

戦略プロポーザル

デジタル社会における 新たなトラスト形成

STRATEGIC PROPOSAL

New Trust Formation in the Digital Society

エグゼクティブサマリー

本プロポーザルは、デジタル化の進展につれて社会のさまざまな場面でトラスト（信頼）がうまく働かなくなってきたという状況・問題を、改善・解決するための新たな仕組み作り、すなわち「デジタル社会における新たなトラスト形成」のための研究開発の推進を提言する。

トラストは相手が期待を裏切らないと思える状態である。リスクがあるとしても、相手をトラストできると、安心して迅速に行動・意思決定ができる。トラストは協力や取引のコストを減らしてくれる効果があり、人々の活動を拡大し、ビジネスを発展させ、ビジネスの生き死にを左右する要因にもなる。

しかし、デジタル化の進展につれて、バーチャルな空間にも人間関係が広がり、複雑な技術を用いたシステムへの依存が高まり、だます技術も高度化してしまった。その結果、デジタル社会と言われる今日において、顔が見える人間関係や人々の間のルールに支えられた「旧来のトラスト」だけではカバーされないケースが拡大し、社会におけるトラストの働きがほころんできている。この問題は、自動運転車、AI エージェント、コミュニケーションロボット、メタバースなどの新技術・新サービスの社会受容を左右し、フェイク・偽装・なりすましなどによる詐欺・犯罪の懸念を高める。

この問題に対処するため、「デジタル社会における新たなトラスト形成」の研究開発推進が強く求められる。この取り組みによって目指すのは、不信・警戒を過度に持つことなく幅広い協力・取引・人間関係を作ることができ、デジタル化によるさまざまな可能性・恩恵がより広がるような社会である。これを実現するには、技術開発だけでなく、制度設計や社会受容も考えた「総合知」に基づく研究開発が必要である。

デジタル化が進展し、上述のようなほころびが生じている状況のなか、断片的に切り取られた情報や対象のある一面しか見ずに、何かを信じ込むことはとても危うい。デジタル社会のトラストは、多面的・複合的な検証によって支えていくべきであり、その検証が適切かつ容易に行えるような仕組みが望まれる。

本プロポーザルでは、この多面的・複合的な検証という考えを具現化する上で助けとなる概念として、「トラストの3側面」を導入する。すなわち、トラストは、どのような側面に着目するかによって「対象真正性」（本人・本物であるか?）、「内容真実性」（内容が事実・真実であるか?）、「振る舞い予想・対応可能性」（対象の振る舞いに対して想定・対応できるか?）という3通りに整理できる。これら3側面のそれぞれに対して「社会的よりどころ」になるものを技術や制度によって用意し、それらを用いた多面的・複合的な検証を、社会の中で広く活用される仕組みとして定着させることが、前述した目指す姿の実現につながる。

そのために必要な「総合知」に基づく研究開発として、本プロポーザルでは、次のような4層から成る研究開発課題への取り組みを連携して推進することを提言する。

- (1) トラストの社会的よりどころの再構築： トラストの3側面（対象真正性／内容真実性／振る舞い予想・対応可能性）で何を社会的よりどころに設定するか、社会的よりどころをどのような技術と制度によって担保するか、に関する研究開発。
- (2) 社会的トラスト形成フレームワーク： 人々が社会的よりどころを容易に使いこなし、トラストできる対象を広げていけるようにするとともに、社会的よりどころが公正・健全に維持されるようにするための研究開発。
- (3) 具体的トラスト問題ケースへの取り組み： デジタル化の進展で生じたトラスト問題ケースに対して、(1)(2)の枠組みを用いた解決や状況改善を実証する研究開発。具体的ケース固有の問題分析・対処と具体的ケースからの(1)(2)の研究開発へのフィードバックも含む。
- (0) トラストに関する基礎研究： (1)(2)(3)の実現とその社会受容のための、社会におけるトラストに

についての理解や、そのデジタル化による影響・変化に関する基礎的な研究開発。

トラストに関わる研究開発は、これまでもさまざまな分野で取り組まれてきている。しかし、それらの間で知見共有や連携がほとんどなく、それぞれはトラスト問題に対して個別的な対処や断片的な状況改善にとどまっている。本プロポーザルは、このような現状から取り組み方の変容を促す。すなわち、デジタル社会における新しいトラストの仕組みとそれによるトラスト問題対策の全体ビジョンを描いて共有し、具体的トラスト問題と共通基礎の両面からの取り組み（4層から成る研究開発課題）を連携して進める。その結果、それらの取り組み全体として社会に貢献する道筋も強化される。

このような変容を促すための方策としては、「分野間の知見共有・連携促進の場作り」、「学際的トラスト研究の継続的活動体制と人材育成策」、「ファンディングプログラムによる推進・加速」を組み合わせることが有効と考える。また、「総合知」による研究開発の推進においては、情報科学分野と人文・社会科学分野の横断的な取り組みに加えて、新たなトラストの仕組みに対して期待するとともに実践において大きな役割を担う産業界からの参画も重要である。

Executive Summary

This proposal recommends the promotion of research and development to create a new mechanism to improve and solve the situation and problem of trusts not working well in various aspects of society as digitalization progresses, namely, "new trust formation in the digital society".

Trust is the state of being able to believe that the other party will not disappoint you. If you can trust the other party, even if there is risk, you can act and make decisions quickly and with peace of mind. Trust has the effect of reducing the cost of cooperation and transactions, expanding people's activities, developing business, and being a factor in whether a business lives or dies.

However, as digitization has progressed, relationships have expanded into virtual spaces, reliance on complex technological systems has increased, and deception techniques have become more sophisticated. As a result, in today's digital society, cases that cannot be covered only by the "old trust" supported by face-to-face relationships and rules among people are expanding, and the function of trust in society is no longer working well. This issue affects the social acceptance of new technologies and services such as autonomous cars, AI agents, communication robots, and metaverse, and raises concerns about fraud and crime through faking, disguising, and spoofing.

To address this issue, promotion of research and development of "new trust formation in the digital society" is strongly urged. The goal of this initiative is a society where a wide range of cooperation, transactions, and relationships can be created without excessive distrust or caution, and where the various possibilities and benefits of digitalization can be expanded. To realize this, it is necessary to conduct research and development based on "convergence of knowledge" that considers not only technological development but also institutional design and social acceptance.

In the situation where trust does not work well due to the problems caused by digitalization, it is very dangerous to believe something based on only fragmented information or one aspect of the subject matter. Trust in the digital society should be supported by multifaceted and composite verification, and a mechanism is desirable to ensure that such verification can be performed appropriately and easily.

This proposal introduces the "three aspects of trust" as a concept that will assist in embodying this idea of multifaceted and composite verification. The three aspects of trust are (a) "object certification", whether the trustee is the real person or the real item, (b) "content authenticity", whether the content is a fact or the truth, and (c) "behavior expectation and acceptance", whether it is possible to expect or accept to the behavior of the trustee. The above-mentioned goal of this proposal will be achieved through the establishment of a system that provides "social bases" for each of these three aspects through technology and institutions, and through the establishment of a multifaceted and composite verification process that is widely utilized in society.

As research and development based on the "convergence of knowledge" necessary for this purpose, this proposal recommends that the following four-tiered approach to research and development targets be promoted in a coordinated manner.

(1) Rebuilding the social bases of trust: Research and development on what to set as the social basis in terms of the three aspects of trust (object certification/content authenticity/behavior expectation and acceptance) and what technologies and institutions should be used to ensure this social basis.

(2) Social trust formation framework: Research and development to make it easier for people to use the social bases of trust and expand the range of trustable objects, as well as to ensure that social trust anchors are maintained in a fair and healthy manner.

(3) Addressing specific trust problem cases: Research and development to demonstrate solutions or situation improvements using (1) and (2) for trust problem cases caused by the digitization. Here, not only solution design for specific problem cases but also feedbacks from analysis of specific cases to the research of (1) and (2) are included.

(0) Basic research on trust: Basic research and development for the realization of (1), (2) and (3) and its social acceptance, for example, research on understanding of trust in society and its impact and change due to digitalization

Trust-related research and development has been tackled in a variety of fields. However, there is little knowledge sharing or collaboration among them, and each is limited to addressing individual cases or improving the situation piecemeal. This proposal encourages a change in approach from this current situation. In other words, we should draw and share an overall vision of a new trust mechanism in the digital society and solutions for actual trust problems, and work together on the research and development agenda consisting of four layers to address both actual trust problems and common foundations. As a result, social contribution scenarios of those efforts will also be strengthened.

In order to achieve the above-mentioned change, we believe it is effective to combine the following measures: (i) setting opportunities for sharing knowledge and promoting collaboration among disciplines, (ii) continuous activity formation for interdisciplinary trust research and human resource development, and (iii) promotion and acceleration through funding programs. In the promotion of research and development based on "convergence of knowledge," in addition to cross-disciplinary efforts between the information science field and the humanities and social sciences, participation from industry which needs a new trust mechanism and plays a major role in the practice is also important.

目次

1	研究開発の内容	1
2	研究開発を実施する意義	4
2.1	現状認識および問題点	4
2.2	社会・経済的効果	12
2.3	科学技術上の効果	14
3	具体的な研究開発課題	16
3.1	トラストの社会的よりどころの再構築	17
3.2	社会的トラスト形成フレームワーク	21
3.3	具体的トラスト問題ケースへの取り組み	22
3.4	トラストに関する基礎研究	24
4	研究開発の推進方法および時間軸	27
付録		30
付録 A	検討の経緯	30
付録 B	国内外の状況	35
付録 C	用語解説	39
付録 D	参考文献	42
コラム一覧		
コラム 1	情報に関する対象真正性と内容真実性	10
コラム 2	トラストの弊害	15
コラム 3	内容真実性：事実や真実は定まるものなのか？	19
コラム 4	AI 応用システムの振る舞い予想・対応可能性	20
コラム 5	デジタル化とトラスト問題の関係	25

1 | 研究開発の内容

本プロポーザルが掲げる「デジタル社会における新たなトラスト形成」とは、デジタル化の進展につれて、社会のさまざまな場面でトラスト（信頼）がうまく働かなくなってきたという状況・問題を、改善・解決するための新たな仕組み作りを意味する。この社会的な仕組み作りには、技術開発だけでなく、制度設計や社会受容も考えた「総合知」に基づく研究開発が必要である。

トラストは相手が期待を裏切らないと思える状態である。リスクがあるとしても、相手をトラストできると、安心して迅速に行動・意思決定ができる。しかし、デジタル化の進展につれて、バーチャルな空間にも人間関係が広がり、複雑な技術を用いたシステムへの依存が高まり、だます技術も高度化してしまった。その結果、デジタル社会と言われる今日において、顔が見える人間関係や人々の間のルールに支えられた「旧来のトラスト」だけではカバーされないケースが拡大し、社会におけるトラストの働きがほころんできている^{[1][2][3]}。この問題意識に基づき、本プロポーザルが目指すのは、「デジタル社会における新たなトラスト形成」によって、不信・警戒を過度に持つことなく幅広い協力・取引・人間関係を作ることができ、デジタル化によるさまざまな可能性・恩恵がより広がるような社会である（図1-1）。

