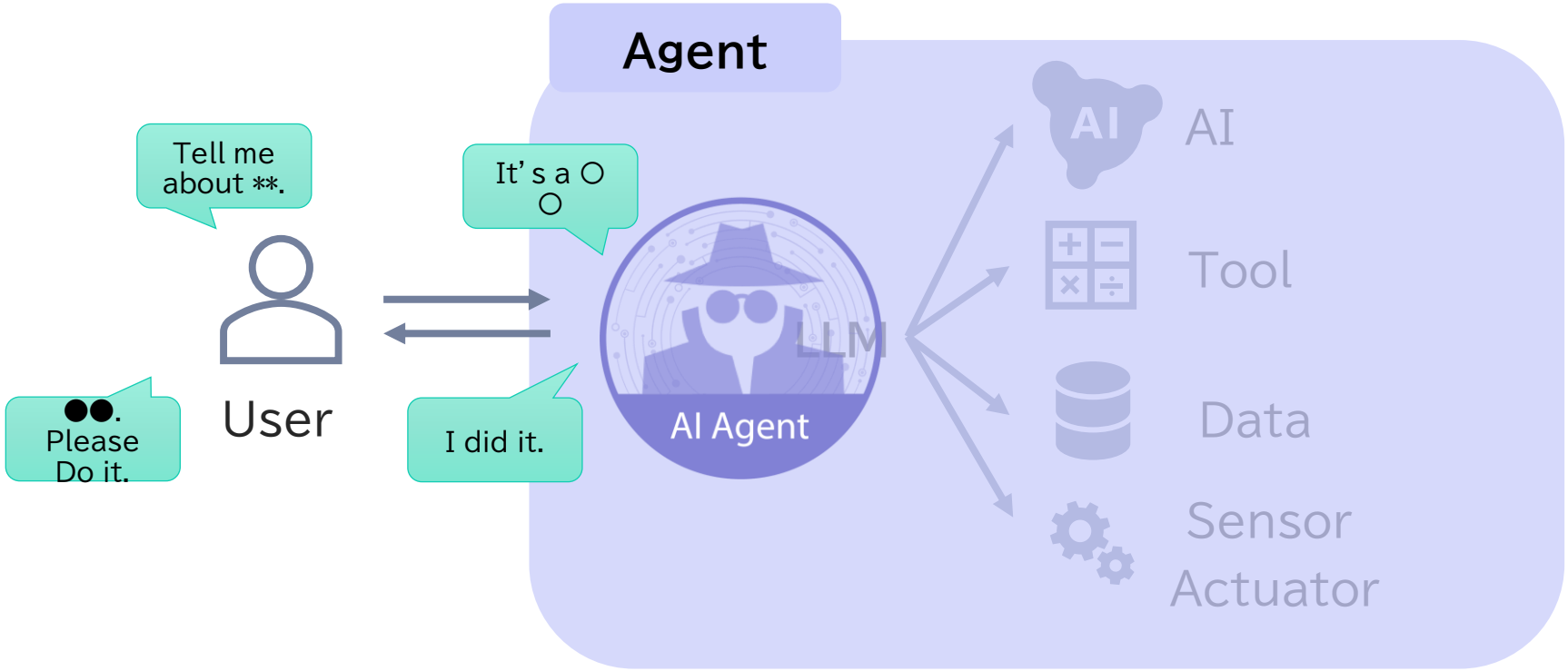
- 自動交渉エージェントによる社会価値提供「内部最適化・個別意思決定の限界の突破」
- 国内外での業界団体（DTC、自律調整SCMコンソーシアム）における当該コンセプト牽引活動
- 自動交渉プロトコルの国際標準化（UN/CEFACT “eNagotiation”）、国際技術コンペ（ANAC SCMリーグ）の主催
- 日本学術会議での提言、電子情報通信学会での委員会設立
- 産総研や他社（非開示）との共同研究、CATENA-Xとの意見交換、COCNからの政策提言

エージェント経済圏におけるAIセーフティとAI品質に関する課題

- ユーザにとっての課題、ベンダーにとっての課題、社会としての課題
- 今後、求められる取り組み

生成AI技術の進展による、業務用エージェントの実用化：Phase 1

「エージェントの体」をとることで、単なるツールやデータ、デバイスが、部下・同僚・代理人へ！！



Why Agent?

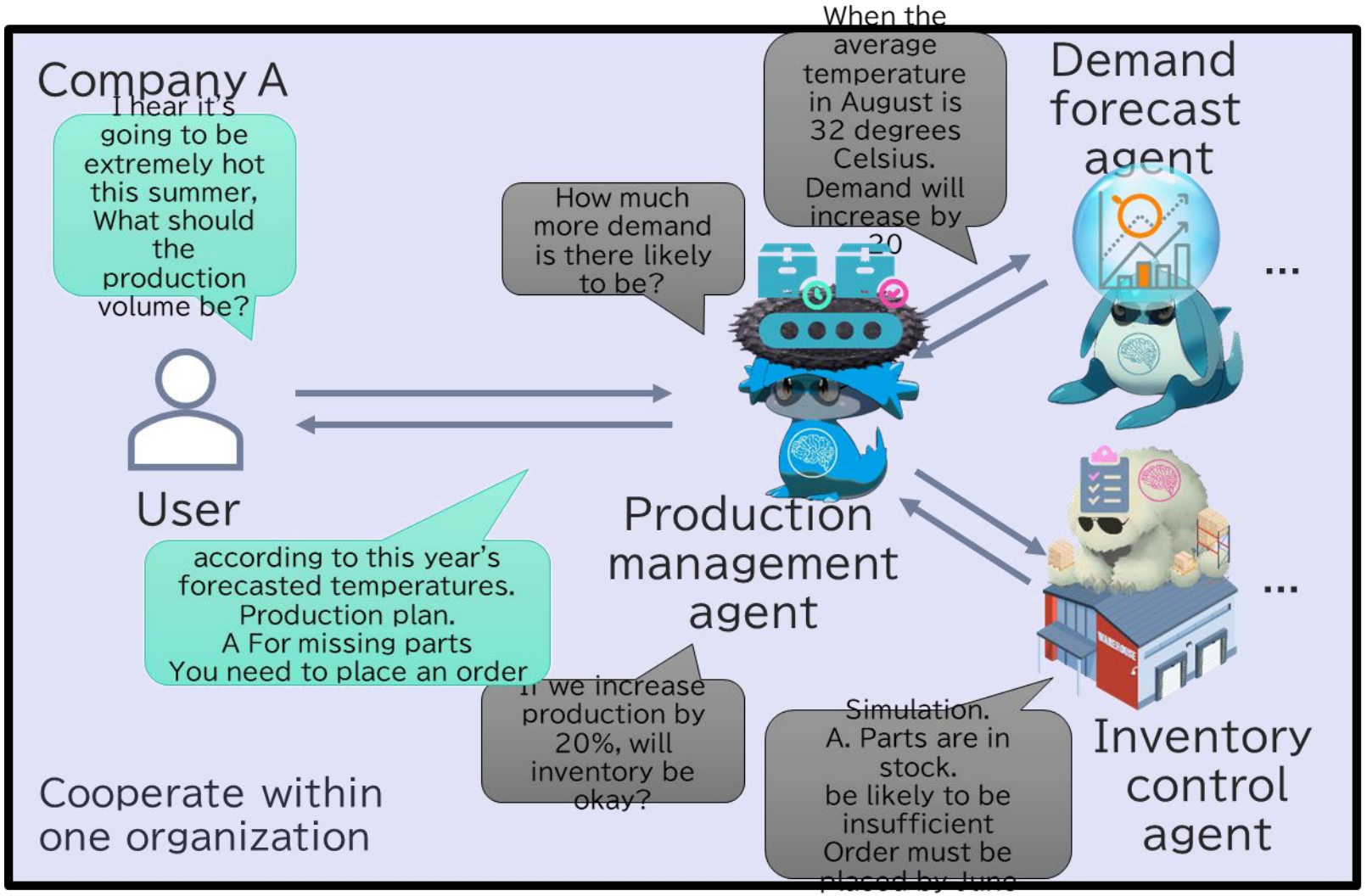
- Convenient to use
- Reusable Units
- Easy to connect
- Simple Business models
- Easy coexistence with Humans
- High Scalability, etc.

Why LLM/MMM?

- By integrating various functions by LLM/MMM,
- Specific solutions can be developed with only vague instructions.
- It will be able to evolve autonomously.

生成AI技術の進展による、業務用エージェントの実用化：Phase 2

社内では複数のエージェントが稼働し、各自の目的達成のためにきちんと協調・連携動作をする



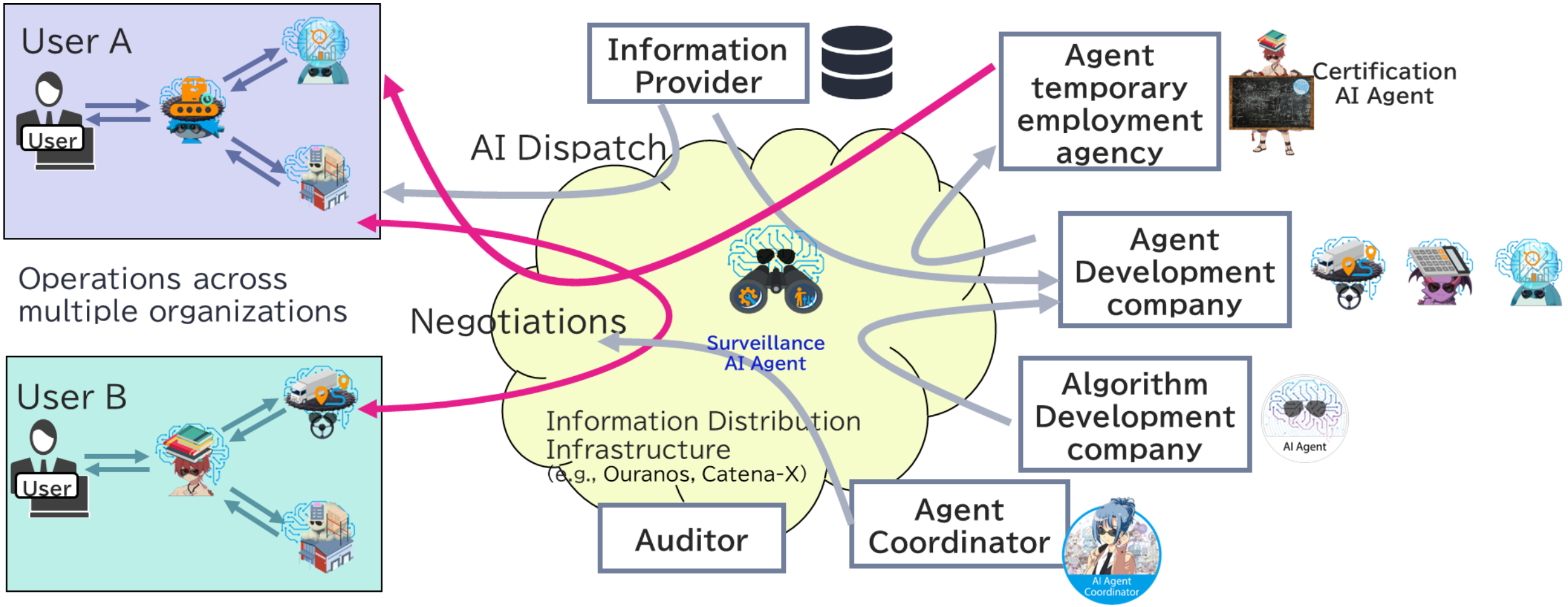
Thanks to the LLM Integration, detailed coordination between systems and the exchange of necessary information can be done autonomously.

