さきがけ
「信頼されるAIの基盤技術」
研究総括説明

2022年4月27日

研究総括　有村　博紀（北海道大学・教授）

科学技術振興機構

戦略目標（概要）

信頼されるAIの基盤技術

- 人工知能(AI)技術を用いたシステムやサービス等のAI技術が社会の中に浸透しつつある。現在，「人間中心のAI社会」*1の実現のために「信頼される高品質AI」*2(Trusted Quality AI) の実現が課題。
- 一方で，深層学習に代表される現在のAI技術がもつさまざまな弱点や限界、社会との整合のために不足する要件が明らかに（脆弱性，予測のブラックボックス性，説明性や納得性，不公平性等）。
- そのため、現在のAIの限界を超えるAI技術そのものを革新的に発展させ、社会からの要請に応え得る根本的な信頼性確保を実現するための研究開発が必要となっている。

(*1) 2019.3 (*2) 2019.6 統合イノベーション戦略推進会議決定

戦略目標（達成目標）

人間中心のAI社会の実現のため「信頼される高品質なAI」の創出に向けた研究開発の推進をはかる

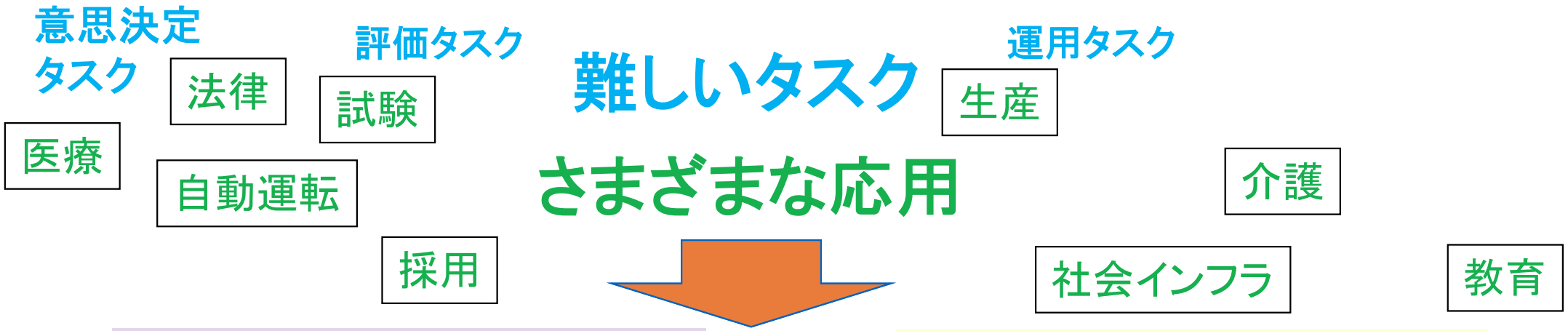
(1) 現在のAI技術の限界を克服する新技術の創出

(2) AIシステムの信頼性・安全性を確保する技術の創出

(3) データの信頼性確保と意思決定・合意形成支援技術の創出

次世代の人工知能技術

基本的な問い：21世紀の社会で、
コンピュータが人間と働いていくために
必要な人工知能技術とは何か？



社会的要請
- 社会的責任
- 人間との協働

- 説明可能性
- 信頼性・有意性
- 公平性・AIシステム

機械学習技術の次の課題!

次世代の人工知能技術

- 現在「AI」としてどちらの方向で議論しているのか

1. 機械学習アルゴリズム

2018年現在、盛んに研究されている

2. “AIシステム”

- 複雑巨大システムへの機械学習アルゴリズムの組込み
- 例：生産プラント、交通インフラ、社会インフラ等

シミュレーション?

制御システム?

3. “汎用AI”

機械学習アルゴリズムを中核として、人間の知能に匹敵する汎用AIの実現を目指す

制御システムの専門家へのインタビュー

- 従来の複雑システムの場合。
- 理論に基づく制御コンポーネントを、非線形のシステムに組み込んで運用する場合がある。
- その場合には、制御理論と現場の経験を総合して、システムを設計し、運用する。

例：The Partnership on AI

PARTNERSHIP ON AI

- 機械学習技術の社会的受容を目指すコンソーシアム。
- 取り組むべき課題として次の6つをあげている

Our Work

- 安全・安心なAI (Safety-Critical AI)
- 公正性、透明性、責任有るAI (Fair, Transparent, and Accountable AI)
- AIと、労働者、経済
- 人々とAIシステムの協調
- 人々と社会へAIが与える影響 (Social and Societal Influences of AI)
- AIと社会利益 (AI and Social Good*)