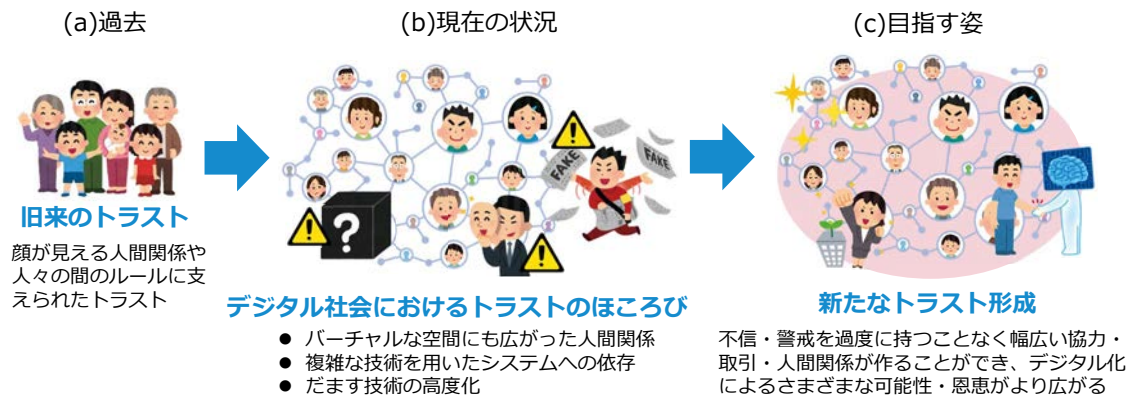


図1-1 デジタル社会におけるトラストの問題認識と目指す姿

デジタル化が進展し、上述のようなほころびが生じている状況のなか、断片的に切り取られた情報や対象のある一面しか見ずに、何かを信じ込むことはとても危うい。デジタル社会のトラストは、多面的・複合的な検証によって支えていくべきであり、その検証が適切かつ容易に行えるような仕組みが望まれる。本プロポーザルでは、この多面的・複合的な検証という考えを具現化する上で助けとなる概念として、「トラストの3側面」を導入する。すなわち、トラストは、どのような側面に着目するかによって「対象真正性」「内容真実性」「振る舞い予想・対応可能性」という3通りに整理できる（表1-1）¹。これら3側面のそれぞれに対して「社会的よりどころ」になるものを技術や制度によって用意し、それらを用いた多面的・複合的な検証を、社会の中で広

1 この「トラストの3側面」に関する整理は、JST CRDS 俯瞰セミナー&ワークショップ^[1]での議論において、大屋雄裕教授（慶應義塾大学）からいただいた示唆に基づく。このうち「振る舞い予想・対応可能性」の説明がやや分かりにくいかもしれない。多くの場合、「対象の振る舞いが想定通り（予想通り・期待通り）か？」と解釈すればよいのだが、想定から外れたとしても、何らかの補償・賠償が約束されているならば大丈夫と思えることがある。それゆえ「想定・対応できるか？」という説明にした。

表 1-1 トラストの3側面

トラストの3側面	どのような側面に関するトラストか
(A) 対象真正性	本人・本物であるか？
(B) 内容真実性	内容が事実・真実であるか？
(C) 振る舞い予想・対応可能性	対象の振る舞いに対して想定・対応できるか？

く活用される仕組みとして定着させることが、前述した目指す姿の実現につながる。

そのために必要な「総合知」に基づく研究開発として、本プロポーザルでは、表 1-2 のような4層から成る研究開発課題への取り組みを連携して推進することを提言する。

表 1-2 4層から成る研究開発課題

研究開発課題	内容
(1) トラストの社会的よりどころの再構築	トラストの3側面（対象真正性／内容真実性／振る舞い予想・対応可能性）で何を社会的よりどころに設定するか、社会的よりどころをどのような技術と制度によって担保するか、に関する研究開発。
(2) 社会的トラスト形成フレームワーク	人々が社会的よりどころを容易に使いこなし、トラストできる対象を広げていけるようにするとともに、社会的よりどころが公正・健全に維持されるようにするための研究開発。
(3) 具体的トラスト問題ケースへの取り組み	デジタル化の進展で生じたトラスト問題ケースに対して、(1)(2)の枠組みを用いた解決や状況改善を実証する研究開発。具体的ケース固有の問題分析・対処と具体的ケースからの(1)(2)の研究開発へのフィードバックも含む。
(0) トラストに関する基礎研究	(1)(2)(3)の実現とその社会受容のための、社会におけるトラストについての理解や、そのデジタル化による影響・変化に関する基礎的な研究開発。

トラストに関わる研究開発は、これまでもさまざまな分野で取り組まれてきている。しかし、それらの間で知見共有や連携がほとんどなく、それぞれはトラスト問題に対して個別的な対処や断片的な状況改善にとどまっている。このような現状に対して、上で述べたような取り組みを通して、図1-2のような変容を促す。すなわち、デジタル社会における新しいトラストの仕組みとそれによるトラスト問題対策の全体ビジョンを描いて共有し、具体的トラスト問題と共通基礎の両面からの取り組みを連携して進めることで、それらの取り組み全体として社会に貢献する道筋も強化されるというのが、本プロポーザルが描くトラスト研究開発の目指す姿である。

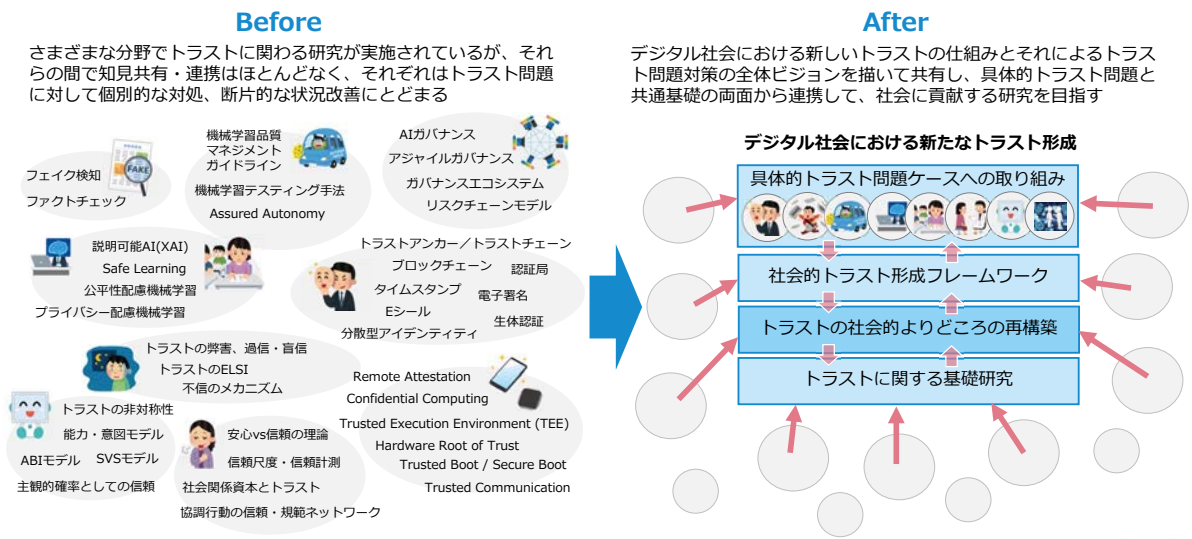


図 1-2 トラスト研究開発の目指す姿

2 | 研究開発を実施する意義

2.1 現状認識および問題点

2.1.1 デジタル社会におけるトラスト問題

第1章でも述べたように、トラスト（信頼）は相手が期待を裏切らないと思える状態である¹。リスクがあるとしても、相手をトラストできると、安心して迅速に行動・意思決定ができる。社会におけるトラストの役割・効果について、ニクラス・ルーマン（Niklas Luhmann）が「社会的な複雑性の縮減メカニズム」^[4]と表現したことはよく知られている。工藤郁子は「取引や協力のコストを減らしてくれる社会関係資本」^[5]、レイチェル・ボツツマン（Rachel Botsman）は「確実なものと同確実なものとの隙間を埋める力であり、未知のものに自信を持つこと」^[6]と記している。また、企業にとってトラストは、ビジネスを発展させ、ビジネスの生き死にを左右する要因になる。

しかし、図1-1に示したように、デジタル化の進展につれて、社会におけるトラストの働きにほころびが見られるようになってきた。「旧来のトラスト」は、顔が見える人間関係や人々の間のルールに支えられたものだったが、デジタル社会と言われる今日、「旧来のトラスト」だけではカバーされないケースが拡大した。その要因として、図1-1(b)に示した3点が挙げられる²。すなわち、第一の要因は、インターネットを介した商取引やSNS（Social Networking Service）などに見られるように、バーチャルな空間にも人間関係が広がったことである。第二の要因は、人工知能（Artificial Intelligence: AI）などの、複雑でブラックボックスとも言われる技術を用いたシステムへの依存が高まったことである。第三の要因は、AI技術によるフェイク画像・動画・音声の生成や、アカウントのなりすましなど、「だます技術」が高度化していることである。

これまでの有識者へのヒアリングやワークショップでの議論^[1]に基づき、デジタル化の進展が生んだトラスト問題として特徴的なケースを図2-1に例示する。インターネットを介して人間関係が広がるとともに、デジタル技術を用いた「だます技術」が高度化したことで、「(a) 仮想世界のトラストに基づく取引」における偽装やなりすましの問題が起き、「(b) メディアにおけるフェイク拡散」も社会問題化している。また、「(c) 自動運転車」「(d) パーソナルAI エージェント」「(e) 人を評価するAI システム」は、ブラックボックス的なAI技術が組み込まれたシステムをトラストできるかという問題、および、そのようなシステムに依存したときに生じる懸念を示したケースである。さらに、対人関係にAI・ロボット・アバターといったものが関わってくることで、対人関係のトラストにこれまでと異なる新たな様相・懸念が生じるケースとして、「(f) 医療意思決定におけるAI セカンドオピニオン」「(g) コミュニケーションロボット」「(h) メタバース内活動におけるトラスト」を挙げた。

このように、デジタル化の進展に伴い、社会のさまざまなシーンでトラストに関わる問題・懸念が生じている。

1 トラストの定義やモデルについては、2.1.3節において、より詳しい説明を試みる。

2 これら3つの要因は、ある主体と他の主体との間に情報の格差（情報の非対称性）を生むことにもつながりやすい。情報の非対称性という面からも、トラストの形成への影響が考えられる。

(a) 仮想世界のトラストに基づく取引

ネットの世界で、リアルには面識のない人たちとの取引や、仮想通貨・デジタル資産を用いた取引が拡大。さまざまな分野で、シェアリングビジネス、不特定多数の中からのマッチングビジネスも立ち上がっている。



仮想世界・デジタルデータの性質を悪用した偽装やなりすまし等の犯罪も起きている。対策も取られているが、常に新しい仕組みが生まれ、新しいリスクが発生し、対策が追いつかない面もある。

(b) メディアにおけるフェイク拡散

最新のAI技術によって人間の認識能力では見破れないフェイクの作成が容易になった。ソーシャルメディアの普及によって、フェイクやデマの拡散が大規模化。



フェイク動画による政治干渉や個人攻撃・棄損等が社会問題化。簡単に人を騙すことができしまい、裁判等での証拠の信頼性も揺らぐ。フェイクの法的規制が強くなると、表現の自由が妨げられる恐れも生じる。

(c) 自動運転車

AI技術を活用した状況認識や運転制御によって、車が運転手なしで走行する地区・状況が広がる。



AI技術はブラックボックスで動作保証や精度保証ができない。安心して乗車できるのか？事故が発生したときに、原因解明や責任の所在はどうなるのか？

(d) パーソナルAIエージェント

個人情報の管理代行をパーソナルAIエージェントに任せるサービスの利用が増えていく。



ブラックボックスで、個人の意図・期待の通りに振る舞うという100%保証はできないのに、個人情報を委ねることができるのか？期待に反する事態が起きた場合の責任の所在はどうなるのか？

(e) 人を評価するAIシステム

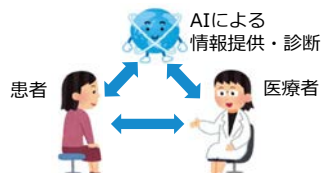
採用試験の一次フィルタリング、人事評価や配属最適化といった人事業務へもAIシステムが応用されつつある。中国では個人の様々な行動履歴を追跡し、信用スコアを算出して、優遇や制限を与えるシステムが稼働している。



学習データに偏りや差別的な要因が含まれていると、不公平で差別的な評価を助長するという問題がある。また、評価アルゴリズムに過剰適合して行動する人々を生み出す。

(f) 医療意思決定におけるAIセカンドオピニオン

従来は患者と医療者の二者関係による医療意思決定だった。そこに、AIによる情報提供や診断が加わり、三者関係による医療意思決定へと変化しつつある。



患者が異なる多様な情報を参照できるようになり、医療者からの説明と異なる情報が得られることもある。三者関係における新たな役割関係とそこでのトラストのあり方が必要になっている。