No need to spend man-hours on system integration with conventional SI.
Inter-system integration that used to take time in conventional SI is no longer necessary.

Easily tackle those complicated business tasks!

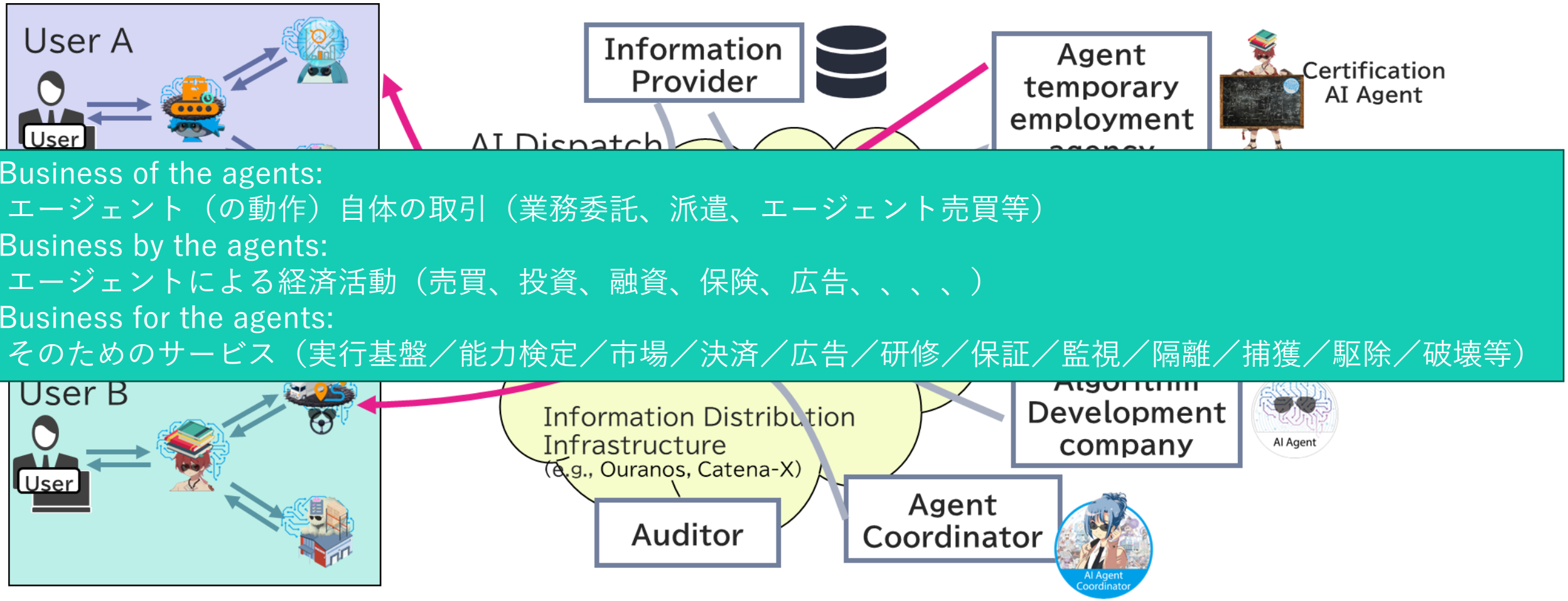
生成AI技術の進展による、業務用エージェントの実用化：Phase 3

各自の目的達成のためには、社外のエージェントとも交渉・調整を行い、適切な合意を形成する



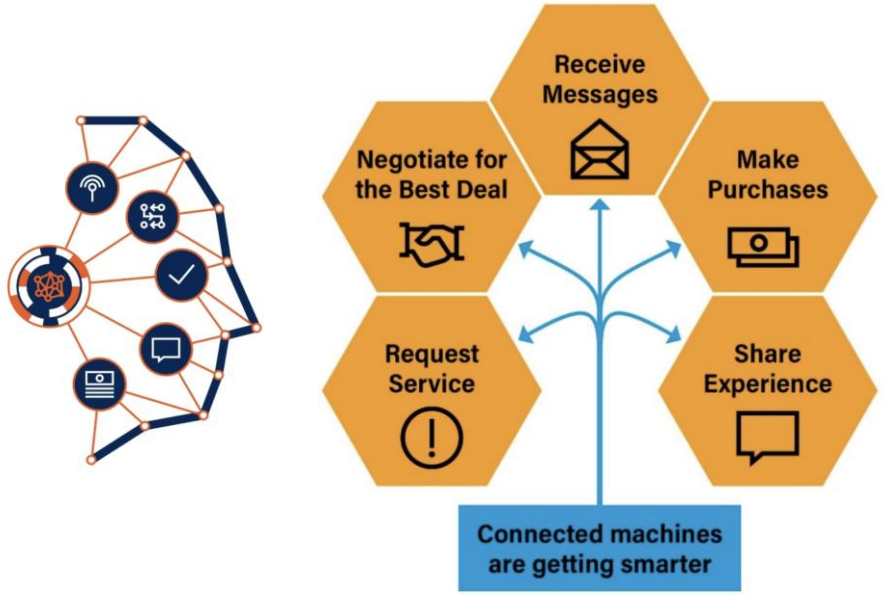
X-as-an-Agent (XaaA)経済圏の形成と進展へ

Organic Ecosystem of Business of the agents, by the agents, for the agents



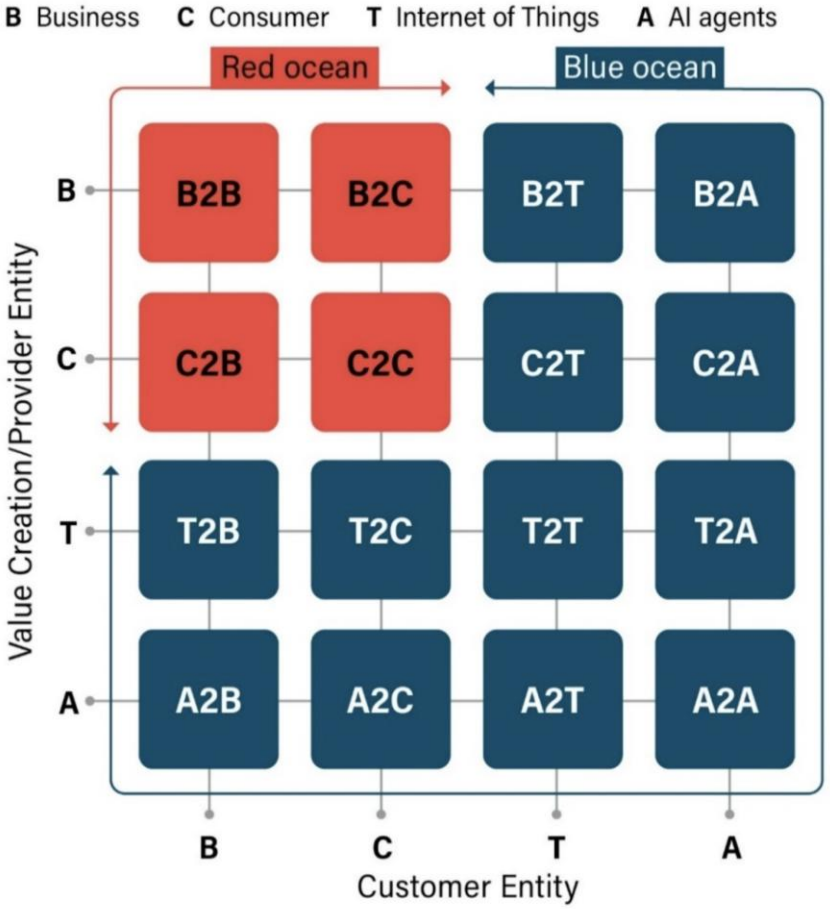
WHEN MACHINES BECOME CUSTOMERS

Some of the Activities that a Machine Customer can Undertake



Trillion Dollars, 2030!
(Hundreds of trillion yen)

Categories of IoT & AI Assistant Extended Business Model Space



AIエージェント・エコノミーの到来と未来展望

参考: NRI、特集　203X：AIで「拡張」する社会拡張する経済、AIエージェント・エコノミーの到来

テクノロジーによる経済の進化

- 従来 → デジタル → プラットフォーム → AIエージェント
- 生成AIが経済構造を変革する

主な特徴

- 経済の重点が「規模」から「深度」へと移行
- AIによる労働力のスケーリング、および多数の顧客への個別サービス提供の実現

経済・産業構造の変化

- AIエージェント市場の顕著な拡大
- 医療、教育、金融サービスの民主化
- 生産性の向上と労働力不足の緩和

課題と対策

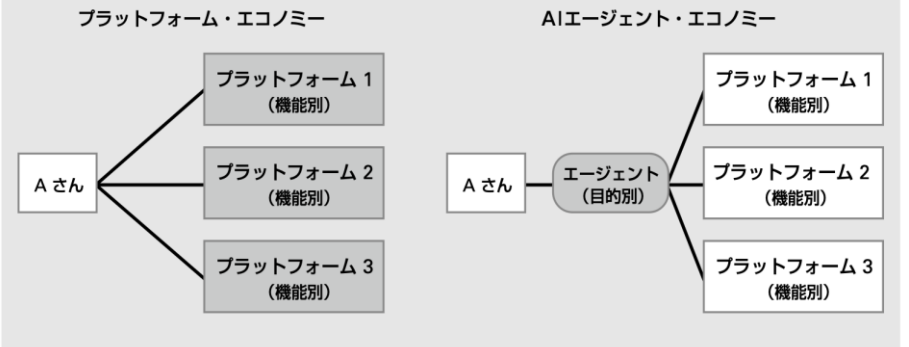
- 経済価値と社会価値の両立を目指したバイデザインアプローチ
- 評価基準の確立、プライバシーおよびセキュリティの整備、標準プロトコルの策定