KDD 2017で同名のWorkshop

2016年創設。

創設メンバー

Amazon, Facebook, Google, DeepMind, Microsoft, IBM, Apple, …

現在、連携先は >50%非営利、13カ国からの、80団体

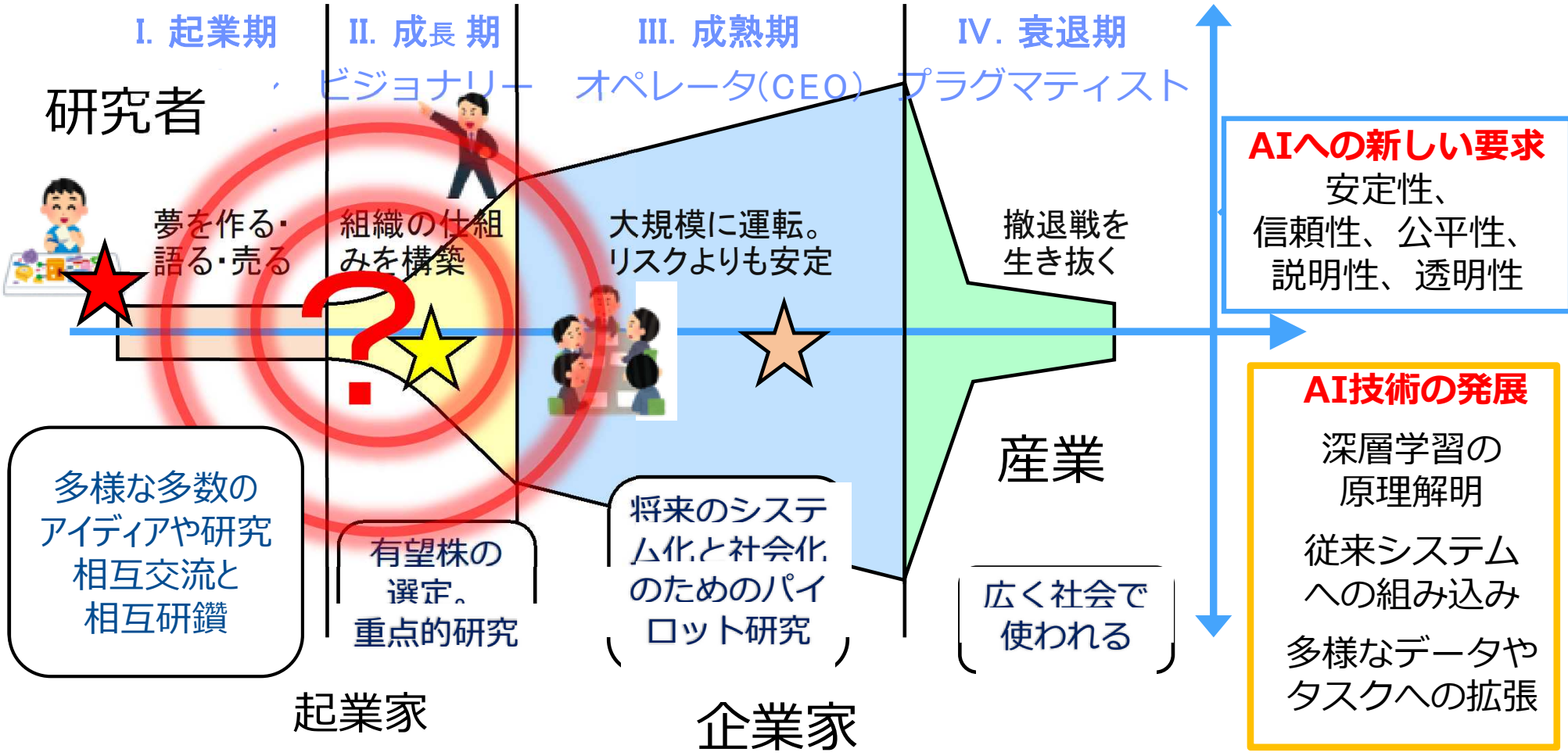
https://www.partnershiponai.org

次世代の人工知能技術は今どこに？

- 現在のAI研究は、発展のどの段階か？

現在の人工知能技術
- 現在は爆発的な発展の変革期

人間中心のAI社会



AIへの新しい要求
安定性、
信頼性、公平性、
説明性、透明性

AI技術の発展
深層学習の原理解明
従来システムへの組み込み
多様なデータやタスクへの拡張

(*) “The four: The Hidden DNA of Amazon, Apple, Facebook, and Google,” Scott Galloway, 2017 (訳 渡会圭子、東洋経済新報社, 2018)