(g) コミュニケーションロボット

親近感を持つ外観や、会話を含むコミュニケーション能力を備えたコンピューターエージェントあるいはロボットが、さまざまなサービスや日常的なタスクにおいて、対人インターフェイスとして広がる。



人間らしい外観を持っていると、人間並みの能力を持っていると期待・過信してしまい、そうでないと失望するといったことが起きやすい。その一方、身近なロボットに、過度な親近感・依存感を持ってしまふタイプの人もいる。

(h) メタバース内活動におけるトラスト

仮想世界の中で自分のアバターを介した新たな人間関係や経済活動が生まれる。



生身の人間や物理的な実体が必ずしも確認できない世界において、リアル世界と同様のトラストが成り立ち得るか？

図 2-1

デジタル化の進展が生んだトラスト問題の例

2.1.2 トラスト研究開発の状況

2.1.1 節でも述べたように、トラストは社会秩序、人間関係、ビジネス上の取引など、さまざまな形で社会における重要な役割・効果を果たしていることから、トラストに関わる研究開発も、幅広い分野でさまざまな取り組みが進められてきた。その研究開発動向に関する俯瞰的な整理については報告書^[1]にまとめたので適宜参照いただくこととし（付録B.1も参照）、ここでは主な研究開発トピックを、図2-2を用いて概括する。

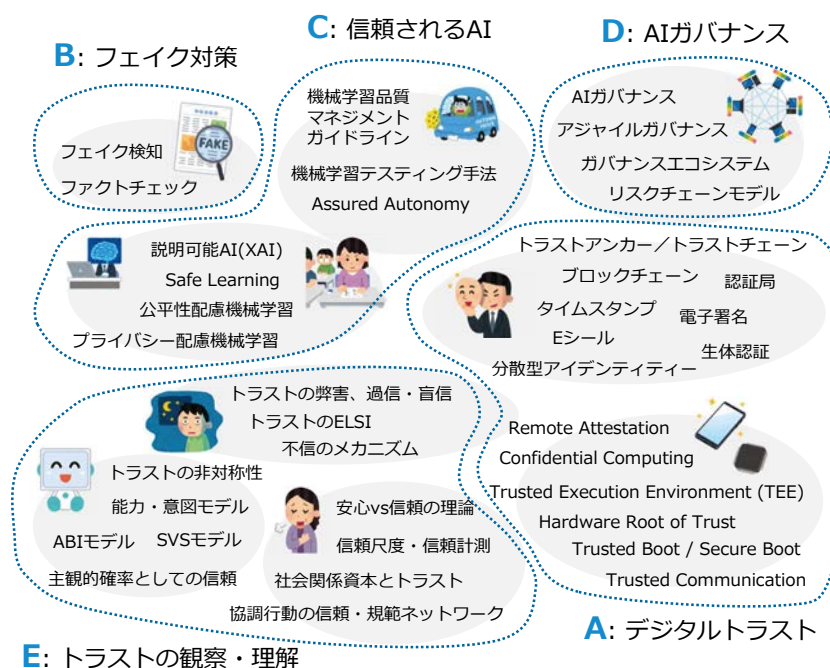


図2-2 トラストに関わる主な研究開発トピック

まず、特に活発に取り組まれているのが図2-2において「A: デジタルトラスト」とまとめたトピック群である^[7]。セキュリティの研究開発と重なる領域であり、ハードウェア的な改ざん防止の仕組み、ブロックチェーン（改ざん防止機能を持つ分散型台帳）、暗号技術に基づく電子署名・認証、個人や機器の認証技術などが含まれる。

「B: フェイク対策」には、近年社会問題化しているフェイクニュースやフェイク画像・動画・音声などの検知や拡散防止のための技術や、ファクトチェックの活動・システムなどが該当する^{[8][9]}。「A: デジタルトラスト」と比べると、何が正しいかを定めることが難しく、「表現の自由」の問題に関わることも多い。

「C: 信頼されるAI」(Trustworthy AI) も近年非常に活発に取り組まれるようになったトピック群で、AIシステムの安全性・信頼性・社会受容性を確保するためのさまざまな技術開発が進められている^{[10][11]}。ブラックボックスとも言われる深層学習（Deep Learning）などのAI技術による判定結果に説明を与える説明可能AI（Explainable AI: XAI）技術、機械学習技術を用いた帰納型のシステム開発におけるテスト・デバッグ手法や品質管理ガイドラインなどソフトウェア工学的な方法論、プライバシーや公平性を確保するデータ分析法など多岐にわたっている。

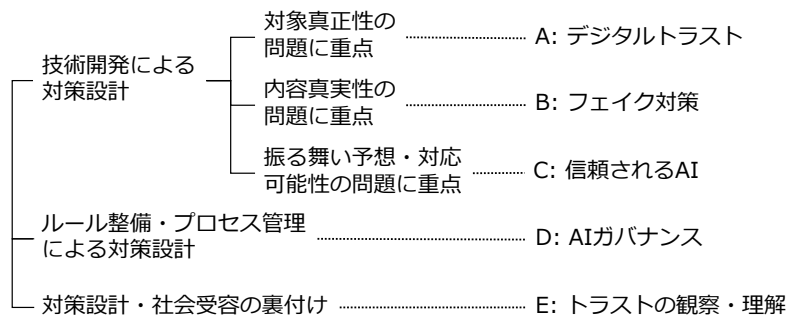
AIシステム開発のための技術群とは別に、運用に関わるガイドラインや活用に関わるルールの整備などを含めた「D: AIガバナンス」と呼ばれる取り組みも進められている^{[2][12]}。

また、哲学、社会学、政治学、心理学、経済学などの分野では、トラストの定義・モデル化、人間の振る

舞い・態度に関する観察・比較実験などが行われてきた。この「E: トラストの観察・理解」の研究では、長年、人間に対するトラストが論じられてきたが、近年は機械・システムに対するトラストも検討対象に含まれるようになってきた^[13]。

以上のように、トラストに関連する異なる切り口からさまざまな研究開発が進められている。研究論文数が近年大きく増加している研究開発トピックも多い。しかし、さまざまな分野で進められているトラスト研究開発の間での知見共有・連携はほとんど見られない。多くの研究開発は、トラストに関わるある一面にフォーカスしたものであることから、トラスト問題に対して個別的な対処や断片的な状況改善にとどまりがちである。この状況は、国内だけでなく海外でも同様である。

このような知見共有・連携がほとんど見られないという状況が生じた背景には、図2-3に示したようなトラスト研究開発における観点の違いがあると思われる。JST CRDS主催で2021年7月29日から10月1日にかけて、さまざまな分野で取り組まれているトラスト研究開発を俯瞰するセミナーシリーズとワークショップ^[1]を開催したところ、多くの参加者から、分野横断で知見を共有することの有用性や、トラストに関する総合的な研究開発のために分野横断で連携することの必要性が指摘され、その議論から前述の表1-1のような「トラストの3側面」という整理を得た。これを図2-2と対応付けると、「A: デジタルトラスト」が対象真正性、「B: フェイク対策」が内容真実性、「C: 信頼されるAI」が振る舞い予想・対応可能性に対応し、A・B・Cはトラストの異なる側面の問題に重点を置いた、技術開発による対策設計の取り組みと位置付けることができる。「D: AIガバナンス」は振る舞い予想・対応可能性に重点を置きつつも、ルール整備やプロセス管理の面から対策を検討している。また、A～Dのような対策設計やその対策の社会受容の裏付けとなるような基礎研究として「E: トラストの観察・理解」が位置付けられる。C・Dは共にAIを扱っているの2つの取り組みの間には連携が見られるが、それ以外は現状、トラスト研究開発の間での知見共有・連携はほとんど見られない。なお、A～Dはデジタル技術と不可分な検討であるのに対して、Eの多くは必ずしもデジタル技術を前提とするものではない。以上をまとめると図2-3のようになる。



[注] おおまかな傾向であり、この見方に収まらない取り組みも存在する

図2-3 トラスト研究開発の観点の違い

2.1.3 トラスト関連政策・プログラムの状況

2.1.2節に示した研究開発トピックの分類に沿って、関連政策・プログラムとして注目されるものを、表2-1にまとめた（付録B.2にも補足情報を記載）。必ずしも網羅したものではなく、日本国内における取り組みをまず押さえた上で、海外での取り組みは特徴的なもののみを挙げたものであるため、国内の取り組みに関する記載が相対的に多くなったが、わが国の政策において「トラスト（信頼）」は一つの訴求点とされていることは特記しておきたい。特に「A: デジタルトラスト」において、2019年のダボス会議で安倍首相（当時）が

「Data Free Flow with Trust (DFFT)」を打ち出したことや、「C: 信頼されるAI」において、統合イノベーション戦略推進会議で閣議決定された「AI戦略2019」で「Trusted Quality AI」が中核研究課題に位置付けられたことは注目に値する。

ただし、同じく「トラスト」という言葉が用いられているものの、図2-3に示したように、それぞれトラストの異なる側面を扱っている。「トラストの3側面」を整理し、それらを総合的に扱うようにするならば、トラストを軸として、より一貫した戦略・政策を打ち出すことが可能になるはずである。

表2-1 注目されるトラスト関連政策提言・プログラム事例

分類	国	政策提言・プログラム事例
A: デジタルトラスト	日本	「Data Free Flow with Trust (DFFT)」(2019年1月ダボス会議、安倍首相) - デジタルトラスト協議会(2020年8月設立) - 内閣デジタル市場競争本部 Trusted Web 推進協議会「Trusted Web ホワイトペーパー Ver.1」(2021年3月)、「同 Ver.2 (案)」(2022年7月) - デジタルガバメント閣僚会議 データ戦略タスクフォース「包括的データ戦略」(2021年6月閣議決定) - デジタル庁 データ推進戦略ワーキンググループ トラストを確保したDX推進サブワーキンググループ(2021年11月～)
	欧州	eIDAS (electronic Identification and Authentication Service) 規則(2014年7月成立、2016年7月施行): トラストサービスの統一基準
B: フェイク対策	米国	国防高等研究計画局(DARPA)「Media Forensics (MediFor)」プログラム、「Semantic Forensics (SemaFor)」プログラム
C: 信頼されるAI	日本	- 統合イノベーション戦略推進会議「AI戦略2019」(2019年6月)における主要な研究開発課題として「Trusted Quality AI」 - 文部科学省2020年戦略目標「信頼されるAI」を受けたJSTプログラム: CREST「信頼されるAIシステム」、さきがけ「信頼されるAI」(2020年度～) - NEDO「次世代人工知能・ロボット中核技術開発事業」において「AIの信頼性」(2020年度～)
	英国	英国研究・イノベーション機構(UKRI)「Trustworthy Autonomous Systems Programme」
D: AI ガバナンス	日本	- 世界経済フォーラム「Rebuilding Trust and Governance: Towards DFFT」白書(2021年3月) - 経済産業省「Governance Innovation Ver.2: アジャイル・ガバナンスのデザインと実装に向けて」(2021年7月)、「AI原則実践のためのガバナンス・ガイドライン Ver. 1.1」(2022年1月) - 日本ディープラーニング協会「AIガバナンスとその評価」研究会報告書「AIガバナンス・エコシステム-産業構造を考慮に入れたAIの信頼性確保に向けて-」(2021年7月)
	欧州	欧州委員会「Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts」(2021年4月)

2.1.4 トラストのモデルとデジタル社会におけるトラスト問題の捉え方

トラストはさまざまな定義^[1]がなされているが、それらにある程度共通な理解として、本プロポーザルでは「トラストは相手が期待を裏切らないと思える状態」と定義した。また、多くの論文に共通して示されているトラストの性質として、次の4点が挙げられている^[14]。