結論・提言

- AIエージェント経済による少子高齢化問題の克服
- AIエージェント経済による日本の持続可能な経済成長の実現を目指すべき

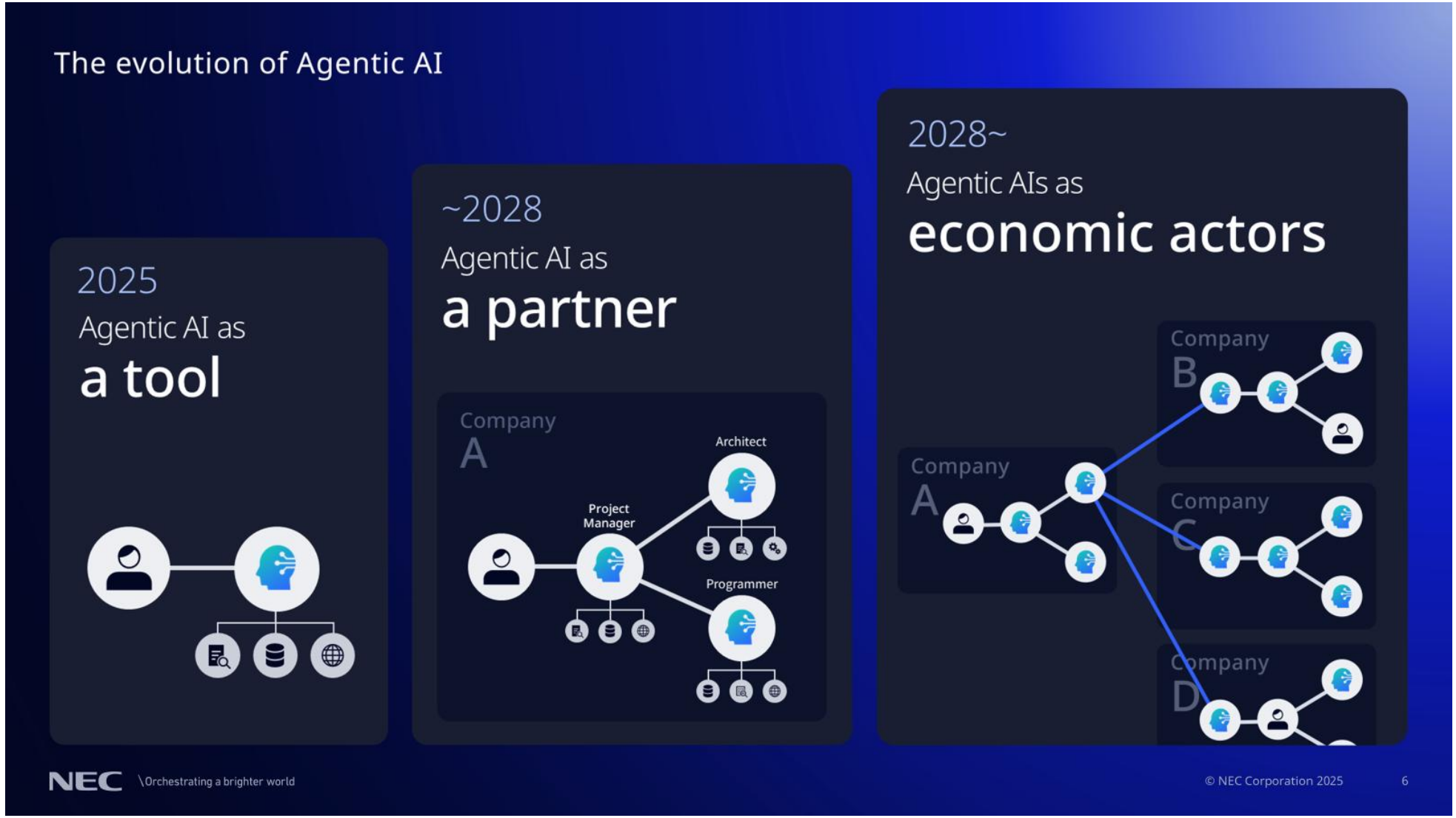
図1　デジタルテクノロジーによる経済の拡張

図2　プラットフォーム・エコノミーとAIエージェント・エコノミーの比較（ユーザーの視点）



経済主体としてのエージェンティックAI

Motoo Nishihara, NEC CTO, MWC2025



エージェント経済圏の勃興
- AIセーフティとAI品質に関する課題 -

生成AI技術の進展による、業務用エージェントの実用化

- Phase 1：単なるツールやデータ、デバイスが、部下・同僚・代理人へ
- Phase 2：社内でのエージェント間の協調・連携
- Phase 3：会社の境界を越えたエージェント間の調整・交渉

エージェント経済圏（X-as-an-Agent Economy）の形成・進展へ→数百兆円@2030

エージェントの、エージェントによる、エージェントのための、ビジネス

- エージェント（の動作）自体の取引（業務委託、派遣、エージェント売買等）
- エージェントによる経済活動（売買、投資、融資、保険、広告、、、、）
- そのためのサービス（実行基盤／能力検定／研修／保証／監視／隔離／捕獲／駆除／破壊等）

NECの取り組み

- **自動交渉エージェントによる社会価値提供「内部最適化・個別意思決定の限界の突破」**
- **国内外での業界団体（DTC、自律調整SCMコンソーシアム）における当該コンセプト牽引活動**
- **自動交渉プロトコルの国際標準化（UN/CEFACT “eNagotiation”）、国際技術コンペ（ANAC SCMリーグ）の主催**
- **日本学術会議での提言、電子情報通信学会での委員会設立**
- **産総研や他社（非開示）との共同研究、CATENA-Xとの意見交換、COCNからの政策提言**

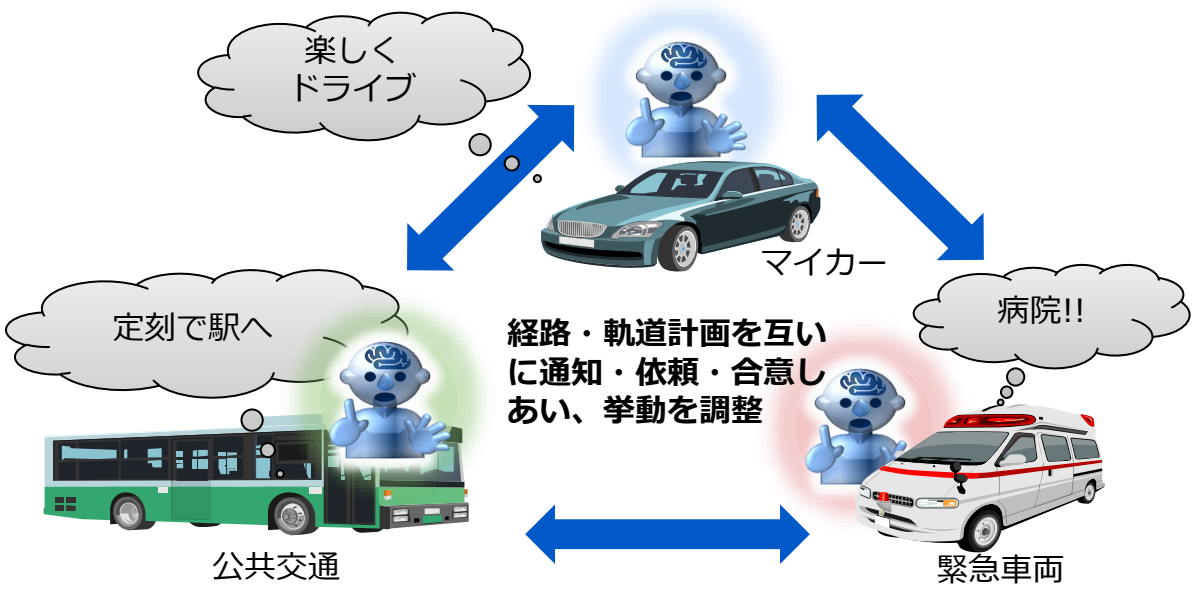
エージェント経済圏におけるAIセーフティとAI品質に関する課題

- ユーザにとっての課題、ベンダーにとっての課題、社会としての課題
- 今後、求められる取り組み

NECの提供する顧客価値・社会価値
「内部最適化・個別意思決定の限界を自動交渉エージェントで突破」

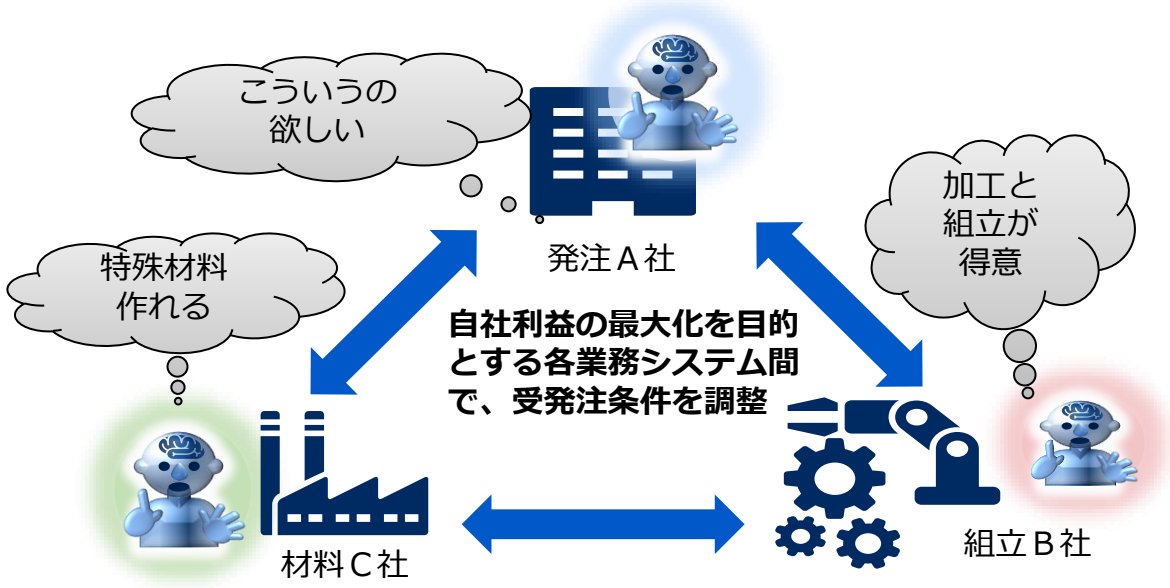
- スマート化されたシステムが広く普及し十分に社会価値を発揮するためには、それらの間の挙動や利害を調整する機能がキーになる　→　内部最適化・個別意思決定の限界を突破
- 既存のデータ共有や協調制御のアーキテクチャは、参加者の内部情報開示や自己決定権剥奪を前提としている　→　経済主体間の調整では、相談や交渉ベースのしくみが必要/非常に有益