さきがけ次世代IoT研究領域（概要）

領域名：信頼されるAIの基盤技術

「人間中心のAI社会」の実現には、社会における多様で重要なタスクをこなし、人間と共に働く「信頼されるAI」が必要。そのために、従来のAI技術の単なる延長でなく、現在のAI技術/AIシステムがもつ本質的な問題点に取り組み、新たな原理・技術・社会実装の発見・発明を通じて、それらを解決する「信頼されるAI」のための革新的なAI基盤技術の創生を目指す。さらに、社会のさまざまな局面で、重要なタスクを遂行するための「AIシステム」の構成原理と実現技術の確立を目指す。

対象とする技術項目

(1) 現在のAI技術の限界を克服するための新しい数理・計算・解析の基盤技術の研究、(2)AIシステムの信頼性・頑健性・透明性・公平性等、社会における新たなAI応用タスクの概念化・定式化、新しい構成原理・実現技術、(3) AIシステムを支えるデータや情報の信頼性保証や、AI技術を用いた意思決定や合意形成の技術 、(5) AIシステムの設計・開発・運用の方法論等

研究総括方針（1/4）

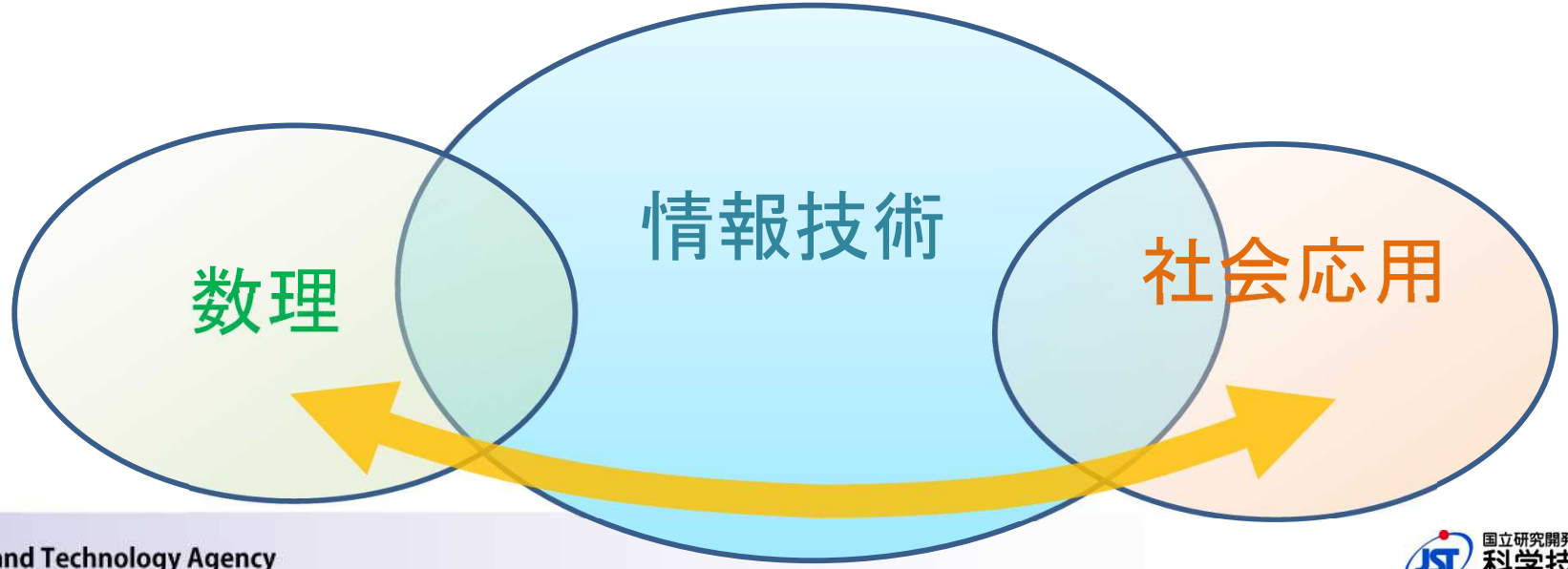
数理・情報技術・社会応用等の多様な立場から、

人間中心のAI社会の実現に資する、

次世代AI技術と周辺技術に関して、

若手研究者による新しい発想に基づいた、

独創的かつ挑戦的な研究を推進する。



研究総括方針（2/4）

提案募集する研究の例

■ 例1： 現在のAI技術の弱点の克服

・現在のAI技術の原理・限界の解明。新しいAIの構成原理。多様なデータやタスクを対象とする新しい数理モデルや、アルゴリズム、システムの提案。

・帰納的情報処理（例：深層学習）と演繹的情報処理（例：記号推論・知識処理・シミュレーション等）を融合したAIの研究。高次の機械学習タスク（因果関係の発見や、反事実説明、統計的有意な予測等）の研究