- Trustor (トラストする側) と Trustee (トラストされる側) という2者間の関係である。

- トラストは危険な状況や不確実な状況に存在する。
- トラストはリスクを取る行動につながる。
- トラストは非常に主観的な問題で、個人と環境の状況・文脈の影響を受ける。

これらを踏まえるとともに「Trusted Web ホワイトペーパー」^[3]も参考にして、本プロポーザルでは、トラストについて図2-4のようなモデル化を行う³。これをベースとして、現状の問題点を述べることで、第3章の研究開発課題の具体化につなげる。

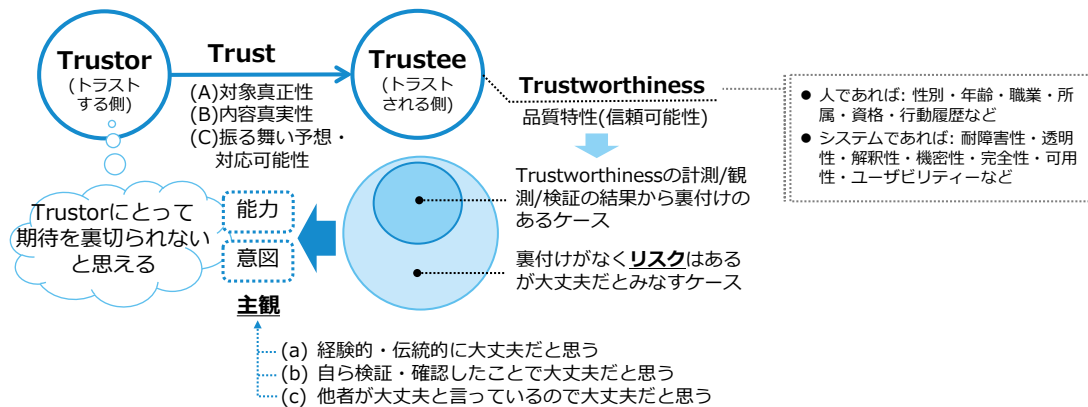


図2-4 トラストのモデル

図2-4に示したトラストのモデルの概要を説明する。

まず、トラストは、TrustorとTrusteeの関係であり、Trustorにとって、Trusteeは期待を裏切らないと思えるときに、トラストが成り立つ。期待を裏切らないと思えるか否かは、Trustorの主観的な判断による。その判断の際に着目され得るTrusteeの品質特性をTrustworthinessと呼ぶ。それらTrustworthinessを計測／観測／検証した結果から裏付けがあって、期待を裏切られないと思えるケースもあれば、必ずしもそのような裏付けはなく、リスクがあるが、大丈夫だとみなすケースもある。この2つのケースの比がどの程度ならば大丈夫と思えるかは、人によって異なる。

期待を裏切られない（大丈夫だ）と思えるかを主観的に判断する際には、Trusteeの能力（期待に応える能力があるか）や意図（期待に応えようとしているか／悪意を持っていないか）などを推測して判断していると言われる⁴。また、この主観的な判断にはいくつかのパターンがあり、(a) 経験的・伝統的に大丈夫だと思う、(b) 自ら検証・確認したことで大丈夫だと思う、(c) 他者が大丈夫と言っているので大丈夫だと思う、といったものが挙げられる。

また、既に言及しているように「トラストの3側面」（第1章の表1-1）として、(A) 対象真正性（本人・本物であるか?）、(B) 内容真実性（内容が事実・真実であるか?）、(C) 振る舞い予想・対応可能性（対象の

3 図2-4のモデルは、さまざまな文献に共通的に見られる見方を取り込んだが、これが定番というわけではなく、あくまでトラストのモデル化の一例である。研究開発課題の説明に際して、本プロポーザルなりのトラストの捉え方として示したが、トラスト研究開発をこれに沿ったものに限定するものではない。

4 ここではトラストの判断条件として能力・意図モデル^[15]を例に挙げたが、他にもABIモデル（Ability：能力、Benevolence：善良さ、Integrity：誠実さ）^[16]ほか、さまざまなモデルが提案されている。なお、これらは、相手（Trustee）が期待を裏切らないかについて、相手の内面・資質などを推し量るものだが、「トラストの3側面」は相手の何をトラストするかを分類したものである。

振る舞いに対して想定・対応できるか?)を分けて考える。「トラストは相手が期待を裏切らないと思える状態」という定義では、(C)の側面(相手の振る舞い)を中心に置いているが、そもそもその相手が本人・本物だという(A)の側面は大前提である。また、相手が人や組織やシステムという場合は、その振る舞いが期待を裏切らないかを考えるが、相手が情報という場合もあり、その場合は、その内容が事実・真実だという(B)の側面を期待する。これは別な言い方をすると、相手が情報の場合、それを使って行動・意思決定をするのは自分であり、それによって自分が招いた結果が期待通りのものかが、内容真実性に左右されるということである。

人が主観的に判断する場合には、おそらく3側面を多面的・複合的に捉えた上で、トラストできるかを総合的に判断していると思われる。例えば、ある新しいサービスを使ってみようかと考えるとき、「そのサービスの

2

研究開発を実施する意義

コラム1

情報に関する対象真正性と内容真実性

対象真正性は「本人・本物であるか?」を意味する。その一例として、対象が人や会社であれば、確かにその本人であるとか、確かにその会社であるとか、契約や手続きの際に確認するし、確認を求められる。また、対象が物理的なモノであれば、例えばしばしば問題になるブランド品の真贋問題は、対象真正性の問題である。

では、対象が情報の場合はどうか。情報の内容が正しいか／否か(事実・真実であるか／否か)は内容真実性の観点であり、オリジナルの情報から改ざんされていないかは対象真正性の観点になる。このとき、フェイク動画を対象とするならば、オリジナルの動画から改ざんされたものではないかというのは対象真正性の観点であり、動画の内容が実際に起きた事実なのかというのは内容真実性の観点になる。したがって、例えばディープフェイクの技術によって、オリジナルの動画中の人物の顔を別人の顔に置き換えたフェイク動画は、対象真正性と内容真実性の両面で問題視される。

この例では対象真正性と内容真実性に類似するところを感じるかもしれない。しかし、対象真正性は基本的に「本人・本物であるか／否か」という2値であり、その「社会的よりどころ」は本人・本物だという証明書(いわば100%保証)が与えられるという形であるのに対して、内容真実性はコラム3で詳しく述べるように、100%保証はあり得ないものであり、その確かさにはグラデーションがあるものとして考えねばならない。このような性質の違いがあり、現状の対象真正性を扱うための枠組み(デジタルトラスト／トラストサービス)から、内容真実性は分けて考えられている。ただ、対象真正性についても、証明書自体は100%保証の形を取っているが、その証明書の発行元がどれくらい信用できるかについてはグラデーションがある。後述する複合的検証のメカニズムでは、対象真正性と内容真実性を合わせて扱う仕組みを考えていく。

仕組み（どのように動いてどのような結果が得られそうか）が信じられるか」（振る舞い予想・対応可能性の側面）、「そのサービスの提供企業が怪しくないか」（対象真正性の側面）、「そのサービスについての評判やレビュー投稿は本当か／ヤラセではないか」（内容真実性の側面）というように、多面的にチェックしようとするであろうし、一つの側面についても複数の情報を突き合わせることもするであろう。このように、いろいろな視点から多面的に関連情報を集め、その一つだけでは確信を持てなくとも、それらを複合的に検証することで、総合的な判断を下すというようなことを、人は行っている。

その一方で、デジタル社会のトラスト問題への技術的な対策や制度的な対策を作る場合は、いったん (A) (B) (C) に分解し、それぞれに対処する技術や制度を作るのが妥当である。実際、2.1.2 節で述べたように、(A) (B) (C) それぞれの問題に対処する研究開発が進められている。その対処の方法は、図 2-4 のモデルに基づくならば、裏付けのあるケースを拡大する（二重円の内側の円の比率が大きくなるようにする）というものが主になる。すなわち、デジタル化の進展に伴い、バーチャルな空間にも広がった人間関係、複雑な技術を用いたシステムへの依存、だます技術の高度化といった要因により、リスクのあるケースが拡大してしまったというのが現在の状況だが、技術や制度による新たな対策を作り出すことで、裏付けのあるケースが拡大されるようにするという流れになる（図 2-5）。

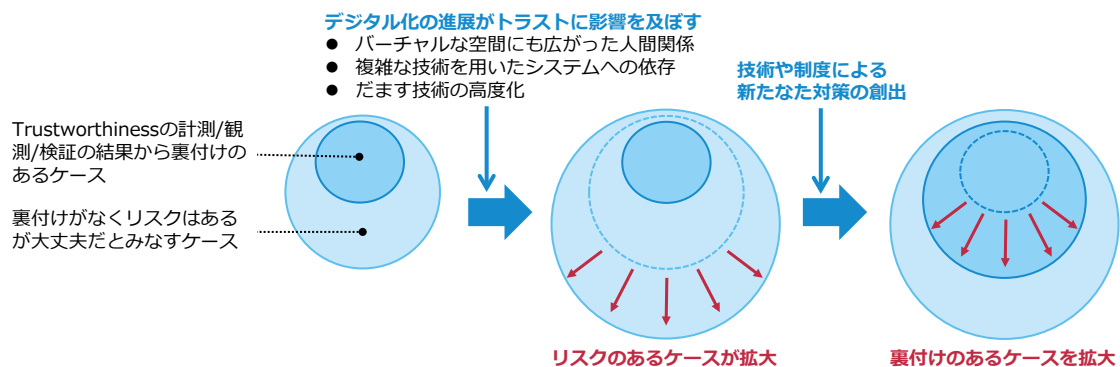


図 2-5 デジタル化の進展と対策による状態の変化

図 2-4 のモデルでも示したように、トラストは最終的に各 Trustor の主観的な判断による。しかし、詐欺のような犯罪を防ぎ、社会の秩序を維持するためには、トラストを支える「社会的よりどころ」を人々がある程度の共通認識として持つことが有効である。図 2-4 のモデルに示した裏付けのあるケースというのは、主にこの「社会的よりどころ」との照合・検証で得られる。「トラストの 3 側面」の各々について、現状の「社会的よりどころ」としてどのようなものが機能しているかを表 2-2 に例示した。しかし、デジタル化の進展に伴い、現状の「社会的よりどころ」では不十分になってきており、表 2-2 にはその問題点も併せて記載した。例えば、内容真実性について、従来は証拠写真・監視カメラ映像などが事実性判断の「社会的よりどころ」になり得たが、AI 技術によって高品質なフェイク映像が簡単に生成できるようになってしまったため、必ずしも確かなよりどころにならなくなってしまった。

表 2-2 トラストの3側面に対する「社会的よりどころ」の例と問題点

トラストの3側面	現状の社会的よりどころ	問題点
(A) 対象真正性 本人・本物であるか？	印鑑・サイン、身分証・鑑定書、 デジタル認証・生体認証など	真正性保証の対象が拡大、デジタル特有の偽造・偽装・改ざんの可能性も拡大。トラスト基 点の信頼性担保にも課題あり。
(B) 内容真実性 内容が事実・真実であるか？	事実性は証拠写真・監視カメラ映像 など、学説は査読制による学術 コミュニティ合意など	AIによるフェイク生成が高品質化したため、写 真・映像の証拠性が揺らぎつつある。そもそも 絶対的真理・事実は定まらず、ファクトチェッ ク可能な対象は限定的。
(C) 振る舞い予想・対応可能性 対象の振る舞いに対して想定・対応 できるか？	人的行為・タスクについては契約・ ライセンスなど、機械・シス テムの動作については仕様書など	ブラックボックスAIでは動作仕様が定義できず、 常にその動作を予見できるわけではない。説明 可能AIも近似的説明であり、保証にはならない。

以上を総括すると、デジタル社会においてトラストの働きがほころんできている状況（2.1.1 節）に関して、特に手を打つべき重要な問題は、「トラストの3側面」の各々について「社会的よりどころ」となるものが十分担保できていないという問題である。さらに、現状、技術的・制度的な対策が「トラストの3側面」ごとにばらばらに取り組みられており、それらの間の知見共有・連携ができていないため、トラスト問題に対して個別的な対処や断片的な状況改善にとどまっていることも問題である。