移動体間での経路・軌道調整



それぞれが目的を円滑に達成

企業間での受発注条件調整



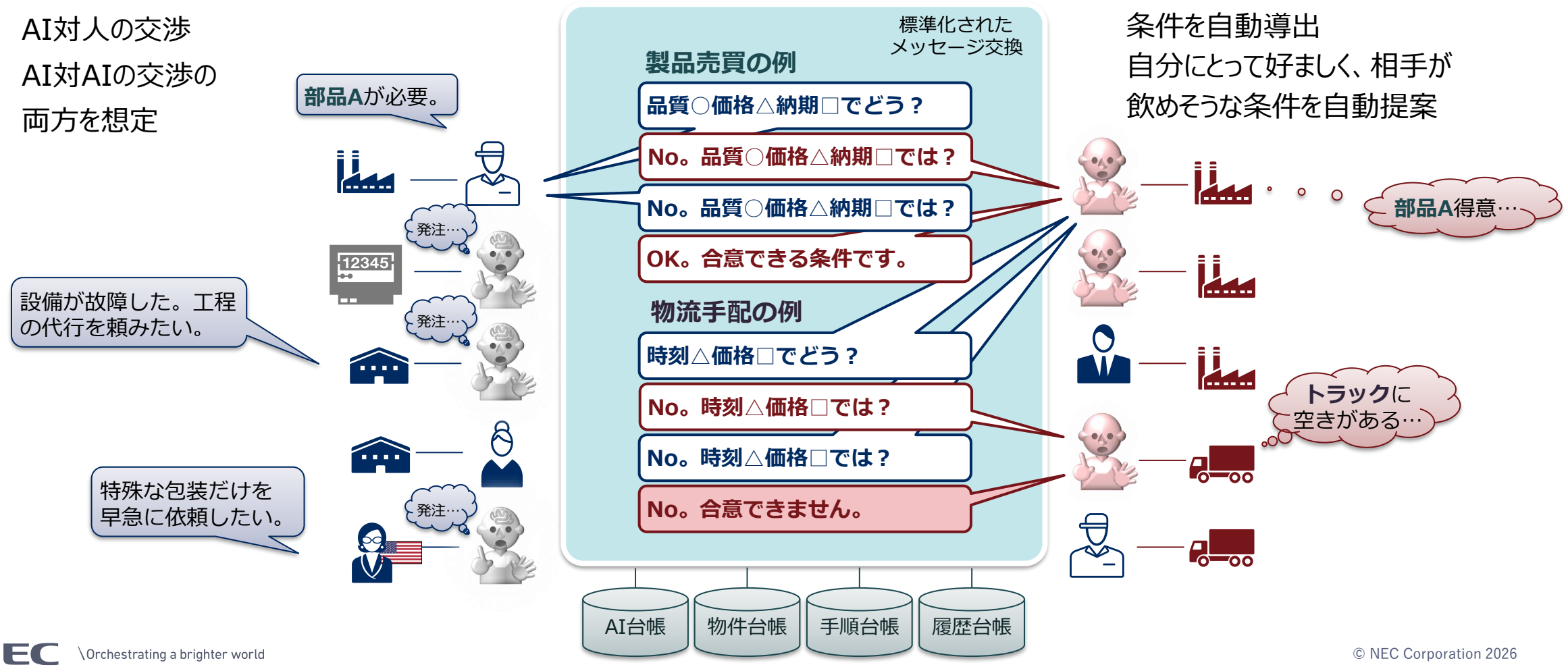
互恵関係を発見／最適化

例：企業間での取引条件の調整

人に代わって自社のために取引先と相談・交渉
→できるだけ良い条件で商談をまとめる／ダメなものは断る

AI対人の交渉
AI対AIの交渉の
両方を想定

譲れない条件、とれると嬉しい
条件を自動導出
自分にとって好ましく、相手が
飲めそうな条件を自動提案

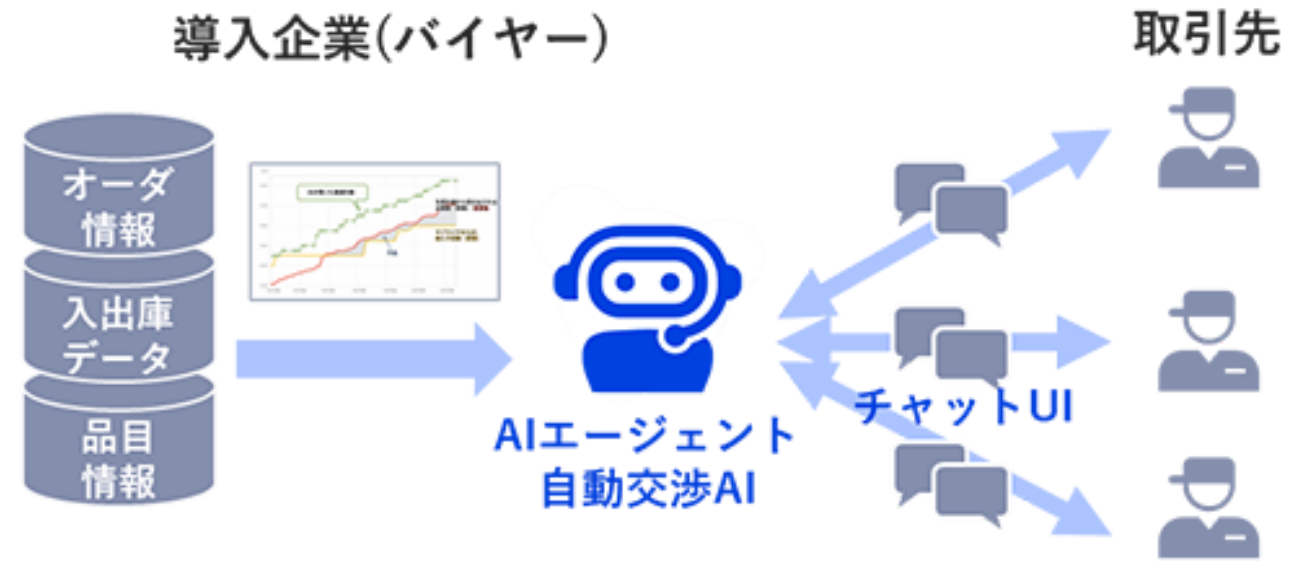


御参考：導入試験の結果

		通常の人間と人間の間での交渉		バイサイドに交渉AIを導入	
		合意までの時間	自動合意率	合意までの時間	自動合意率
セルサイドからの交渉開始要求	周辺機器	3 hours - 2 days	0	33 sec.	100% (3/3)
	電子部品	3 hours - 2 days	0	4 minutes 49 seconds	85.71% (18/21)
バイサイドからの交渉開始要求	周辺機器	3 hours - 2 days	0	1 minute 53 seconds	80.95% (17/21)
	電子部品	3 hours - 2 days	0	38 sec.	100% (95/95)
whole		3 hours - 2 days	0	**1 minute 17 seconds**	**95%** (133/140)

https://jpn.nec.com/press/202512/20251202_01.html

2025年12月サービス提供開始



交渉案の生成

- 納期・数量調整が必要なオーダを自動検出
- 必要に応じて納期・数量の交渉案を自動生成

取引相手との交渉実施

- AIエージェントを介し取引先に交渉案を提示
- 取引先との交渉をインタラクティブに実行

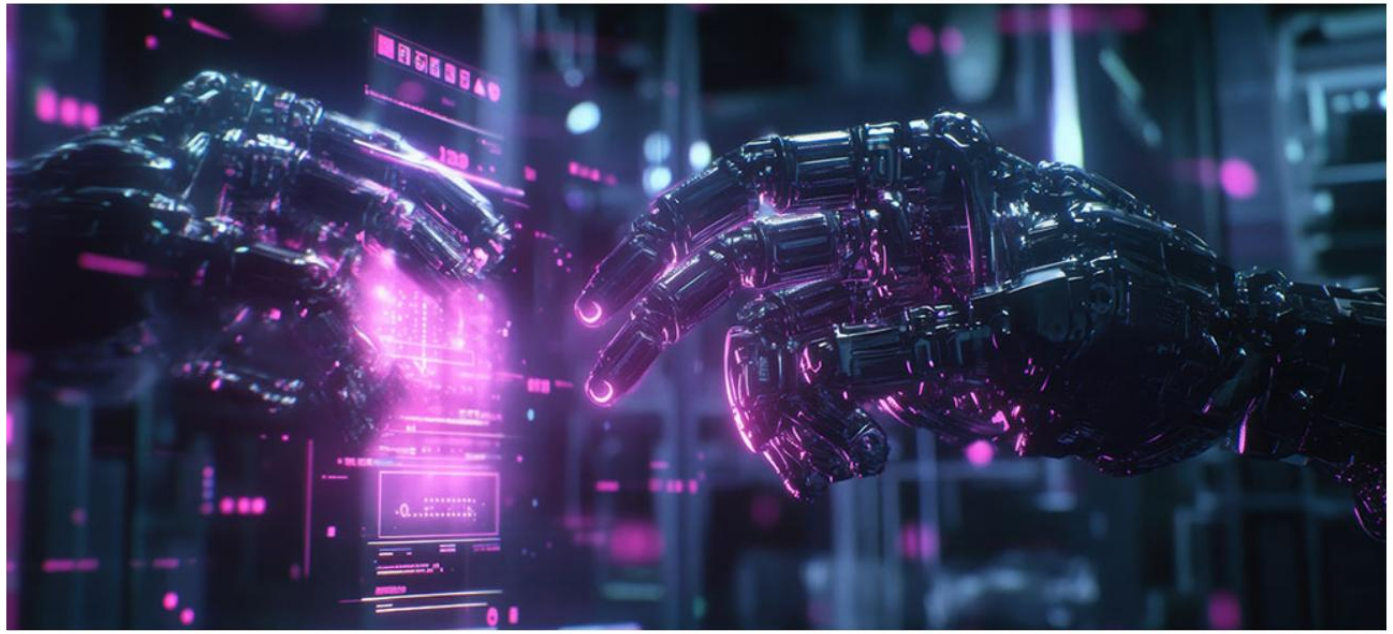
NEC 調達交渉AIエージェントサービス

NECの取り組み：国際業界団体でのテストベッド活動

自動交渉とデジタルツインとマルチエージェントの融合プラットフォーム

Digital Twin Consortium で公式テストベッド NEGOTIATE として採択(2025年7月）

→色々なユースケースを募集中



Solving Cross-Boundary Coordination Through Intelligent Digital Negotiation

Digital twins and multi-agent generative systems (MAGS) securely automate negotiate agreements across organizational boundaries.

