CREST
「信頼されるAIシステムを支える基盤技術」
研究総括説明

2022年4月27日

研究総括　相澤　彰子

科学技術振興機構

本領域がスタートした背景

戦略目標（概要）

「信頼されるAI」

令和2年度における科学技術振興機構の戦略事業の戦略目標（5件）のうちの1つ

- 様々な形でAI技術が社会の中に浸透しつつある。
- 一方で現在のAI技術について、AIシステムの信頼性・安全性やデータ自体の信ぴょう性に関わる懸念が指摘されている。
- 今後のAIの進化と信頼性確保のための幅広い基盤技術の研究開発が必要

研究領域の背景

AI技術の信頼性や安全性等に関する懸念

- AI技術そのもの、特にその中心技術である深層学習について、出力結果の**説明性や納得性**が不十分、データに含まれている**バイアス**を学習してしまう、**未知・想定外ケースや環境変化**に対してぜい弱である、**文脈や常識**の理解ができていない、等
- AI技術が応用されたシステムやサービスについて、既存のソフトウェア工学等の方法論では**システム全体の信頼性や安全性、品質を保証**することができないため、新たな方法論が必要
- データ自体について、**フェイクの流通や改ざん**の恐れや、これらフェイクの作成・流通や改ざんにAIが悪用されるといった問題

戦略目標（達成目標）

「人間中心のAI社会原則」に基づいた「信頼される高品質なAI」（Trusted Quality AI）の創出に向けた研究開発を推進する。具体的には以下の3つの達成を目指す。

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及び意思決定・合意形成支援技術の創出

CREST
「信頼されるAIシステムを支える基盤技術」研究領域（概要）

領域名（略称）：信頼されるAIシステム

本研究領域は、人間が社会の中で幅広く安心して利用できる「信頼される高品質なAI」の実現につながる基盤技術の創出やそれらを活用したAIシステムの構築を行います。研究にあたっては、人間中心のAIシステムに関する信頼性や安全性等の定義や評価法の検討に取り組み、AIシステム全体としてその要求や要件を満たす技術の確立を目指します。

Towards Our Digital Future

勢い衰えず

2020.3

US:

- DARPA「Explainable Artificial Intelligence (XAI)」2017- (80億円)
- DARPA「Media Forensics」「Semantic Forensics (deep fakes)」
- DARPA「AI Next Campaign」(contextual reasoning) 2018（2億ドル以上）

EU: Trusted AI

- European Union Regulations on Algorithmic Decision Making and a “Right to Explanation” https://arxiv.org/pdf/1606.08813.pdf
- NL4XAI? Horizon2020 https://nl4xai.eu/

China

- Responsible AI

JPN:

- 「人間中心のAI社会原則」（イノベーション戦略推進会議）
- 人工知能学会「人工知能学会　倫理指針」
- NEDO「人工知能の信頼性に関する技術開発」

2021.3

Source: The National Security Commission on Artificial Intelligence, Report on March 1, 2021

“With a remarkable increase of investments in the global AI industry over the past five years and an unprecedented amount of general R&D dollars being invested worldwide, there is no AI slowdown in sight—only new horizons for deployed AI.”

Frontiers of AI Technology.

The next decade of AI research will likely be defined by efforts to incorporate existing knowledge, push forward novel ways of learning, and make systems more robust, generalizable, and trustworthy.[11] Research on advancing human-machine teaming will be at the forefront, as will improvements in hybrid AI techniques, enhanced training methods, and explainable AI.

領域ロゴです

信頼される AI システム

Trusted quality AI systems

領域のタイムライン（公募）

2020公募（1回目） → 第1期スタート（5.5年）

2021公募（2回目） → 第2期スタート（5.5年）

2022公募（3回目）

参加者からみる領域の概要

2020年度(1期)採択課題

氏名	所属機関	役職	研究課題名
伊藤 孝行	京都大学 大学院情報学研究科	教授	ハイパーデモクラシー:ソーシャルマルチエージェントに基づく大規模合意形成プラットフォームの実現
乾 健太郎	東北大学 大学院情報科学研究科	教授	知識と推論に基づいて言語で説明できるAIシステム
越前 功	国立情報学研究所 情報社会相関研究系	教授	インフォデミックを克服するソーシャル情報基盤技術
後藤 真孝	産業技術総合研究所 人間情報インタラクション研究部門	首席研究員	信頼されるExplorable推薦基盤技術の実現
森 健策	名古屋大学 大学院情報学研究科	教授	あいまい性を表現するReliable Interventional AI Robotics