また、技術的な対策だけでなく、制度的な対策も合わせて考えるべきであることや、トラストは最終的に各人の主観的な判断によることを踏まえると、情報科学系・工学系の研究開発だけでなく、人文・社会科学系の知見も含めた「総合知」による取り組みが求められる。例えば、トラストするか否かを左右する内面的・外在的な要因にはどのようなものがあるか、不信はどのようなプロセスで生まれるか、だまされるときはどのような心理状態にあるか、社会秩序や人々の協力関係とトラストはどのように影響し合うかなど、心理学・社会学・政治学・哲学などの分野で取り組まれているトラストという行為／態度／現象の観測・理解・解明に関わる研究成果・知見は、技術的・制度的な対策を考える上で有用であることは間違いない。また、法学や経済学では、どのようなルール・法制度が社会や人々に対してどのような影響・変化を生み出すかに踏み込んだ取り組みがなされており、対策設計にもつながる。さらに、比較文化学など、国や文化的背景による社会や人々の受け止め方の違いやメンタリティーの違いを理解することは、新しいトラストの仕組みを日本国内からさらにグローバルに展開できるものか考える上で不可欠であろう。しかしながら、現状、情報科学系・工学系の技術開発が対策の中心に検討が進んでいるものの、トラストに関する人文・社会科学系の知見は必ずしも十分に生かされていないように思われる。

2.2 社会・経済的效果

本プロポーザルでは、デジタル化の進展に伴い、社会におけるトラストの働きがほころんできているという問題意識に対して、「デジタル社会における新たなトラスト形成」によって、不信・警戒を過度に持つことなく幅広い協力・取引・人間関係を作ることができ、デジタル化によるさまざまな可能性・恩恵がより広がるような社会を目指している（図 1-1）。これは、わが国が掲げている「Society 5.0」の社会像を実現するために不可欠な取り組みである。デジタル化が進展しても、トラストが成り立たなければ、図 2-1 に例示したようなさまざまな問題が発生する。フェイク・偽装・なりすましなどによる詐欺・犯罪を防ぎ、自動運転車、AI エージェント、コミュニケーションロボット、メタバースなどの複雑な技術を用いた新サービスの社会普及を促進

するなど、デジタル社会の健全な発展と安全安心（経済安全保障も含む）をもたらすために、トラストの形成が不可欠である。

トラストはビジネス関係にも不可欠で、競争要因としても重視されており、デジタル化が進展した産業・経済において、表2-3のようなさまざまな貢献を果たす。この表に示した貢献内容は、「旧来のトラスト」にも共通する面があるが、デジタル化の進展に伴い、「旧来のトラスト」だけでは貢献が成り立たなくなっており、「新しいトラストの仕組み」に置き換わることで、デジタル技術を活用したさまざまなサービス・製品において、これらの貢献が成り立ち得る。したがって、このようなさまざまな貢献が波及する関連市場も幅広い。そのごく一部ではあるが、表2-4に参考市場規模を示す。

表2-3 産業・経済の発展を支えるトラスト

トラストの貢献	貢献シーンの例
取引・協力できる相手の拡大	サプライチェーン拡大、シェアリングエコノミーなど
先端技術を用いたビジネス加速	新技術・新ビジネスに対する社会受容の促進など
ビジネス意思決定の迅速化・不安低減	経営判断、災害時・非常時判断など
消費者・取引参加者の安心	偽装・偽造リスクの低減、評判・レビューの健全化など

表2-4 波及する関連市場の一例

関連市場規模	出典	トラストの貢献
電子商取引市場（国内・2020年実績）：B2C 19.3兆円、B2B 334.9兆円、C2C 2.0兆円	経済産業省商務情報政策局：令和2年度産業経済研究委託事業「電子商取引に関する市場調査」報告書（2021年7月） https://www.meti.go.jp/press/2021/07/20210730010/20210730010.html	商取引にトラストは不可欠、問題が起さればビジネスが停止する
電子商取引市場（世界・2021年推定）：5,000億ドル	OECD Business and Finance Outlook 2019: Strengthening Trust in Business https://www.oecd-ilibrary.org/sites/35736da3-en/index.html?itemId=/content/component/35736da3-en	同上
アイデンティティ／デジタルトラスト市場（国内・2024年推定）：424.6億円	IDC Japan コラム第1回「アイデンティティ／アクセス（IAM）管理市場概況と展望」（NTTコムウェア） https://sc.nttcom.co.jp/column/iam/iam-market-overview-outlook/	デジタルトラスト（対象真正性）に関わる
メタバース市場（世界・2030年推定）：1兆-2兆ドル ※複数推定の中の一例	三菱総合研究所「メタバースの概要と展望 第2回：メタバース経済への期待と課題」 https://www.mri.co.jp/knowledge/column/20220601.html	新IT・新サービスに対するトラストが普及・利用を左右
ビジネスメール詐欺損失額（米国・2021年実績）：24億ドル	FBI Internet Crime Report 2021 https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf	今後ディープフェイク悪用詐欺の急増が予想され対策急務（内容真実性）

2.3 科学技術上の効果

本プロポーザルの実現は、トラストに関わる研究開発について、図1-2に示したBefore／Afterのような変容をもたらす。すなわち、これまでトラストの異なる側面からばらばらに取り組まれていた状況（Before）から、全体ビジョンを描いて共有し、「トラストの3側面」を多面的・複合的に扱う取り組み（After）に変容させる。これはトラスト研究の再定義であり、これによって、個別的な対処や断片的な状況改善にとどまっていたものが、デジタル社会における新たなトラスト形成へと結集し、社会に貢献する研究の道筋・インパクトも強化される。

このようなトラストに関する研究開発は、科学と社会の接点、特にデジタル技術と社会に関わる研究分野に幅広く波及する。ELSI（Ethical, Legal and Social Issues：倫理的・法的・社会的課題）やRRI（Responsible Research and Innovation：責任ある科学・イノベーション）とも表裏一体となり、さまざまな研究分野の共通基礎と位置付けられるような新たな学術領域が創出される。

この研究分野は、情報科学、工学、社会学、心理学、哲学、政治学、経済学、医学など、極めて幅広い分野にまたがる学際的な研究分野であり、第6期科学技術・イノベーション基本計画が掲げる「総合知」を活用して推進する研究分野である。したがって、この研究分野の推進は、「総合知」を活用する研究開発の推進と、それに取り組む人材育成や研究コミュニティの活性化につながる。実際、トラスト関連研究は、これまでいろいろな分野で個別に研究されていたが、2021年に企画・開催したセミナーシリーズやワークショップは、さまざまな分野のトラスト関連研究者が集まり、非常に活発な議論がなされる有意義な場となった^[1]。これをきっかけとして、さらに発展させることによって総合的なトラスト研究コミュニティの形成が見込まれる。

さらに、わが国の科学技術政策、特に情報技術や総合知に関わる研究戦略を、トラストを軸として、より強力に発信することが可能になる。現状、日本政府は、DFFT（Data Free Flow with Trust）を掲げたデータ戦略、Trusted Quality AIを掲げたAI戦略を発信している。これをさらに推し進め、「トラストの3側面」を多面的・複合的に扱う方向性を打ち出すならば、より総合的で一貫性のあるトラストを軸とした日本の研究戦略を構築でき、国際的にも先行したコンセプトを発信できる。

コラム2

トラストの弊害

トラストすることで、迅速に行動・意思決定ができ、取引や協力のコストが低減されるといった良い効果があると述べたが、トラストすることには弊害もある。

図2-4や図2-5に二重円で示した通り、トラストする状況は「裏付けがなくリスクがあるが大丈夫だとみなす」部分を含んでいる。これは、その部分について思考停止することを意味する。ただし、思考停止せずに、常にあり得る全ての可能性を考えて裏付けを取ることは不可能であり、そのリスクを気にし始めたら、全く眠れないような状態になってしまう。

二重円の内側の円の比率が十分大きくないとトラストしない慎重な性格の人がいる一方、内側の円の比率がかなり小さくても思考停止してトラストしてしまう性格の人もあるだろう。さらに、人は自分の信じたいものばかりに目を向けてしまうという「確証バイアス」を持っているので、そのような条件に当てはまるものは、あまり深く考えずにトラストしてしまいやすい。その結果として起こる「過信」「妄信」は、リスクの大きいトラストだと言える。このような状態に陥ると、悪意を持った他者にだまされやすくなるし、フェイクニュースを信じてしまいやすくなる。他者からその信じているものを否定されると、かえって頑なになってしまう「バックファイア効果」も起きやすい。

無条件に「過信」「妄信」してしまうことなく、適度な警戒・検証をしつつ、トラストの効果を享受していくのが望ましい。新しいトラストの仕組み作りにおいては、トラストの弊害を抑制することも考えておきたい。

3 | 具体的な研究開発課題

第1章の冒頭で述べたように、デジタル化の進展につれて、社会のさまざまな場面でトラスト（信頼）がうまく働かなくなってきたという状況・問題を、改善・解決するための新たな仕組みを作ることが、本プロポーザルが掲げる「デジタル社会における新たなトラスト形成」である。これによって、不信・警戒を過度に持つことなく幅広い協力・取引・人間関係を作ることができ、デジタル化によるさまざまな可能性・恩恵がより広がるような社会を目指す。

この問題認識と目指す社会の姿について、第1章では図1-1のような絵を描いた。これをトラスト関係図というレベルに抽象化したイメージ図が図3-1である。この図では、TrustorやTrusteeになる主体（人・組織・システム・Webサイトなど）が丸で表されており、それらの間に生まれるトラストが矢印で表されている。矢印は「裏付けのしっかりしたトラスト」と「裏付けの弱いトラスト」という2種類を区別して表した。図2-4中の二重円で説明すると、外側の円に対する内側の円の比率が大きいものが「裏付けのしっかりしたトラスト」で、比率が小さいものが「裏付けの弱いトラスト」を意味する。「(a) 旧来のトラストが主になっている状況」では、「裏付けのしっかりしたトラスト」が成り立っているが、関係の範囲は狭い。「(b) デジタル化の進展でリスクが高まっている状況」では、関係範囲が広がったが、「裏付けの弱いトラスト」が増えた。「(c) 新たなトラスト形成によって目指す姿」では、「裏付けのしっかりしたトラスト」の関係範囲がさらに広がり、「裏付けの弱いトラスト」の比率が減少する。

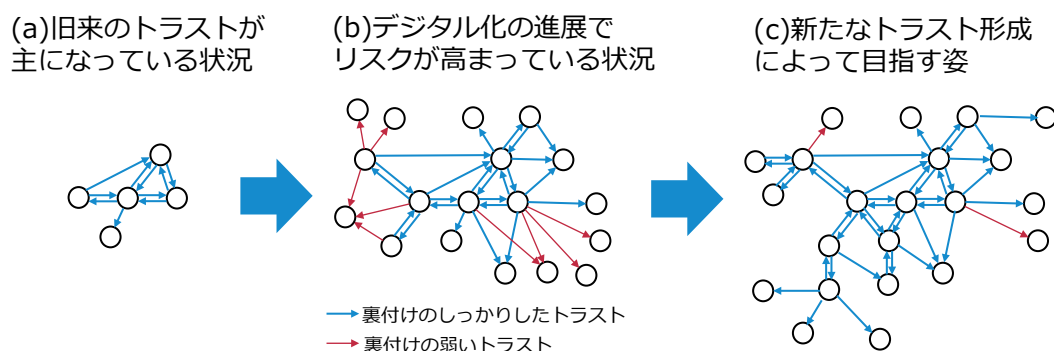


図3-1 トラスト関係の状況変化

図1-1や図3-1における(b)の状況が生まれた要因は、第1章・第2章で述べてきたように、デジタル化の進展に伴って「バーチャルな空間にも広がった人間関係」「複雑な技術を用いたシステムへの依存」「だます技術の高度化」といったことが起きたためである。そして、2.1.4節の最後に総括したように、(b)の状況において特に手を打つべき重要な問題は、以下の3点だと考える。