・人間の脳情報処理や認知に関する知見に基づく新しいAI原理。

これら従来の研究トピックにとどまらず，将来の人間中心のAI社会の基盤技術として、革新をもたらすようなあらゆる研究テーマ．

■ 例2： AIシステム・サービスの信頼性・安全性を確保するための研究

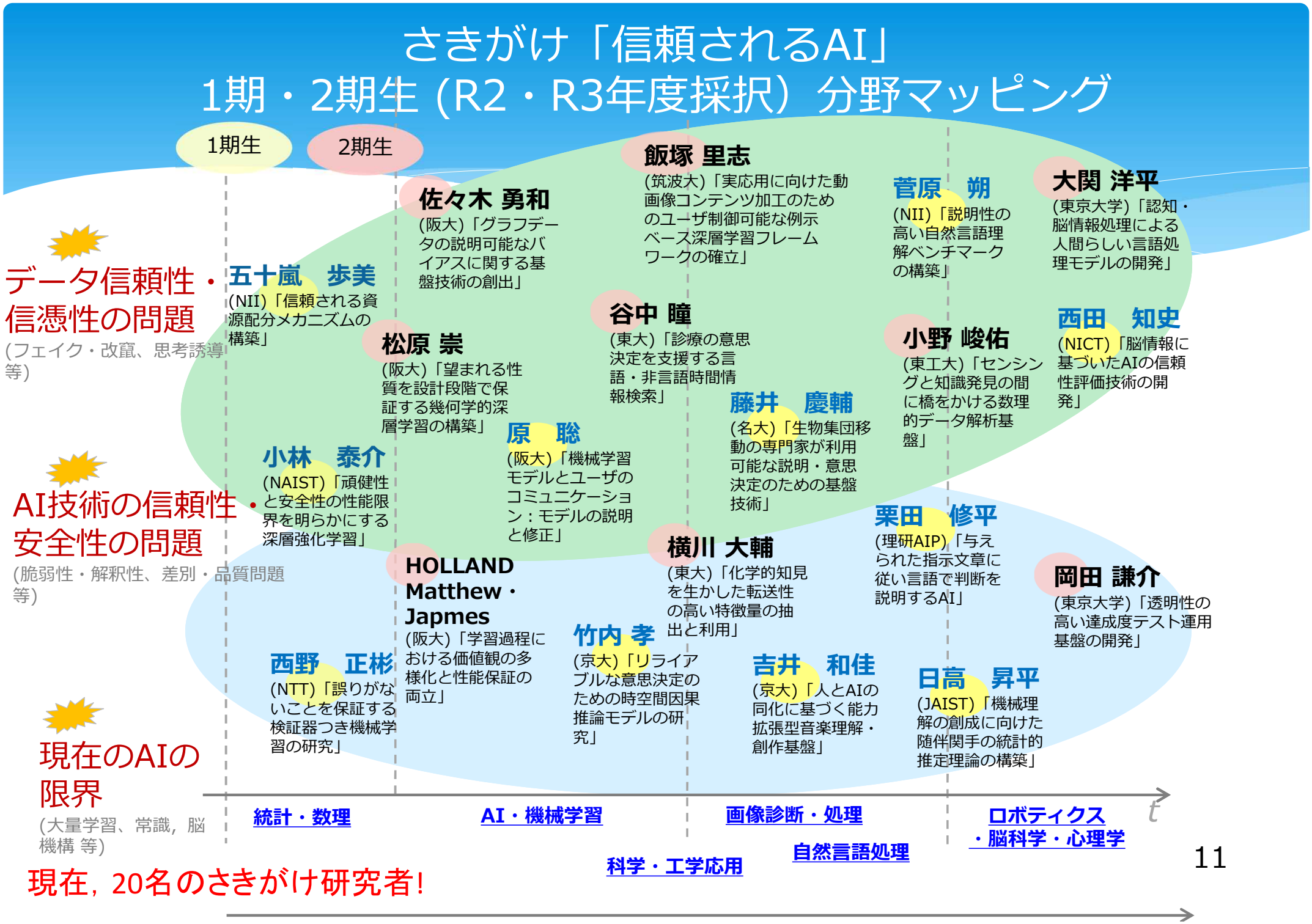
・AI技術を中核とする情報システム（以降、AIシステム）の設計・開発・運用や、品質評価、安全設計、説明性、透明性、公平性等に関わる新しい数理・方法論の研究。