The testbed contributes to industry advancement in the following ways:

1.Coordination-Within-Competition:

2.Utility Evaluation:

3.Decision-Making Flexibility:

4.Resource Allocation:

5.Institutional Policy Adherence:

Member, Lead Developers

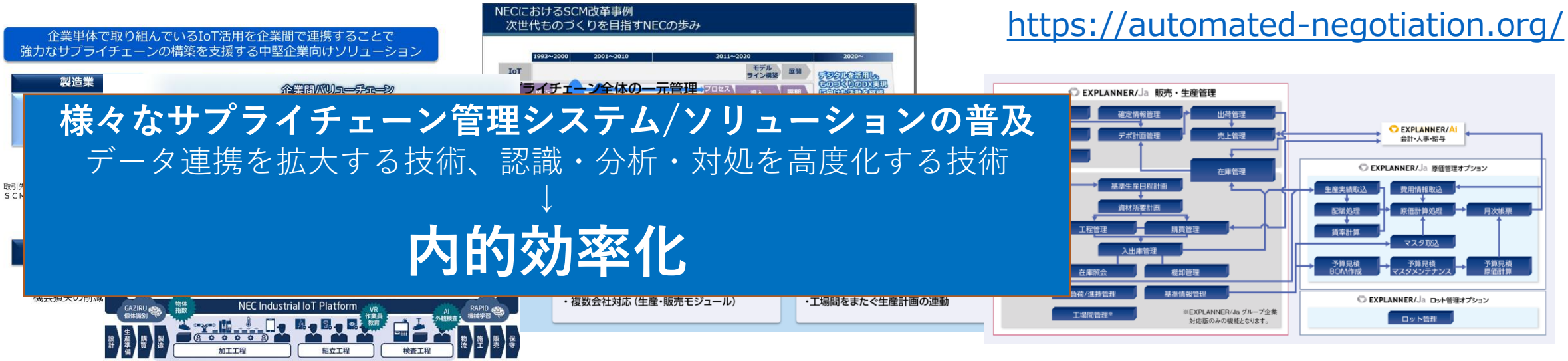
NEC \Orchestrating a brighter world　XMPRO

https://www.digitaltwinconsortium.org/initiatives/digital-twin-testbeds/automated-negotiation-digital-twins-mags/

NECの取り組み：自律調整SCMコンソーシアム設立・運営

Automated Negotiation SCM Consortium

https://automated-negotiation.org/



様々なサプライチェーン管理システム/ソリューションの普及
データ連携を拡大する技術、認識・分析・対処を高度化する技術
↓
内的効率化

✖

自動交渉システム/ソリューションの発展
取引相手との相談・交渉を支援・自動化する技術
↓
外部との調整の効率化

＝

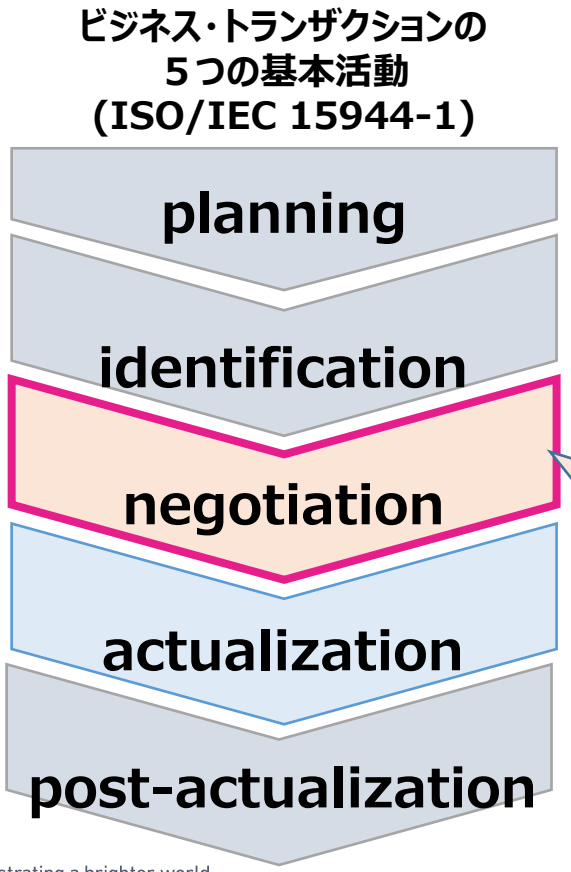
サプライチェーンの劇的な効率化へ

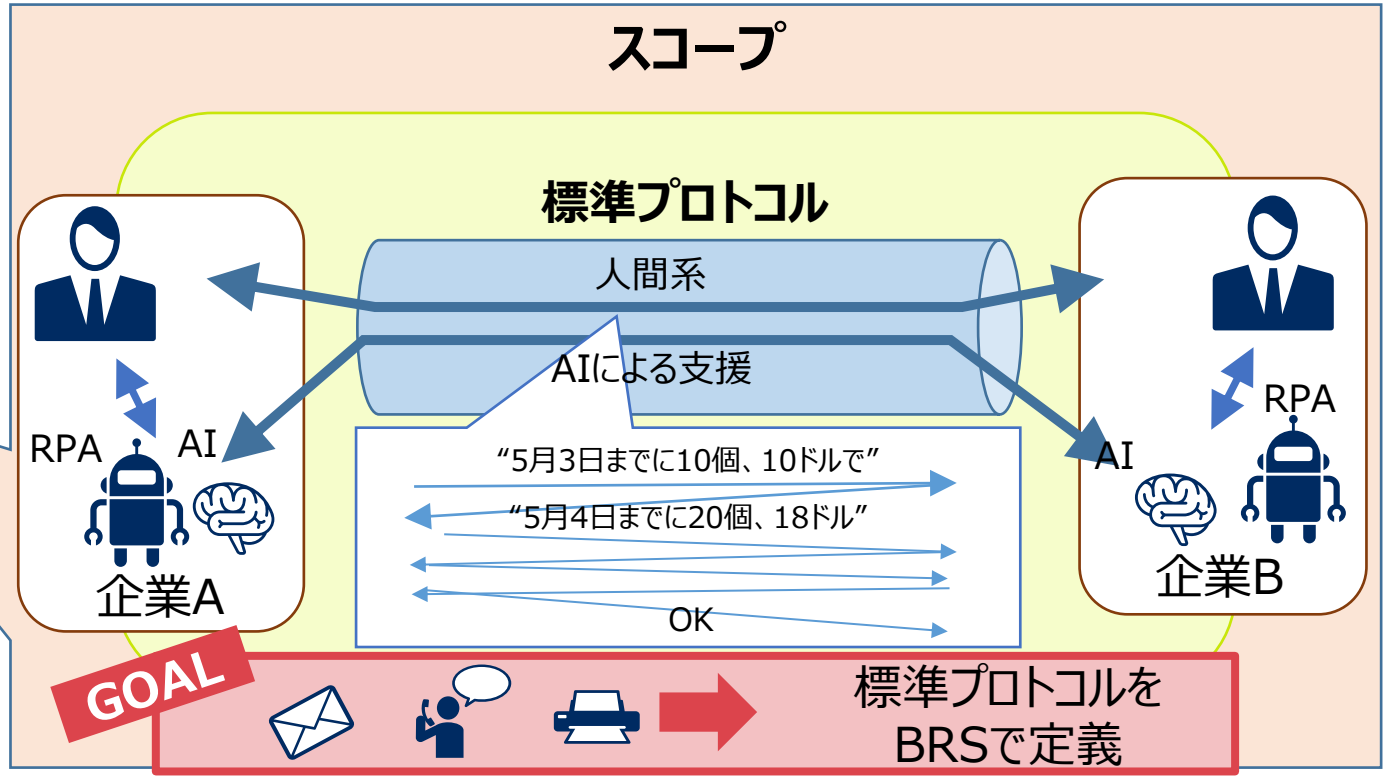
そのための技術課題の解決は進んできた。業務面の課題を解決し社会価値を実現する。
→実用的な調整業務フローへの組み込み方＝意思決定モデルを整理し普及させていく
　商習慣や関連業務、自己決定権の確保、内部情報の非開示、、、等の勘案が必要

NECの取り組み：自動交渉プロトコル国際標準化（非エージェンティック版）

国連の標準化団体（UN/CEFACT）に eNegotiation PJ を提案し採択　→24年標準公開
https://uncefact.unece.org/display/uncefactpublic/E+Negotiation

- 現在はメール・電話・FAXなどで実施されている交渉を、EDI（電子データ交換）化するための標準策定
- 当該標準に従ったEDIを通して、人・RPA・AIによる交渉トランザクションが、独立に開発されたシステムの間でも相互接続／相互運用可能になる。