1. 「トラストの3側面」の各々について「社会的よりどころ」となるものが十分担保できていない。
2. 技術的・制度的な対策がばらばらに取り組まれており、トラスト問題に対して個別的な対処や断片的な状況改善にとどまっている。
3. 情報科学系・工学系の技術開発を中心に対策の検討が進んでいるが、人々の主観や社会受容が絡むトラストに関する人文・社会科学系の知見は必ずしも十分に生かされていない。

以上を踏まえ、本プロポーザルでは、図1-1や図3-1における状況(b)から状況(c)へ変えるため、図3-2に示す4層から成る研究開発課題への取り組みを連携させて推進することを提案する。4層のうち「(1) トラストの社会的よりどころの再構築」が、上記問題1・2を改善するための中核コンセプトであり、これによってトラスト関係（図3-1の矢印）を結んでよい相手か否かの判断がしやすくなる。さらに、4層を連携させて研究開発を進めることが、上記問題2・3を改善するとともに、社会に価値をもたらす研究開発の道筋を与えることになる。

以下、4層の研究開発課題のそれぞれの内容を示す。

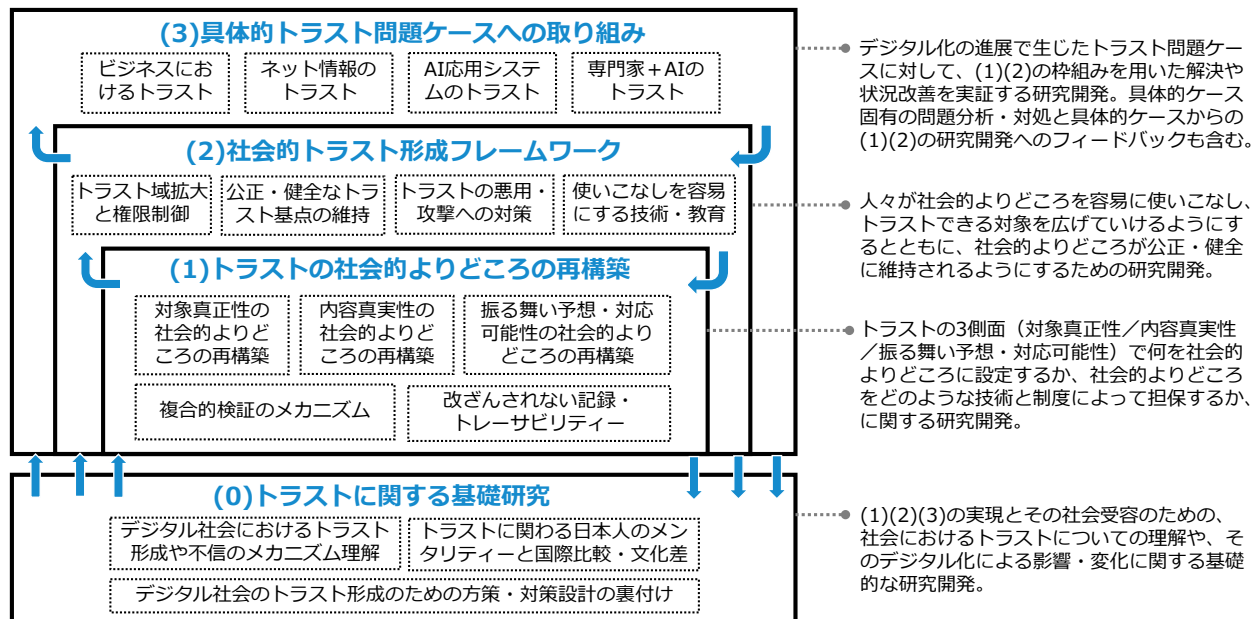


図3-2 4層から成る研究開発課題の構成

3.1 トラストの社会的よりどころの再構築

第1層「トラストの社会的よりどころの再構築」は、トラストの3側面（対象真正性／内容真実性／振る舞い予想・対応可能性）で何を社会的よりどころに設定するか、社会的よりどころをどのような技術と制度によって担保するか、に関する研究開発である。図3-2には、この研究開発課題をより細分化した例として「対象真正性／内容真実性／振る舞い予想・対応可能性それぞれの社会的よりどころの再構築」「複合的検証のメカニズム」「改ざんされない記録・トレーサビリティ」を挙げた¹。以下、それぞれについて説明する。

- 対象真正性／内容真実性／振る舞い予想・対応可能性それぞれの社会的よりどころの再構築：**これは表2-2に示した問題点の解決・対処を目指す取り組みである。すなわち、「トラストの3側面」のそれぞれについて、従来も「社会的よりどころ」になるものが用意されていたが、デジタル化が進展した状況に

¹ 図3-2において、4層の研究開発課題(0)(1)(2)(3)のそれぞれについて、それをさらに細分化した例を示したが、これは一例であり、細分化した課題として考えられるものが、これらに限定されるものではない。

において、従来のよりどころ（表2-2にあるように例えば身分証・鑑定書・監視カメラ映像・仕様書・契約書など）だけでは十分ではなくなってきたので、「社会的よりどころ」となるものを強化・再構築することの必要性が高まっている。これは何か新しい証明書を用意すればよいという簡単なものではなく、何に対して誰がどのように保証するのか、それがどのような仕組みで検証できるのか、それがどの程度まで確かな保証だとみなせるのか、社会的に通用するものになるのか、など十分に吟味される必要がある。このような「社会的よりどころ」が拡充されると、Trustorにとって、相手（Trustee）をトラストできるか否かの判断をよりの確にできるようになり、だまされてしまうリスクを減らすことが可能になる。

- **複合的検証のメカニズム：** 上記で「社会的よりどころ」の強化だけでなく再構築の必要性にも言及しているのは、「社会的よりどころ」に対する考え方の変化が念頭にある。この点に関わるのが「複合的検証のメカニズム」である。表2-2に示した「現状のよりどころ」は、もともとそれが一つあればトラストして大丈夫と思えることが期待された。しかし、デジタル化で生じた問題点や、コラム1・コラム3でも触れたような100%保証の困難性を踏まえると、対象のある一つの側面や断片的に切り取られた情報だけから何かを信じ込むことはとても危うい。その一方で、一つ一つは100%保証ができなくとも（もちろん一つ一つのよりどころの保証の確かさはできる限り高めるべきである）、多面的に取得した複数通りの情報を、複合的に検証し、総合的に判断するならば、その危うさ（リスク）は軽減できる。そこで、「社会的よりどころ」に基づく検証は複合的に扱い、総合的に判断することが容易なメカニズムを備えることが望ましい。なお、多面的・複合的に収集・検証するというのは、「トラストの3側面」を合わせて扱うことと、1つの側面について複数通りの情報を扱うことの両方を含む。このような多面的・複合的検証に合わせて「社会的よりどころ」を体系化・再構築することが必要になると考える。
- **改ざんされない記録・トレーサビリティ：** 対象真正性については改ざんされていないことを検証する必要があり（コラム1参照）、振る舞い予想・対応可能性については振る舞いの履歴を検証する必要がある（コラム4参照）。そのために、改ざんされない記録とトレーサビリティの担保が求められる。ブロックチェーン技術が活用可能であろうし、必ずしもブロックチェーンのような分散型技術で保証しなくとも、改ざんできないように厳密な運用体制を作るというアプローチもあり得る。

3

具体的な研究開発課題

コラム3

内容真実性：事実や真実は定まるものなのか？

トラストの一つの側面として「内容が事実・真実であるか？」を、本プロポーザルでは「内容真実性」と呼ぶ。だが、これは、100%間違いがないと保証できる絶対的な事実・真実が定められるという性質のものではない。図3-3に示すように、事実性は、ある発生した事象に対して、観測された結果が観測装置の特性・精度や観測者の認知バイアスに左右されるため、複数の観測でその結果・報告が異なるということが起こり得る。さらに、さまざまな事象についての観測された結果の集合から解釈・発見したことを研究論文として投稿し、他の専門家たちからの査読を経て採択されると、一定の確からしさを持つ言説として認められるが、その言説は絶対的な真実として認められたわけではない。科学研究の世界では、採択された研究論文の言説が後続研究によって否定されたり、修正されたりすることはたびたび起こる。

だからと言って、内容真実性を問うことが無意味なはずはない。裁判において事実認定は不可欠であるし、フェイクニュースに対するファクトチェック活動も社会的意義が大きい。情報内容の真偽、すなわち内容真実性は、絶対的なものではなく、ある種の社会的合意と考えるのが妥当であろう。それゆえ、情報を出す側は、単に真偽を表記・主張するだけでなく、その根拠が何か、その確かさがどの程度でどのような検証が行われたのか、といったメタ情報も併記したり、簡単に確認できるようにしたりすることが望ましいし、情報を受ける側は、一つの手がかりだけで真偽を決めつけず、多面的・複合的に判断するようにした方がよい。

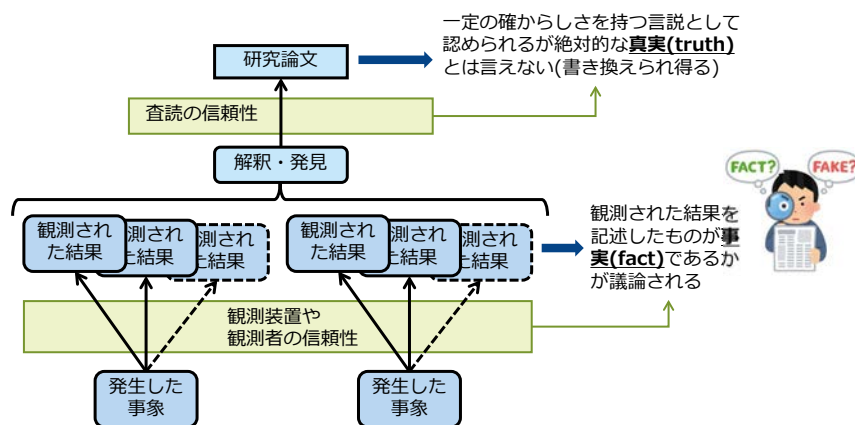


図3-3 内容真実性に関わる要因

コラム4

AI 応用システムの振る舞い予想・対応可能性

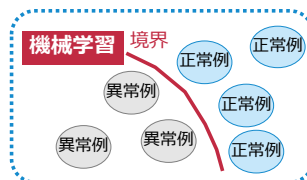
現在のAI 応用システムには、主に深層学習などの機械学習技術が用いられている。従来のプログラミングによるシステム開発は演繹型であるのに対して、機械学習によるシステム開発は帰納型である^[10]。すなわち、プログラミングでは、手続きや判定ルールを明示的に書くことでシステムの動作が決まるのに対して、機械学習では、データを例示すると、それをまねて自動的に判定ルールができ、システムの動作が決まる（図3-4）。

プログラミングによる
演繹型システム開発

```
if 条件a then 処理A
else 処理B
```

条件a: 温度センサー > 90℃
(異常と判定する条件)

手続きや判定ルールを明示的に書くことでシステムの動作が決まる

機械学習による
帰納型システム開発

データを例示すると、それをまねて自動的に判定ルールができ、システムの動作が決まる

図3-4 演繹型システム開発と帰納型システム開発

演繹型システム開発では、どのような条件・状況のときにシステムがどのように動作するかが、プログラム（コード）として明確であり、仕様書としても記述される。したがって、基本的には仕様書の通りに動作するとみなすことができ、仕様書は振る舞い予想・対応可能性のよりどころになる。

しかし、帰納型システム開発では、そのような仕様書がなく、どのような条件・状況のときにシステムがどのように動作するか、起こり得るすべてのケースを事前に知ることは原理的に不可能である。このとき何が振る舞い予想・対応可能性のよりどころになり得るかが課題になっている。システムのテストをできる限り網羅的・効果的に行う手法の開発、そのような事前のテストが十分行われたかを含む開発プロセスの審査、事前にすべての可能性を押さえられない代わりに事後対策としての責任の明確化や保険の設計、それらの基礎情報としてのシステム動作履歴の記録・検証などが検討されている。なお、演繹型システム開発においても、大規模化・複雑化したシステムには、同様の課題が存在する。