2021年度(2期)採択課題

氏名	所属機関	役職	研究課題名
鹿島 久嗣	京都大学 大学院情報学研究科	教授	人とAIの協働ヒューマンコンピューテーション基盤
高前田 伸也	東京大学 大学院情報理工学系研究科	准教授	D3-AI: 多様性と環境変化に寄り添う分散機械学習基盤の創出
竹内 一郎	名古屋大学 工学研究科	教授	AI駆動仮説の静的・動的信頼性保証と医療への展開
山田 誠二	情報・システム研究機構 国立情報学研究所	教授	納得感のある人間-AI協調意思決定を目指す信頼インタラクションデザインの基盤構築と社会浸透

領域アドバイザー

選考・領域運営／研究推進・評価を担当

氏名	所属 （五十音順）
岡田 浩之	玉川大学 工学部　教授
奥村 学	東京工業大学 科学技術創成研究院　教授
神嶌 敏弘	産業技術総合研究所　主任研究員
佐藤 洋一	東京大学 生産技術研究所　教授
辻　ゆかり	NTTアドバンステクノロジ株式会社　IOWN推進室　取締役／室長
福田 雅樹	大阪大学社会技術共創研究センター　教授
福水 健次	統計数理研究所 数理・推論研究系　教授
村上 祐子	立教大学 大学院人工知能研究科　教授
盛合 志帆	情報通信研究機構 サイバーセキュリティ研究所　上席研究員
横尾 真	九州大学 大学院システム情報科学研究院　主幹教授
若宮 直紀	大阪大学 大学院情報科学研究科　教授
鷲崎 弘宜	早稲田大学 理工学術院基幹理工学部　教授

2022年4月時点

AIPネットワークラボ

2016年度に開始した文部科学省「AIP※プロジェクト」をJSTと理研が一体的に推進
JST ： AIPネットワークラボ（戦略的創造研究推進事業のAI関連領域）
理研 ： 革新知能統合研究センター（AIPセンター）

※ AIP：Advanced Integrated Intelligence Platform

文部科学省 MINISTRY OF EDUCATION, CULTURE, SPORTS, SCIENCE AND TECHNOLOGY-JAPAN
AIPプロジェクト

人工知能/ビッグデータ/
IoT/サイバーセキュリティ
統合プロジェクト

理化学研究所
革新知能統合研究センター
（AIPセンター）
杉山センター長

一体的に推進

JST AIPネットワークラボ　ラボ長：江村克己

AIP ネットワークラボ

CREST

- バイオDX（岡田 総括）
- S5基盤ソフト（岡部 総括）
- 信頼されるAIシステム（相澤 総括）
- 数理的情報活用基盤（上田 総括）
- 共生インタラクション（間瀬 総括）
- 人工知能（栄藤 総括）

さきがけ PRESTO

- 社会変革基盤（栗原 総括）
- ICT基盤強化（東野 総括）
- 信頼されるAI（有村 総括）
- 数理構造活用（坂上 総括）
- IoT（徳田 総括）
- 人とインタラクション（暦本 総括）

ACT-X

- 数理・情報のフロンティア（河原林 総括）
- AI活用学問革新創成（國吉 総括）

戦略目標からみる領域の概要

戦略目標（達成目標）

「人間中心のAI社会原則」に基づいた「信頼される高品質なAI」(Trusted Quality AI)の創出に向けた研究開発を推進する。具体的には以下の3つの達成を目指す。