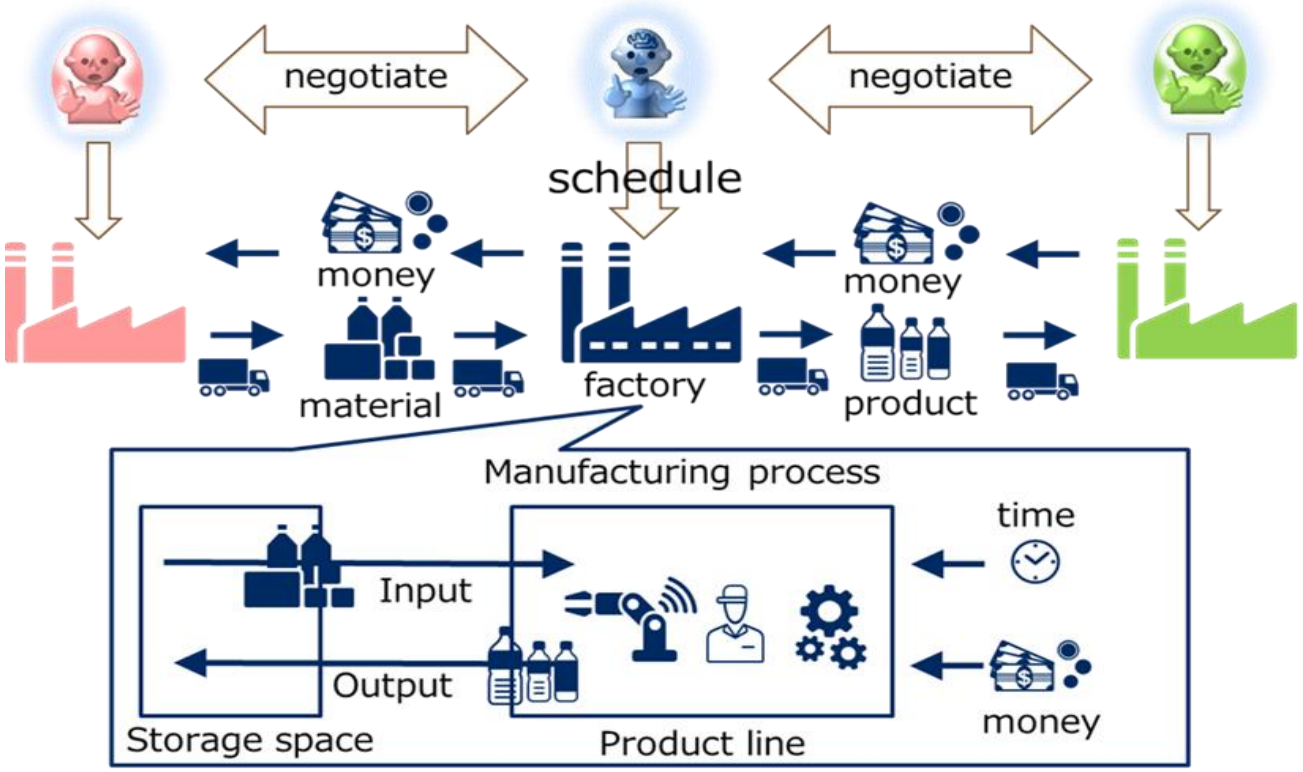




NECの取り組み：自動交渉の国際競技会でSCMリーグを主催

2019年から Automated Negotiation Agent Competition で、交渉技術を競うSCMリーグを年次開催
出場者は製造業者の交渉AIを提出。仮想経済空間内で、AI同士で交渉による部材/製品の売買、
それらを用いた自社工場での製造計画立案を行い、一定期間後に最大利益を獲得したAIが勝者。

参加チームは8チーム（2019年）→22チーム→51チーム→76チーム（2022年）と**年々増加**
２０２４年は７２チーム



Rank	Agent	Earned money
2		
3		
4		

当該リーグのポータルページ
https://scml.cs.brown.edu/

NECの取り組み：アカデミアへの提言・学会での方向性主導

- **日本学術会議　未来の学術振興構想(2023年版)　グランドビジョン⑧-61**
  「AI/人間共存社会における新しいコミュニケーションパラダイムの実現」執筆（今年度改版予定）

  https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-25-t353-3-61.pdf

  提案者：川添 雄彦（（一社）電子情報通信学会）、森永 聡（日本電気株式会社）

  人間社会が長年培ってきた「コミュニケーション」を
  より高次にAI-人間、AI-AI間で実行するための技術・規範・ルールが必要
  → 通信保障、セキュリティ、相互交渉、倫理・法制度などのレイヤーで構成される
  新しいコミュニケーションスタック

- **電子情報通信学会　企業イニシアティブ委員会**
  「AIが相互運用される社会システム分科会」設立

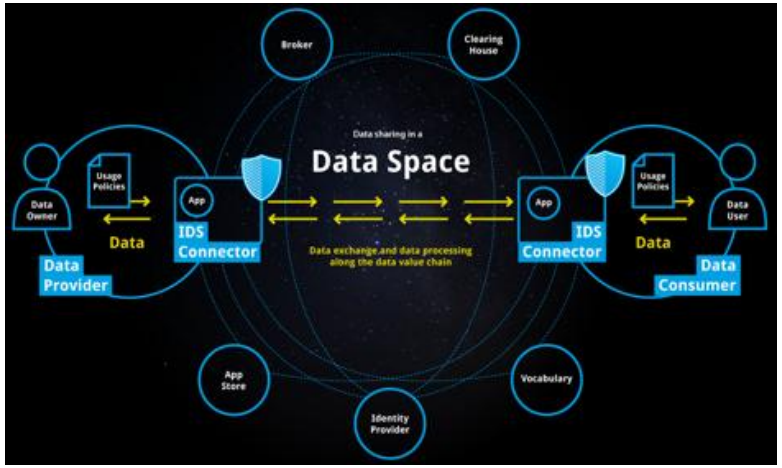
  https://www.ieice.org/jpn_r/activities/kigyo-initiatives/ai/

  委員長：森永　聡（日本電気株式会社）

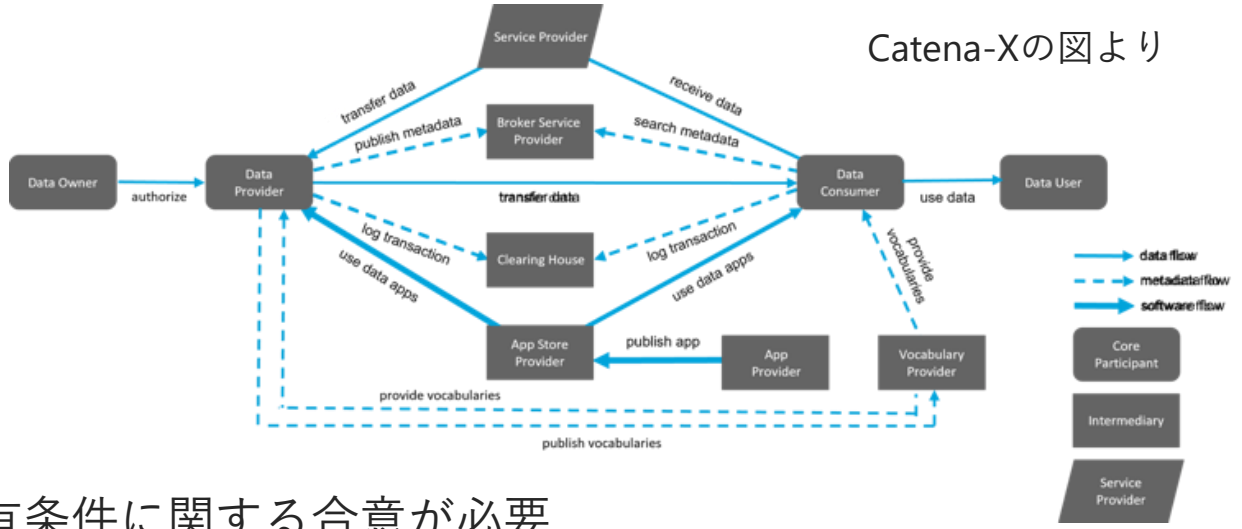
  本分野の課題抽出、解決方法について様々な分野の専門家の協力のもと、下記に取り組んでいく
  - 関連技術・知見の創造と共有
  - 解決のためのエコシステムの組成
  - 社会実装の推進（標準化を含む）

NECの取り組み：関連共同研究、Catena-X等との対話

データ共有条件に関する Data Provider/Consumer 間での合意形成



Catena-Xの図より



- データの出し手と受け手の間では、データ共有条件に関する合意が必要
  - どのデータ、サンプルサイズ、納期、使用期限、使用目的、対価、持ち出し可否、複製可否、再販可否、、、、
  - 例えば、Catena-X では IDSAが定めた交渉手順に従うことが標準とされている
- 現状は人力対応で高コスト。データスペース活用拡大の重大なボトルネックになっている。

↓

自動交渉技術による合意形成の超効率化、データスペース活用の劇的な拡大へ

データサンプル一個でも、人手を介さず、一瞬で、きめ細かい条件まですり合わせ

より良い条件で、ダイナミックに、データ取得／収益化

データスペース上での X-as-an-Agent 経済圏形成

データだったら、

「あなたの保有しているデータベースの中の＊＊をください」
「＊月＊日に＊＊＊に行って、＊＊＊＊の画像をとってきてください」
「来年１年間、＊＊＊のセンシングを行って送ってください」
「このデータに対して、＊＊＊処理をした結果を送ってください」
「＊＊＊の条件で、＊＊シミュレーション／実験した結果を送ってください」
「＊＊＊で学習させたＡＩモデルを送ってください」
「＊＊に関して＊＊を＊＊ＡＩモデルに生成させた結果を送ってください」

、、、

データじゃなくても、データスペース上で同じように取引条件を自動交渉することも可能

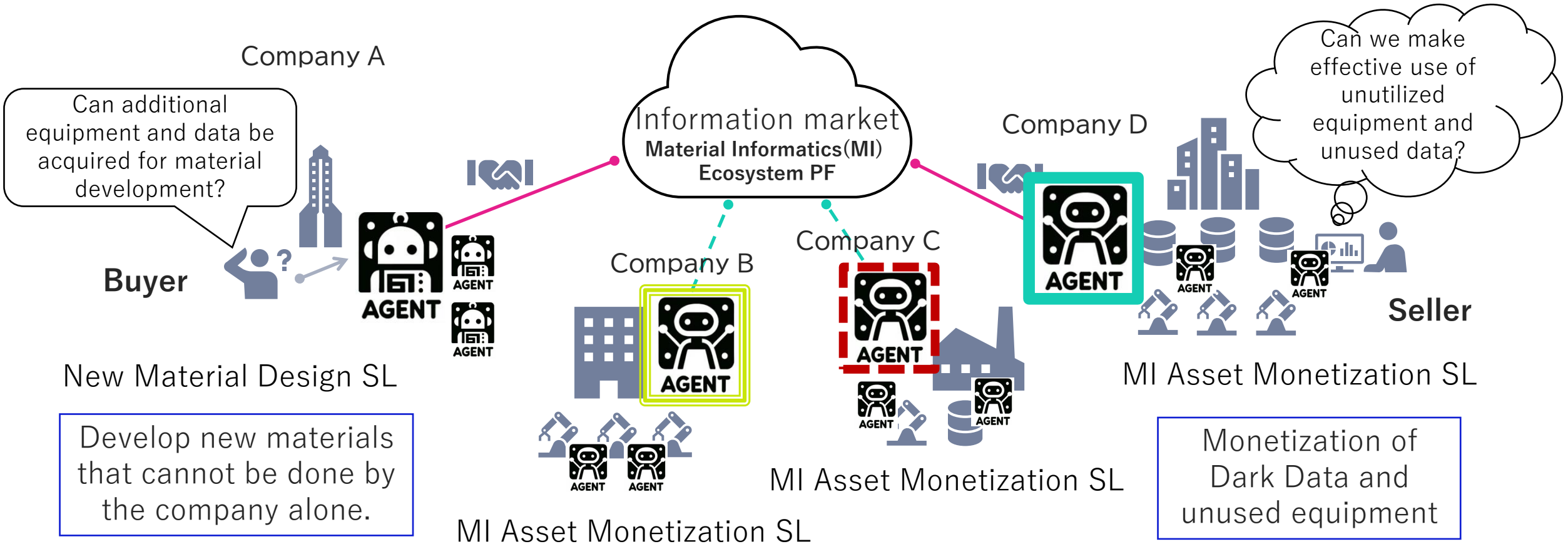
業務プロセスが何もかもＡＩエージェント化していくトレンドと掛け合わせると、、、、、、、、、、、