また、説明可能なAI（XAI）の技術開発も行われているが、XAIによって得られるのは近似的な説明なので、説明から外れるシステム動作は発生し得る。したがって、XAIはシステム動作の傾向理解に役立つが、動作保証を与えるものではない。

3.2 社会的トラスト形成フレームワーク

第2層「社会的トラスト形成フレームワーク」は、人々が「社会的よりどころ」を容易に使いこなし、トラストできる対象を広げていけるようにするとともに、「社会的よりどころ」が公正・健全に維持されるようにするための研究開発である。図3-2には、この研究開発課題をより細分化した例として「トラスト域拡大と権限制御」「公正・健全なトラスト基点の維持」「トラストの悪用・攻撃への対策」「使いこなしを容易にする技術・教育」を挙げた。以下、それぞれについて説明する。

- トラスト域拡大と権限制御： 従来のデジタルトラストでは、対象真正性を支えるトラストの基点（トラストアンカー）として、認証局（Certification Authority：CA）とそれが発行するデジタル証明書が位置付けられ、そこからトラストチェーンをつないでいくことで、トラストできる対象範囲（トラスト域）を拡大する。このようなトラストアンカー&トラストチェーンの考え方は基本的なものであり、「社会的トラスト形成フレームワーク」の設計においても有効と考える。世界経済フォーラムの「Rebuilding Trust and Governance: Towards Data Free Flow with Trust (DFFT)」白書^[2]においても、基本的な考え方として採用されている。これを対象真正性だけでなく、内容真実性や振る舞い予想・対応可能性の側面も含めた「複合的検証のメカニズム」と組み合わせて発展させる。また、対象（Trustee）がトラストできたときに、その対象にどのような権限（データへのアクセスや何らかの操作を許すかなど）を与えるかについても、適切な制御が行えるように考えていく必要がある。
- 公正・健全なトラスト基点の維持： トラストの基点は、何が本物か、何が事実かといった保証を与えるが、これが権力者や特定勢力に支配されてしまうと、彼らに都合が良いように本物や事実とするものを定めることが可能になってしまう。そのような事態を招かないように見張るための分散監視・相互監視、および、透明性確保の仕組みを考える必要がある。
- トラストの悪用・攻撃への対策： トラストの「社会的よりどころ」の構築は、TrustorがTrusteeをトラストできるか否かを、よりの確かつ容易に判定できるようにしようとするものであるが、その外側から、トラスト関係を悪用しようとする行為や、トラスト関係を攻撃する行為などが行われる可能性がある。例えば、親密さを演じつつ詐欺を狙う行為、良好なトラスト関係が築かれている状態に対してフェイクなどを用いて棄損する行為などが考えられる。これへの対策の一例として、その類の行為を検知するとか、行動履歴を分析して疑わしい者をスコアリングするなどが考えられる。
- 使いこなしを容易にする技術・教育： トラストの「社会的よりどころ」が強化・再構築されたとしても、人々がそれらを使いこなせなければ、社会において新たなトラストがうまく機能しているとは言えない。そこで、トラストできるかを判断する相手・対象について、「社会的よりどころ」による検証結果を簡単に得られるシステムの使い勝手や、その結果の理解や使い方についての教育・啓発なども重要である。

ここで、第1層「トラストの社会的よりどころの再構築」と第2層「社会的トラスト形成フレームワーク」を合わせた仕組みについて、その適用前後（Before/After）を対比したおおまかなイメージ図を図3-5に示す。この図の左側（Before）では、従来の「社会的よりどころ」に基づいているものの、対象真正性、内容真実性、振る舞い予想・対応可能性のそれぞれで疑念が生じていることを示している。

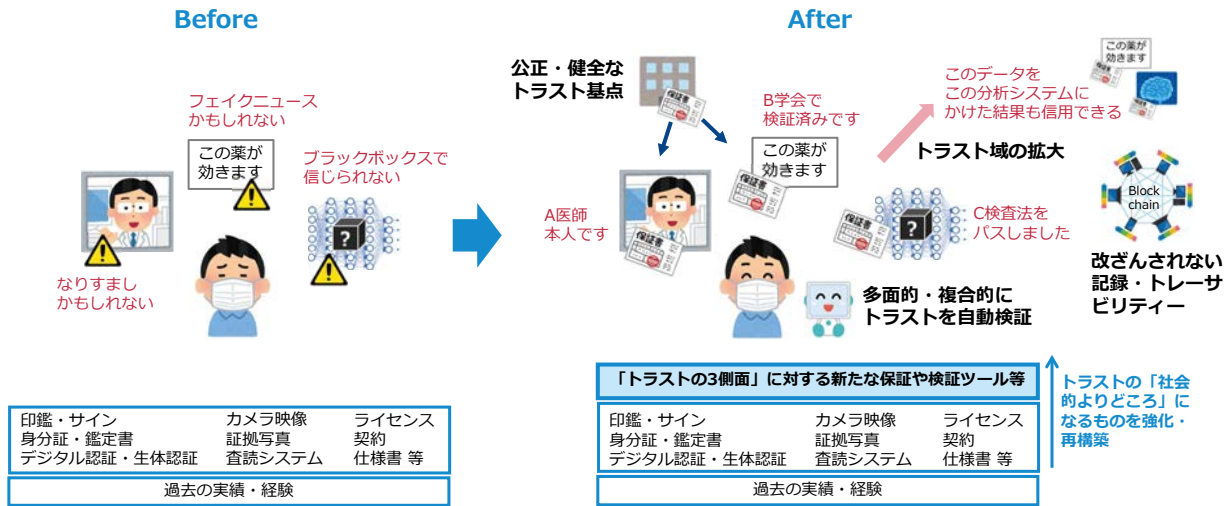


図 3-5 デジタル社会のトラストの仕組み（Before/Afterのイメージ図）

図の右側（After）では、3.1節・3.2節で述べた対策が組み込まれた仕組みによって、状況が改善される様子（実現できた場合の期待）を示している。まず、「トラストの3側面」について「社会的よりどころ」となる新たな保証や検証ツールが拡充される。それらを用いることで、図中にあるように、疑念が生じていた3点についての保証が与えられるようになる。図中では複雑な表現を避けて簡略化したが、実際には3側面それぞれで複数の観点から複合的な検証が行われる。図中のように3つ程度であれば、総合的な判断を人間が行うことも簡単だが、その数が増えると、多数の検証結果から総合的な判断を行うことは、人間にとって必ずしも容易なことではない。多面的・複合的な自動検証をサポートするAI エージェントのようなシステムも有用になってくる。また、そのような検証を通して、トラストできる可能性の高い人・組織、データや分析法などが分かったならば、それを手がかりにして、トラストできる対象範囲を広げていくことも可能になる。そのような仕組みを支えるものとして、公正・健全なトラスト基点や改ざんされない記録・トレーサビリティといったものも重要なことは既に述べた通りである。

なお、図3-5では、人を中心に置いて、他の人やシステムや情報に対するトラストを扱ったが、多面的・複合的な自動検証が可能になるならば、それをシステム-システム間で実行し、自動的にトラストチェーンをつなぎ、トラスト域を拡大することも考えられる。

3.3 具体的トラスト問題ケースへの取り組み

第3層「具体的トラスト問題ケースへの取り組み」は、デジタル化の進展で生じたトラスト問題ケースに対して、第1層・第2層の枠組みを用いた解決や状況改善を実証する研究開発である。具体的ケース固有の問題分析・対処と具体的ケースからの第1層・第2層の研究開発へのフィードバックも含む。図3-2には、この研究開発課題をより細分化した例として「ビジネスにおけるトラスト」「ネット情報のトラスト」「AI 応用システムのトラスト」「専門家+AIのトラスト」を挙げた。表3-1には、そのそれぞれについて、現状の関連する取り組みに触れるとともに、本プロポーザルで目指す姿の具現化のために考えられる発展の方向性を示した。細分化したターゲット例は、これら4つに限定されるものではないが、図2-1の8つのトラスト問題ケースを踏まえつつ、「トラストの3側面」に対する多面的・複合的な検証に発展させるシナリオとして、3側面のそれぞれから発展させるシナリオ3つと、3側面以外にも重要な要因に関わるシナリオ1つを設定したものである。

表3-1 具体的トラスト問題ケースへの取り組みのターゲット例

ターゲット例	内容・方向性	該当する図2-1の例
ビジネスにおけるトラスト	現状、デジタル庁のサブWGやTrusted Web推進協議会では、サプライチェーンやワークフローにおける企業や個人の対象真正性の確保を中心としたトラストフレームワークの設計を進めている。その発展として、さらに内容真実性や振る舞い予想・対応可能性を取り込む方向性が考えられる。	(a) 仮想世界のトラストに基づく取引 (h) メタバース内活動におけるトラスト（ここでは経済活動面に着目）
ネット情報のトラスト	フェイク問題が深刻化するSNSなどネット情報空間における内容真実性について、100%の真実・事実は確保し得ないが、一定の裏付けを持つものとそうでないものが区別される仕組み・運営体制が望まれる。その中では、対象真正性や振る舞い予想・対応可能性の側面も複合的に扱うことが必要になる。	(b) メディアにおけるフェイク拡散
AI応用システムのトラスト	Trustworthy AIに関する標準化や適合性評価（認証）によって、AI応用システムの振る舞い予想・対応可能性を一定程度確保する取り組みが進みつつある。AI応用システム・サービスの中では、振る舞い予想・対応可能性に加えて、対象真正性や内容真実性も複合的に扱うことが必要になってくる。	(c) 自動運転車 (d) パーソナルAI エージェント (e) 人を評価するAIシステム (g) コミュニケーションロボット
専門家+AIのトラスト	専門家（人）vs. 専門家AIが起こり得る状況において、ユーザーを含めた3者関係をどのような形にしていけるか。特に医療意思決定の現場で検討が進められており、トラストの3側面だけでなくHAI（Human-Agent Interaction）的な要因も含めて、3者関係におけるトラスト構築を検討する方向性になる。	(f) 医療意思決定におけるAI セカンドオピニオン

なお、「ネット情報のトラスト」と「AI応用システムのトラスト」について、情報科学技術の研究開発として取り組むためのJST CRDSからの戦略プロポーザルは、過去2018年に発信している²。しかし、今回は情報科学技術の研究開発として具体的な社会課題にフォーカスした取り組みだけでなく、制度設計による対策や社会受容の観点も深く配慮した「総合知」による取り組みに広げ、デジタル社会のトラスト問題に対する共通的な枠組みを志向している。

また、トラストは産業上も重要であることは2.2節で述べた通りであるが、産業界へのインタビューから、トラストを確保するためのプロセス（手順）が求められていることが分かった。従来、システム開発においてバグを完全になくすことは困難だが、それをできる限り減らすために、どのようなプロセスを経るのが望ましいかという方法論が作られてきた。デジタル社会における新しいトラストの仕組みにおいても、同じようなものが求められることはよく理解できる。そこでは図2-4のモデルにおけるTrustworthinessの検証プロセスの妥当性が主要な論点となり、このプロセスに対する社会からのトラスト（いわば「プロセスのトラスト」）を確保することが必要になる。その際に、求められるトラストのレベルや検証すべきTrustworthinessの種類は応用依存と考えられるので、上で述べたような具体的なトラスト問題ケースでは、それぞれにおいて「プロセスのトラスト」をどのように確保するかという点も、重要な課題である³。