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及び意思決定・合意形成支援技術の創出

提案募集する研究（1/3）

※一例であり募集課題はこれに限りません

（1）「信頼されるAI」の実現に向けた発展的・革新的なAI新技術

- 深層学習のような帰納的な処理と知識・言語による推論・プランニング等の演繹的な処理を最適に融合させたAI技術の研究
- 大量教師データが与えられなくても、実世界環境との相互作用を通して、知識獲得・成長するAI技術の研究
- 人間の脳情報処理や認知発達過程に関する知見に基づく新しいAI原理の研究　等

提案募集する研究（2/3）

※一例であり募集課題はこれに限りません

（2）AIシステムに社会が期待する信頼性・安全性を確保する技術

- 判断・推論の根拠を説明できるAIシステムを実現するための技術の研究
- データ拡張やデータバイアス除去やデータ匿名化などデータを加工する技術の研究
- 未知・想定外ケースや環境変化にも頑健なAIシステムを実現するための技術の研究
- AIシステム全体の信頼性・安全性の確保、品質保証を可能とする技術の研究　等

提案募集する研究（3/3）

※一例であり募集課題はこれに限りません

（3）人間中心のAI社会に向けたデータの信頼性確保及び人間の主体的な意思決定支援技術

- データ改ざんやねつ造（フェイク）等を検知し対処する技術の研究
- 人間が主体性・納得感を持って、適切かつ迅速に判断を下したり合意を形成したりすることを支援する技術の研究　等

応募にあたっての留意点

研究期間と研究費

- 研究期間は約5.5年間（2022年10月から2028年3月末まで）
- 研究期間全体における研究費は3億円(間接経費を除く)が上限
- 必要に応じて研究加速等の支援を実施
- フランスANRとの共同提案においても、ANR側の予算規模にかかわらずCRESTの基準で応募してください。

2022年度 募集スケジュール

- 募集〆切　：　2022年6月7日（火）正午（厳守）
- 書類選考期間　：　6月中旬～7月上旬
- 書類選考結果の通知　：　7月中旬～7月下旬
- 面接選考会　：　8月3日
- 選考課題の通知・発表　：　8月下旬～ 9月下旬
- 研究開始　：　10月1日

※募集〆切の日付け以外は全て予定です。今後変更となる場合があります

選考の観点(CREST共通)(1/2)

- CRESTの各研究領域に共通の選考基準は、以下の通りです。(a.～d. の全ての項目を満たしていることが必要です)。

a. 戦略目標の達成に貢献するものであること。

b. 研究領域の趣旨に合致していること。

c. 独創的であり国際的に高く評価される基礎研究であって、今後の科学技術イノベーションに大きく寄与する卓越した成果が期待できること。

d. 以下の条件をいずれも満たしていること。

・ 研究提案者は、研究遂行のための研究実績を有していること。

・ 研究構想の実現に必要な手掛かりが得られていること。

・ 研究提案書において、①研究構想の背景(研究の必要性・重要性)、②研究提案者の実績(事実)、及び③研究構想・計画の3者を区別しつつ、それぞれが明確に記述されていること。

選考の観点（CREST共通）(2/2)

d. 以下の条件をいずれも満たしていること。(続き)

- 最適な研究実施体制であること。研究提案者がチーム全体を強力に統率して責任を負うとともに、主たる共同研究者を置く場合は研究提案者の研究構想実現のために必要不可欠であって、研究目的の達成に向けて大きく貢献できる十分な連携体制が構築されること。
- 研究提案者の研究構想を実現する上で必要十分な研究費計画であること。
- 研究提案者及び主たる共同研究者が所属する研究機関は、当該研究分野に関する研究開発力等の技術基盤を有していること。

フランスANR(国立研究機構)との日仏共同提案について

1. 日仏の科学研究における協力促進を目的に、2022年度のCRESTの提案募集では、当研究領域において通常の研究提案に加えて、日仏共同研究グループによる共同研究提案を募集します
2. 日仏の研究代表者で1つの共同研究提案書(英語、CREST-ANR共通書式)を作成し、JST(日本)とANR(フランス)にそれぞれ申請していただきます