→　データスペースはX-as-an-Agent (XaaA) 経済圏の最有力培地

＃前述の自律調整SCMコンソでも「情報流通WG」が活動開始

ユースケース例：MI情報コンソーシアム

- Building trading agreements between material developers and material data holders, simulation servicers, experiment contractors…



エージェント経済圏の勃興
- AIセーフティとAI品質に関する課題 -

生成AI技術の進展による、業務用エージェントの実用化

- Phase 1：単なるツールやデータ、デバイスが、部下・同僚・代理人へ
- Phase 2：社内でのエージェント間の協調・連携
- Phase 3：会社の境界を越えたエージェント間の調整・交渉

エージェント経済圏（X-as-an-Agent Economy）の形成・進展へ→数百兆円@2030

エージェントの、エージェントによる、エージェントのための、ビジネス

- エージェント（の動作）自体の取引（業務委託、派遣、エージェント売買等）
- エージェントによる経済活動（売買、投資、融資、保険、広告、、、、）
- そのためのサービス（実行基盤／能力検定／研修／保証／監視／隔離／捕獲／駆除／破壊等）

NECの取り組み

- 自動交渉エージェントによる社会価値提供「内部最適化・個別意思決定の限界の突破」
- 国内外での業界団体（DTC、自律調整SCMコンソーシアム）における当該コンセプト牽引活動
- 自動交渉プロトコルの国際標準化（UN/CEFACT “eNagotiation”）、国際技術コンペ（ANAC SCMリーグ）の主催
- 日本学術会議での提言、電子情報通信学会での委員会設立
- 産総研や他社（非開示）との共同研究、CATENA-Xとの意見交換、COCNからの政策提言

エージェント経済圏におけるAIセーフティとAI品質に関する課題

- **ユーザにとっての課題、ベンダーにとっての課題、社会としての課題**
- **今後、求められる取り組み**

AIセーフティとAI品質：ユーザーにとっての課題

- 与えられた仕事を、どれくらいちゃんとできるのか
- 余計なこと、変なことをしないか
- 現場の人間や、他のAIエージェントとうまく連携できるのか
- 社内のシステムをうまく使いこなせるか
- ばかの一つ覚えではなく、仕事の仕様や環境が多少変わっても、ちゃんと対応できるのか
- 渉外や商談で、カウンターパートより賢くふるまえるか
- 上記のような評価は、ユーザーである自分がやらないといけないのか
- 誰かが出している評価結果を信じていいのか
- その評価は、自分の現場でもそのまま通用するのか
- 変なことをし始めたら止めることができるのか
- 全停止ではなく、部分停止や機能限定ができるのか
- うまく動かなかったり、変なことをやったり、特に対外的に迷惑をかけたとき、ユーザーである自分が責任を取るのか
- 変な挙動について、ちゃんと証拠を残せるのか
- 誰が悪いのか、責任分解できるのか
- 変なこと・困ったことが起きた／起き始めたことを検知できるのか
- そういうとき、現場で応急処置できるのか
- 自社の大事な情報が外に漏れていかないか
- 使いこなすのが難しくないか
- 覚えることは、少なければ少ないほどいい
- 自分がこのAIを十分に使いこなせるのかが心配
- ちゃんと保守され、改版・改善され続けるのか
- AIが停止しても、クリティカルな業務を継続できるのか
- 他部門や取引先、他のAIエージェントとの関係で、過度に相手に適応しすぎて、自分や自部門に無理がこないか
- この安全性・品質で、対価に見合っているのか
- AIがやってしまったことについて、上司や関連部門にちゃんと説明・言い訳できるのか

AIセーフティとAI品質：ユーザーにとっての課題

評価それ自体の観点

業務遂行品質
- 与えられた仕事を、期待水準の品質でできるか
- 商談・渉外などで不利にならないか
- 仕様変更や環境変化に対応できるか

協調性・相互作用品質
- 現場の人間と連携できるか
- 他のAIエージェントと干渉・暴走しないか
- 社内システムを正しく使えるか

挙動の安全性・制御可能性
- 余計なこと、変なことをしないか
- 変なことを「し始めた」ことを検知できるか
- 全停止ではなく、部分停止・機能限定ができるか
- 現場で応急処置できるか

情報・セキュリティ面の安全性
- 自社の重要情報が外部に漏れないか

評価自体以外の観点

評価の“主体”の問題
- 評価は、ユーザーが自分でやる必要があるのか
- 誰かの評価結果を、どこまで信じていいのか
- その評価は、自分の現場でも通用するのか

説明責任・責任分解の問題
- AIがやってしまったことを
  上司や関連部門に説明・言い訳できるのか
- 変な挙動の証拠を残せるのか
- 誰が悪かったのか、責任分解できるのか
- 特に対外的に迷惑をかけた場合、
  ユーザーである自分が責任を取るのか

継続運用・レジリエンスの問題
- ちゃんと保守され、改版・改善され続けるのか
- AIが停止しても、クリティカルな業務を継続できるのか

人間側の問題
- 使いこなすのが難しくないか
- 覚えることが多すぎないか
- 自分が十分に使いこなせるのか不安

AIセーフティとAI品質：ユーザーにとっての課題

評価それ自体の観点

業務遂行品質

[×、○] • 与えられた仕事を、期待水準の品質でできるか
[×、△] • 商談・渉外などで不利にならないか
[×、○] • 仕様変更や環境変化に対応できるか

協調性・相互作用品質

[×、△] • 現場の人間と連携できるか
[△、×] • 他のAIエージェントと干渉・暴走しないか
[×、○] • 社内システムを正しく使えるか

挙動の安全性・制御可能性

[○、△] • 余計なこと、変なことをしないか
[△、△] • 変なことを「し始めた」ことを検知できるか
[×、×] • 全停止ではなく、部分停止・機能限定ができるか
[×、×] • 現場で応急処置できるか

情報・セキュリティ面の安全性

[○、△] • 自社の重要情報が外部に漏れないか

凡例：[AISIで明示的取組、AISTで明示的取組]

評価自体以外の観点

評価の“主体”の問題

• 評価は、ユーザーが自分でやる必要があるのか [×、×]
• 誰かの評価結果を、どこまで信じていいのか [×、×]
• その評価は、自分の現場でも通用するのか [×、×]

説明責任・責任分解の問題

• AIがやってしまったことを
上司や関連部門に説明・言い訳できるのか [×、△]
• 変な挙動の証拠を残せるのか [△、△]
• 誰が悪かったのか、責任分解できるのか [×、×]
• 特に対外的に迷惑をかけた場合、
ユーザーである自分が責任を取るのか [×、×]

継続運用・レジリエンスの問題

• ちゃんと保守され、改版・改善され続けるのか [×、△]
• AIが停止しても、クリティカルな業務を継続できるのか [×、×]

人間側の問題

• 使いこなすのが難しくないか [×、△]
• 覚えることが多すぎないか [×、×]
• 自分が十分に使いこなせるのか不安 [×、×]

AIセーフティとAI品質：ベンダーにとっての課題

- 期待される品質水準が、案件ごとに違いすぎる
- どこまでを保証して、どこからを免責にすべきかわからない
- 期待仕様どおりの性能を本当に出せているか
- 商談・交渉など、性能が数値化しづらい領域で品質をどう説明すればいいのか
- 想定外の使われ方・組み合わされ方をどこまで面倒を見るべきか
- 脱獄（**jailbreak**）されて、想定外のことをやらされないか
- 契約していない目的外使用をされないか
- **AI**エージェントが鹵獲・リバースエンジニアリングされて、ノウハウや競争優位を奪われないか
- 他社**AI**やユーザー側の設計ミスで事故が起きたとき、自社の責任になるのか
- 学習やアップデートで挙動が変わったとき、毎回再評価が必要なのか
- ライバルはどうやっているのか
- 安全対策・ログ・監査を厚くすると、性能・**UX**やコストが成立しなくならないか
- 説明可能性を高めることで、知財やノウハウが漏れないか
- 競合製品と、安全性や品質をフェアに比較してもらえるのか
- 使おうとしているサードベンダーの**AI**やエージェントが、ちゃんとつながるのか
- インテグレータとしてサードベンダー品を使う場合、ユーザーとしての課題がすべて自分事として降ってくる
- 事故が起きたとき、誰が矢面に立つのか
- 安全・品質対応をどこまでやれば、ビジネスとして成り立つのか
- 将来の規制を見越して設計すると、今は売れなくならないか
- 監督当局から、理不尽な情報開示命令（競争力毀損・対応コスト過大・不公平）が出てこないか

AIセーフティとAI品質：ベンダーにとっての課題

評価それ自体の観点

性能・業務有効性の品質
- 期待仕様どおりの性能を出せているか
- 案件ごとに異なる品質期待にどう応えるか
- 数値化しづらい業務の品質説明するか

安全性・悪用耐性
- 有害・危険な挙動を抑制できているか
- 脱獄（jailbreak）や目的外使用を防げているか
- 想定外の使われ方・組み合わせへの耐性

更新・学習による挙動変化の管理
- 学習・アップデートによる挙動変化の把握
- 変更のたびに再評価が必要になるのか
- 安全対策と性能・UX・コストのバランス

監査・証跡・説明可能性
- 何が起きたかを後から再現・説明できるか
- 監査・説明要求に耐えられるか
- 説明性向上が知財・ノウハウ流出につながらないか

相互運用性・統合品質
- サードベンダー品との相互接続性
- 組合わせた結果の挙動をどこまで自社が担保するのか

公平な比較・評価可能性
- 競合と安全性・品質をフェアに比較してもらえるか
- 評価軸や比較方法が不透明・恣意的にならないか

知財・競争力の保全（リバースエンジニアリング）
- 鹵獲・解析され、ノウハウや競争優位を奪われないか
- 説明・監査要求が実質的な技術開示にならないか

評価自体以外の観点

責任境界・免責設計
- どこまでを保証し、どこからを免責にすべきか
- 他社AIやユーザー起因の事故でも責任を負うのか
- 事故時に誰が矢面に立つのか

SI／インテグレータ構造の問題
- サードベンダー品を使うと、ユーザーとしての課題がすべて自分事になる
- 評価・説明・責任・停止判断が集中しないか

ビジネス成立性
- 安全・品質対応をどこまでやれば事業として成立するか
- 真面目に対応するほど競争上不利にならないか

制度・規制リスク
- 将来規制を見越した設計が、今の競争力を損なわないか
- 監督当局から不公平・過剰な情報開示を求められないか

AIセーフティとAI品質：ベンダーにとっての課題

評価それ自体の観点

凡例：[AISIで明示的取組、AISTで明示的取組]