・未知・想定外ケースや環境変化、データのバイアス、悪意ある攻撃等に対して頑健なAIシステム・サービスを実現するための研究

■ 例3：社会におけるAIシステムのための周辺技術

多数の参加者・利害関係者が参加するAI技術を用いたサービスやシステムの数理・アルゴリズム・制度設計等の研究（例：マッチングや、オークション、クラウドソーシング等）。納得性や，公平性，頑健性な等の望ましい性質をもった意思決定や合意形成の数理や技術。フェイク、データ改ざん等の不正（AIによるものを含む）を検知し、対処する技術の研究。

さきがけ「信頼されるAI」
1期・2期生（R2・R3年度採択）分野マッピング



現在，20名のさきがけ研究者!

領域アドバイザー

氏名（敬称略）	所属	役職
石川 冬樹	情報・システム研究機構 国立情報学研究所 アーキテクチャ科学研究系	准教授
宇野 毅明	情報・システム研究機構 国立情報学研究所 情報学プリンシプル研究系	教授
浦本 直彦	（株）三菱ケミカルホールディングス データ＆先端技術部	部長
大野 和則	東北大学 未来科学技術共同研究センター	教授
岡崎 直観	東京工業大学 情報理工学院	教授
鹿島 久嗣	京都大学 情報学研究科	教授
佐久間 淳	筑波大学 システム情報系	教授
櫻井 祐子	名古屋工業大学 大学院工学研究科	教授
佐倉 統	東京大学 大学院情報学環／理化学研究所 革新知能統合研究センター	教授／チームリーダー
谷口 忠大	立命館大学 情報理工学部	教授
長井 志江	東京大学 ニューロインテリジェンス国際研究機構	特任教授
藤吉 弘亘	中部大学 工学部	教授
松井 知子	統計数理研究所 モデリング研究系	教授
持橋 大地	統計数理研究所 数理・推論研究系	准教授
森永 聡	日本電気（株） データサイエンス研究所	上席主席研究員

関連分野（順不同）：人工知能・機械学習（理論・先端技術開発・応用）・数理統計・自然言語処理・画像処理・信号処理・ロボティックス・情報知覚・創発システム・認知科学・神経脳科学・ソフトウェア工学・データマイニング・アルゴリズム・エージェント（メカニズムデザイン）・集合知・科学技術と社会・自然科学応用など（応募分野は上記に限りません）

運営の基本方針

年2回の領域会議を中心に

さきがけは自由な挑戦の場：

- 短期的成果でなく、失敗をおそれず、挑戦的な研究を

さきがけは切磋琢磨の場：

- 未来のリーダーになる研究者同士のネットワークづくり

さきがけは成長の場：

- 領域アドバイザーからの助言（研究構想、計画、キャリアアップ等）。
- 関係する人文・企業・行政等の多様な関係者との交流。
  CREST等の領域との連携。

研究者・アドバイザの有志による研究会（つながり会）なども行っています
（2022年4月 東京，京都 ）

応募の留意点

- 短期的な成果でなく、それぞれの研究分野に大きなインパクトを与えるような成果を目指してほしい
- 将来の「信頼できるAI」のために、どのように現在の問題点・困難を解決し、どのような新しい概念や技術を作り出したいかを、申請者自身の言葉で語ってほしい
- ご自身の目標実現に際して、どのような領域参加者や他領域と、どのような連携をしたいかを自由に記載いただけるとうれしい

研究期間と予算

- 研究期間：3.5年以内
- 研究費の規模： 4,000万円（直接経費）上限
- 採択数： 10件程度 X 3期（令和2、3、4年予定）

詳しくはJSTの募集要項をご覧ください

公募スケジュール

研究提案募集

募集締切 | **5月31日（火）正午**

〆切後は提案を一切受理しませんのでご留意下さい

書類選考会	7月12日(火)
書類選考通過者への連絡期限	7月19日(火)
面接選考会 ※ 具体的な面接日時についてはJSTから指定させていただきます。あらかじめご了承ください。	8月1日(月)、2日(火)
研究開始	10月以降