- JST、ANR両機関に申請されることが審査の要件となります。必ず両機関に申請をしてください。(ANR申請受付期間:2022年2月21日(月)～5月9日(月)10:00 CEST)
- ANRとJSTが各々提案の審査を行った後、両機関で協議の上採択を決定します。
- CRESTにおける選考では、日仏共同研究提案と通常の研究提案とを分けずに審査します。どちらか一方が有利になることはありません。採択後も通常のCREST課題と同様に研究を推進します。
- 研究代表者は日仏共同提案と通常のCRESTの提案の両方を申請することはできません。
- CRESTへの応募の際に、ANRに提出した日仏共同研究提案の内容を変更することはできません。
- 詳細やその他の留意事項は、WEBをご確認ください。

想定する研究の進め方

- 本領域の重要なミッションとして以下があります
  - AI研究にかかわる多様な研究者・ステークホルダーを巻き込み、「信頼性」の定義・要件を検討
  - 個別の要素技術のみならず、人間中心の信頼されるAIシステムを構築
  - AIの信頼性に関する新たな研究コミュニティの創出を志向
- 上記の実現のため、領域内のみならず、同じ戦略目標の下に実施する、さきがけ「信頼されるAIの基盤技術」を始めとする、領域外との連携によるコミュニティ作りを積極的に行うことを推奨します
- R3年度からは更に日仏共同提案を募集し、海外との連携・協働を図っていきます。(詳細は後述)

応募にあたっての留意点（1/2）

- チーム構成
  - 個別の要素技術の発展のみならず、人間中心の信頼されるAI システムをどう構築するかが研究の重要なポイントとなります
  - チームは、1つの課題の解決を目指す構成でも、複数の課題の解決を目指す構成でも構いません
  - 必要に応じて様々な分野・セクターの幅広いメンバー構成を検討して下さい
- 研究のインパクト
  - 領域に参加する研究チームはハイインパクトな研究成果に関する目標を自ら設定して研究開発を推進します
  - 応募にあたっては、信頼されるAI の実現に向けた、提案研究の学術や社会へのインパクトを明らかにして下さい

応募にあたっての留意点（2/2）

- 領域全体のバランスの考慮
  - 本年度の公募においても、領域全体として「概要」にある3つの研究開発項目を展開できるよう、研究テーマのバランスや理論と実装のポートフォリオも考慮しながら選考を進めたいと考えています
  - 本領域の趣旨、目標を十分に勘案していただき、領域全体が活性化されるような研究が数多く提案されることを期待しています

募集要項の補足

- 様式2，頁1，「1．要旨」には、次の項目に関する記載を含めてください。
  - ① 想定する信頼性の定義及び評価法
  - ② 研究成果のインパクト（誰がどのように使うのか？）
  - ③ 研究体制がベストチームであること及びその理由

補足

CREST TQAIS research map (as of 2020)

(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) データの信頼性及び意思決定・合意形成支援技術の創出

D3-AI

インフォデミック克服

AI駆動仮説の信頼性保証

言語で説明できるAIシステム

Interventional AI robotics

ヒューマンコンピュテーション基盤

Explorable推薦基盤

ハイパーデモクラシー

人間-AI協調意思決定

AIシステム中心（基盤・理論）

人間中心（協調）

Human and AI Systems in terms of Human Awareness

Intentional	Unintended	Unconscious
something caused by somebody's malicious intention	**something that goes (unexpectedly) wrong and delivers undesirable results**	**something caused (unconsciously) by limitations of human abilities**