[×、○] **性能・業務有効性の品質**
- 期待仕様どおりの性能を出せているか
- 案件ごとに異なる品質期待にどう応えるか
- 数値化しづらい業務の品質説明するか

[○、△] **安全性・悪用耐性**
- 有害・危険な挙動を抑制できているか
- 脱獄（jailbreak）や目的外使用を防げているか
- 想定外の使われ方・組み合わせへの耐性

[×、△] **更新・学習による挙動変化の管理**
- 学習・アップデートによる挙動変化の把握
- 変更のたびに再評価が必要になるのか
- 安全対策と性能・UX・コストのバランス

[△、△] **監査・証跡・説明可能性**
- 何が起きたかを後から再現・説明できるか
- 監査・説明要求に耐えられるか
- 説明性向上が知財・ノウハウ流出につながらないか

[×、△] **相互運用性・統合品質**
- サードベンダー品との相互接続性
- 組合わせた結果の挙動をどこまで自社が担保するのか

[×、×] **公平な比較・評価可能性**
- 競合と安全性・品質をフェアに比較してもらえるか
- 評価軸や比較方法が不透明・恣意的にならないか

知財・競争力の保全（リバースエンジニアリング） [×、×]
- 鹵獲・解析され、ノウハウや競争優位を奪われないか
- 説明・監査要求が実質的な技術開示にならないか

評価自体以外の観点

責任境界・免責設計 [×、×]
- どこまでを保証し、どこからを免責にすべきか
- 他社AIやユーザー起因の事故でも責任を負うのか
- 事故時に誰が矢面に立つのか

SI／インテグレータ構造の問題 [×、×]
- サードベンダー品を使うと、ユーザーとしての課題がすべて自分事になる
- 評価・説明・責任・停止判断が集中しないか

ビジネス成立性 [×、×]
- 安全・品質対応をどこまでやれば事業として成立するか
- 真面目に対応するほど競争上不利にならないか

制度・規制リスク [△、×]
- 将来規制を見越した設計が、今の競争力を損なわないか
- 監督当局から不公平・過剰な情報開示を求められないか

AIセーフティとAI品質：社会にとっての課題

政府・規制当局

- 厳しい規制をしても、結局は規制の緩い外国で AI が稼働するだけではないのか。
- 日本だけ真面目に規制して、国内企業だけが不利にならないか。
- 個別の AI をきちんと規制・監督しても、群れとして不適切な現象が起きないか。
- 一方向に挙動が集中して、渋滞や暴落のような事象が起きないか。
- 群れとしての学習の相互作用で、談合のような違法な挙動を発見して実行し始めないか。
- 誰も悪意を持っていないのに、違法と同等の結果が出ないか。
- 規制に反する AI 等を、本当に効率よく検出できるのか。
- 望ましくない「群れとしての挙動」を、効率よく検出できるのか。
- まだグレーな段階で、何かできるのか。
- 望ましくないと判断する基準を、事前に決められるのか。
- 検出できたとして、誰が介入していいのか。
- どこまで介入していいのか。できるのか。
- 介入のインパクトや副作用が心配だ。
- AI の活動をすべて止めることはできない。
- 善良な利用まで巻き込んでしまわないか。
- 全停止ではなく、部分的に止めたり、制限したりできるのか。
- 悪い AI を逮捕する、隔離する、破壊するといったことが本当にできるのか。
- そもそも、何を「悪い AI」として特定するのか。
- 悪事の証拠を、ちゃんと集められるのか。
- 群れとして起きた現象について、誰の責任なのか特定できるのか。
- 検知できなかったこと自体を、後から責められないか。
- ダンピングなど、他国からの攻撃的な行為を AI エージェント経済圏として検知・防御できるのか。

AIセーフティとAI品質：社会にとっての課題

消費者／市民

- 自分が知らないところで、AI が勝手に判断や交渉をしていないか不安だ。
- 人間相手なら事情を汲んでくれたはずのことを、AI だと容赦なく最適化されていないか。
- AI の判断で不利益を被ったとき、誰に文句を言えばいいのか分からない。
- 「AI が判断しました」と言われた瞬間に、話し合いが終わってしまわないか。
- 知らないうちに、信用スコアのようなもので扱われていないか。
- AI 同士で話が進み、人間には結果だけが通知されるようになっていないか。
- 何が起きたのか説明されても、専門用語ばかりで納得できない。
- 自分にとって不利な結果でも、異議を申し立てる方法が分からない。
- 便利にはなっているが、その分、逃げ場がなくなっていないか。
- 気づかないうちに、別の目的に使われていないか。
- 少数派の不利益が見過ごされないか。
- IT 弱者に不利を押し付けられる社会になりそうだ。
- 社会が豊かになっているはずなのに、その恩恵を受けられない人が出てきそうだ。
- 自分が何が欲しいか、何がしたいか、どんな政治的意思を持つかといったことが、巧妙に誘導されて、自由意思を AI にコントロールされるのではないか。
- ものすごい弱肉強食の社会になって、すこしでも変なことをすると、骨までしゃぶり取られるのではないか。
- それが当たり前の世代は、人を人と思わない経済活動をするようになるのではないか。
- AI が引き起こす、人の行動変容のほうが危険なのではないか。
- 賢いし人間と区別がつかない。本気で詐欺やられたら絶対騙される自信ある。

周辺プレイヤー（実行基盤、認証、決済、、、等の業者）にとってもAIセーフティやAI品質は重大な課題であるが本日は時間の都合で省略

AIセーフティとAI品質：社会にとっての課題

	政府・規制当局	消費者・市民	AISI	AIST
見えない決定	執行できない	納得できない	△	△
責任の空白	処分できない	文句を言えない	×	×
群れの暴走	市場が壊れる	弱肉強食	×	△
介入不能	止められない	逃げられない	×	×
価値判断不在	ルール化不能	排除される	×	×
自由意思侵食	情報戦	操作される	×	×
人間の変質	社会不安	倫理崩壊	×	×
国際競争	規制空洞化	影響不可視	△	×

AIセーフティとAI品質：AISI,AISTでカバーしきれていない課題

	ユーザー	ベンダー	社会（政府・市民）
評価の主体・信頼	誰の評価を信じてよいか	評価軸が恣意的・案件ごと	信頼できる評価主体がない
責任分解・帰属	自分がどこまで責任を負うのか	他社AI・SI構造で責任集中	処分・帰責ができない
説明責任	上司・取引先に説明できない	説明が知財流出につながる	市民が納得できる説明がない
運用中の制御	部分停止・応急処置ができない	更新のたびに再説明・再評価	介入権限・正当性が未定義
群れ・相互作用	他AIに過度適応して歪む	組合せ結果まで保証できない	市場暴走・談合の創発
違法以前の兆候	変だと感じても止められない	グレー利用を止めると不利	グレー段階で何もできない
国際実効性	海外AIとの競争で不利	規制対応コストが不均衡	規制が国外で空洞化
自由意思・誘導	—	—	意思・欲望・政治的判断の操作
包摂・弱者保護	IT弱者が使いこなせない	低収益層は切り捨てがち	排除・分断が拡大
人間側の変質	—	—	倫理・行動規範の変容
詐欺・なりすまし	騙されたとき防げない	責任を負わされかねない	見分け不能社会への不安

AISI / AIST は
「AIが安全か／品質が高いか」は扱うが、「ユーザーが責任を負えるか」は扱っていない。
「安全に作れているか」は見るが、「安全に売れるか／責任を負えるか」は見ていない。
「危険なAIを作らない」が中心で、「危険になった社会をどう統治するか」は対象外。
「人がAIに騙されないか」は一部扱うが、「社会がAIに慣れすぎること」は扱っていない。

AIセーフティとAI品質：取り組むべき課題

XaaA経済圏の健全な勃興の大前提
→AI の安全性・品質に関する標準の制度化と、第三者機関による検査・認証の仕組み

ベンダー
- 「安全に売れる」ことを前提とした設計・契約
- 組み合わされる前提での品質開示
- 真面目にやるほど損をしない市場づくりへの関与

消費者・市民
- 「知らないうちに決まる」ことへの関心と声
- IT弱者・少数派の視点を社会に残す
- 人間側の行動変容への自覚

ユーザー
- 「自社で引き受ける責任」と「外に出す責任」の切り分け
- 現場で使える停止・縮退ルールの整備
- AIに合わせすぎない業務・組織設計

政府・当局
- 「違法」以前の兆候を扱う制度設計
- 全停止以外の介入手段の制度化
- 国際的な非対称性を前提にしたルール設計

AISI・AIST
- 単体AI評価から「相互作用・群れ」への拡張
- 技術評価を「社会で使える説明」に変換する枠組み
- 「危険になった後」を前提とした設計指針

エージェント経済圏の勃興
- AIセーフティとAI品質に関する課題 -

生成AI技術の進展による、業務用エージェントの実用化

- Phase 1：単なるツールやデータ、デバイスが、部下・同僚・代理人へ
- Phase 2：社内でのエージェント間の協調・連携
- Phase 3：会社の境界を越えたエージェント間の調整・交渉

エージェント経済圏（X-as-an-Agent Economy）の形成・進展へ→数百兆円@2030

エージェントの、エージェントによる、エージェントのための、ビジネス

- エージェント（の動作）自体の取引（業務委託、派遣、エージェント売買等）
- エージェントによる経済活動（売買、投資、融資、保険、広告、、、、）
- そのためのサービス（実行基盤／能力検定／研修／保証／監視／隔離／捕獲／駆除／破壊等）

NECの取り組み

- 自動交渉エージェントによる社会価値提供「内部最適化・個別意思決定の限界の突破」
- 国内外での業界団体（DTC、自律調整SCMコンソーシアム）における当該コンセプト牽引活動
- 自動交渉プロトコルの国際標準化（UN/CEFACT “eNagotiation”）、国際技術コンペ（ANAC SCMリーグ）の主催
- 日本学術会議での提言、電子情報通信学会での委員会設立
- 産総研や他社（非開示）との共同研究、CATENA-Xとの意見交換、COCNからの政策提言

AIセーフティとAI品質に関する課題

- ユーザにとっての課題、ベンダーにとっての課題、社会としての課題
- 今後、求められる取り組み

\Orchestrating a brighter world