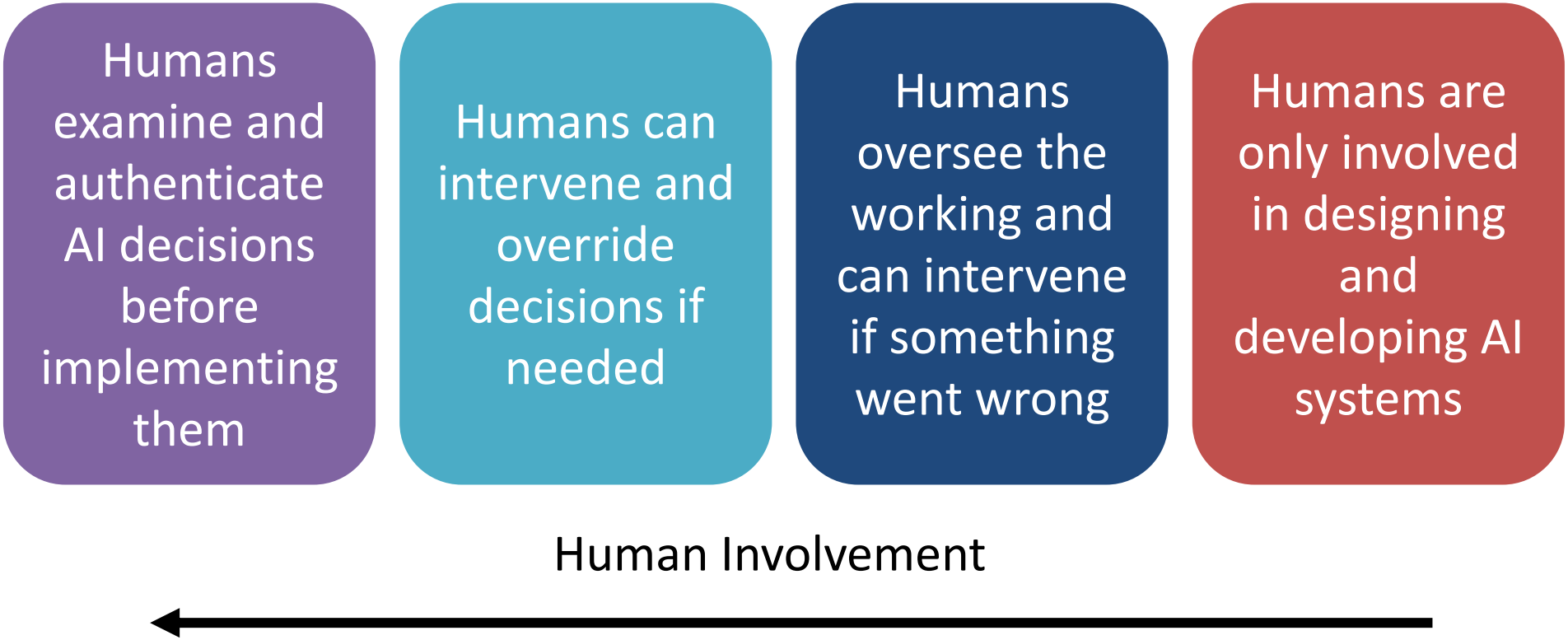
Human Awareness

security attacks, fake, spam, plagiarism, etc.	implementation error, mis-interpretation, unexpected inputs/context	cognitive and social bias, …
悪意ある攻撃からの防御	人間の作業の支援	人間の認知的限界の克服

Technical Aspects of Trusted AI Systems

Human and AI Systems in terms of Human Involvement

Kaur, Davinder, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. "Trustworthy Artificial Intelligence: A Review." ACM Comput. Surv., 39, 55 (2): 1–38.
Figure 2: Different levels of human involvement in making AI systems trustworthy.



AIシステム

＝高度に複雑な情報システム

※統合イノベーション戦略推進会議
「人間中心のAI社会原則」

- 情報分野の研究であれば、たいがいのものはスコープに入る

- しかし、どんな研究でもよいわけではない
- よい研究であれば、何でもよいわけではない

「信頼される」AIシステムとは？

- どのような要件を満足することなのか？
- それをどのように評価したらよいのか？
- それは多くの人の理解を得られるのか？

どのような人が応募するべきか？

- 信頼性という切り口に興味がある人
  → 本領域で想定しているAIシステムの間口は広いので、是非ご検討下さい！

- CRESTへの初挑戦者：
  → 大歓迎。「信頼されるAI」のキーワードのもとで研究を構想して頂けることは何よりありがたいです

- CRESTやさきがけの卒業組：
  → 大歓迎。研究テーマとして継続的なものや単なる発展研究は想定していないので、構想を立てる段階で注意が必要

- 本CRESTは3年目で最後の募集となります。沢山のご提案をお待ちしています。

皆様からのチャレンジングな
提案をお待ちしております！