- 2 2018年にJST CRDSの戦略プロポーザルとして、「ネット情報のトラスト」に対しては「複雑社会における意思決定・合意形成を支える情報科学技術」^[17]、「AI応用システムのトラスト」に対しては「AI応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」^[10]を発行している。これらの提言に沿った推進策として、JST CREST「信頼されるAIシステムを支える基盤技術」（2020年度～）、JST さきがけ「信頼されるAIの基盤技術」（2020年度～）、NEDO「人工知能の信頼性に関する技術開発事業」（2019年度実施、2020年度からは「人と共に進化する次世代人工知能に関する技術開発事業」に発展・移行）などのプログラムも実施されている。
- 3 そのような取り組みの一例として、「AI応用システムのトラスト」の分野では、AIプロダクト品質保証（QA4AI）コンソーシアムにおいて、自動運転、AI-OCR、産業用プロセス、Voice User Interface、生成系システムという具体的なユースケース毎に、品質保証のためのガイドライン（品質保証のための指標・手順・チェックリストなどを含む）^[18]がまとめられている。

3.4 トラストに関する基礎研究

第0層「トラストに関する基礎研究」は、第1層・第2層・第3層の実現とその社会受容のための、社会におけるトラストについての理解や、そのデジタル化による影響・変化に関する基礎的な研究開発である。図3-2には、この研究開発課題をより細分化した例として「デジタル社会におけるトラスト形成や不信のメカニズム理解」「トラストに関わる日本人のメンタリティーと国際比較・文化差」「デジタル社会のトラスト形成のための方策・対策設計の裏付け」を挙げた。表3-2には、そのそれぞれについて、より具体的な内容の例を示した。

表3-2 トラストに関する基礎研究のターゲット例

ターゲット例	より具体的な内容の例
デジタル社会におけるトラスト形成や不信のメカニズム理解	- デジタル社会における情報の非対称性とその発生・緩和要因 - デジタル社会における不信のメカニズムとその要因 - リアルな人間とは異なる対象（システム／デジタル情報／AI エージェント／アバターなど）に対するトラストや不信の特性 - トラスト形成における言語情報・非言語情報の作用 - 進化論・発達科学の視点でのトラストの起源とデジタル環境変化の影響 - リテラシーの副作用、おおむね確かなメディア作り（ネットメディアのガバナンス） - 情報空間・フェイクニュースの生態系 - フェイクニュースを作りネットで拡散する人間の心理 - デジタル技術と政治の共犯 - パブリックトラスト／プライベートトラスト、刹那的トラスト／継続的トラスト、見知らぬ人とのトラスト／知り合いの中のトラスト
トラストに関わる日本人のメンタリティーと国際比較・文化差	- 日本人の主体性なきパターンリズム（決めてくれるのを待つ傾向）：人権や公平性などの価値観（正義）を掲げる欧米と比較して曖昧な日本の価値観 - 主体的判断によるトラスト vs. 制度に支えられたトラスト - AI のトラスト：道具型 AI として押さえ込む vs. 自律型 AI としてのトラスト
デジタル社会のトラスト形成のための方策・対策設計の裏付け	研究開発課題の第0層と第1～3層との間の共同検討：上記2つのターゲット例に該当するようなトラストの観察・理解に関する研究成果が、第1～3層の取り組み（技術や制度による方策・対策）の裏付け・根拠として活用されるようにする

第一に、トラストの「社会的よりどころ」や新たなトラストの仕組みが社会においてうまく働くのかを予見したり評価したりして、その社会受容性や効果を判断していくための基礎研究として、「デジタル社会におけるトラスト形成や不信のメカニズムの理解」は重要である。トラスト形成に関わる要因、あるいは、逆にトラストを失う／得られない不信の要因を知り、また、それらがデジタル化の進展によってどのような影響・作用を受けているのかを観察・理解することは、デジタル社会における新たなトラストの仕組みを実現していくために役立つ。

第二に、トラストに対する考え方や受け止め方は、国の価値観や文化にも依存するので、「トラストに関わる日本人のメンタリティーと国際比較・文化差」を理解しておく必要がある。日本人のメンタリティーを踏まえた社会受容を考えるとともに、海外・他国との違いも意識しつつ、トラストを支える仕組みとして国際的に共通なものにできる部分と、日本固有のものになる部分を押さえておくことも必要である。

これら二つについては、これまで実施してきたセミナーシリーズやワークショップ（付録Aを参照）から示唆された研究開発課題や検討すべき切り口を、表3-2中の「より具体的な内容の例」に挙げた。ただし、第0層「トラストに関する基礎研究」の研究開発課題がこれらに限定されるわけではない。2.1.4節の最終段落にも例示したように、心理学・社会学・政治学・哲学・法学・経済学・比較文化学など、さまざまな人文・

社会科学分野の研究成果・知見の貢献がこの基礎研究に期待される。

さらに、表3-2における第三点として、「デジタル社会のトラスト形成のための方策・対策設計の裏付け」を挙げた。これは上記の二つと第1層から第3層の取り組みをつなげることを意図している。図2-3に示した整理に当てはめるならば、第1層から第3層の取り組みは「技術開発による対策設計」と「ルール整備・プロセス管理による対策設計」に該当し、第0層は「対策設計・社会受容の裏付け」に該当する。第0層は「トラストの観察・理解」に軸足を置いた人文・社会科学系の取り組みが中心になり、かつ、必ずしもデジタル技術を前提とした研究開発でなくともよいが、第1層から第3層の取り組みの裏付けになってこそ意味がある。既公開の報告書^[1]や付録Bに示すように、「トラストの観察・理解」に関する研究は古くから取り組まれており、そのさらなる発展とともに、デジタル社会のトラスト形成のための方策・対策設計の裏付けに資することが望まれる。そのためには、第0層の取り組みと第1層・第2層・第3層の取り組みとの間の連携・共同検討に踏み込むことが有効であろう。例えば計算社会科学や行動経済学など、人文・社会科学系と情報科学系・工学系を横断した研究分野もあり、分野横断の連携はいっそうの活発化が期待される。

なお、これら第0層の研究開発課題は人文・社会科学系の取り組みが中心になるが、人文・社会科学系の貢献は第0層だけでなく、第1層・第2層・第3層においても期待される。

コラム5

デジタル化とトラスト問題の関係

本プロポーザルは、デジタル化の進展によって新たに生じたトラスト問題やより深刻化したトラスト問題にフォーカスしている。ここでは、一つの見方として、図2-4のトラストモデルをベースに、問題の捉え方や対策の位置付けについて整理してみたい。

図3-6では、図2-4のトラストモデルにおいてデジタル化の影響を受ける点として、次の2点を挙げた。

第一に、Trustor－Trustee間インターフェイスにデジタル技術が介在する。具体的には、相手（Trustee）がネットの向こう側にいたり、相手がAIシステムだったり、メタバース内で相手がアバターだったりする。その際、相手は本人でない（なりすまし）かもしれないという不安を持ったり、相手の表情・非言語情報が分からなかったりするといったことがトラスト形成にも影響すると思われる。

第二に、Trusteeの品質特性であるTrustworthinessの計測や観測や検証、および、その妨害やかく乱にデジタル技術が使われる。そこでは、フェイクの画像・音声・映像かもしれない、出典・根拠の怪しい情報かもしれない、偽造・改ざんされた本人証明や本物証明かもしれない、偏った面しか見ていない／見せられていないかもしれない、巧みな会話と表情分析を用いて誘導されているかもしれない、テストが手抜きでシステムが思いもよらない誤動作をするかもしれない、などといった懸念が生じ得る。

TrusteeのTrustworthinessの計測や観測や検証の結果は、Trustorにとって、Trusteeをトラストできるかの判断材料になる。そして、本プロポーザルの第3章に

示した4層の研究開発課題、その中でも特に第1層・第2層は、この判断材料を多面的に拡充し、人々が簡単に複合的検証を行うことができ、妨害・かく乱もされにくくすることに、技術的・制度的な対策の中心がある。そのためには「トラストの3側面」に対する「社会的よりどころ」を拡充することがまず必要になるが、それだけでなく、それらのトラストの基点を適切に担保し、どのようなTrustworthinessについてどのような計測・観測・検証のメカニズムを用意するかを定め、トラストできる対象を評価して広げていく手順を定めることなども含むフレームワーク全体を設計していくが必要になる。

なお、デジタル化の進展によって新たに生じたトラスト問題やより深刻化したトラスト問題への対策として得られたものが、デジタル化以前の問題にも有効に適用できるケースもあるであろう。

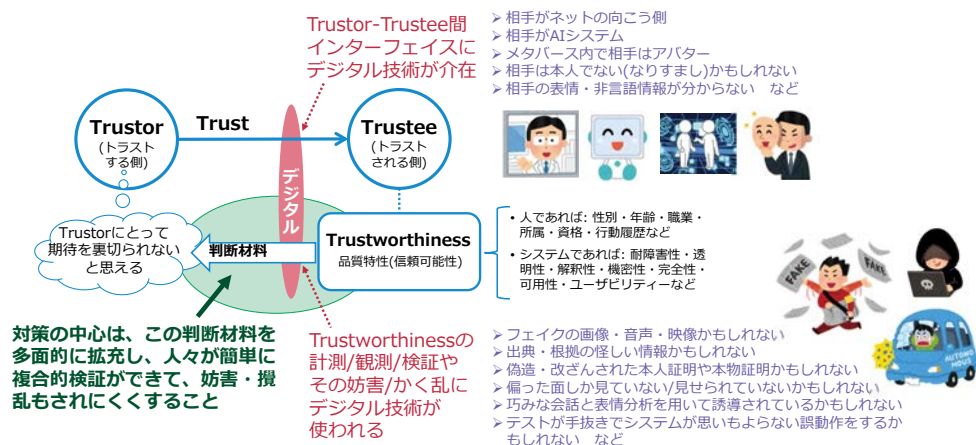
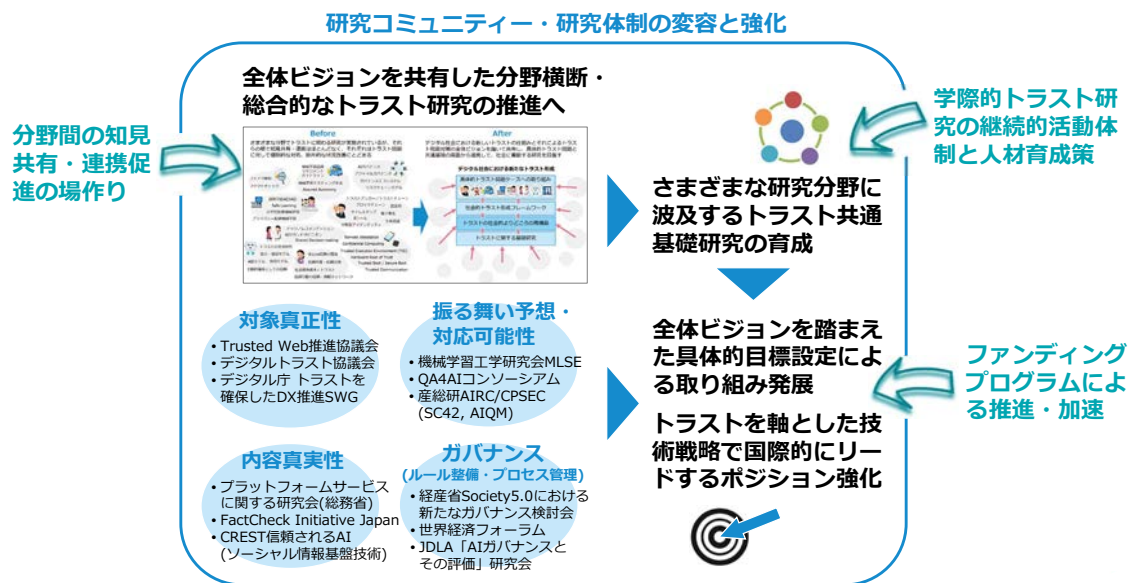


図 3-6 デジタル化の影響と対策の中心

4 | 研究開発の推進方法および時間軸

第1章の図1-2に、トラスト研究開発の目指す姿を Before / After の対比によって示した。すなわち、これまでいろいろな分野で個別に取り組まれていたという状況から、デジタル社会における新しいトラストの仕組みとそれによるトラスト問題対策の全体ビジョンを描いて共有し、具体的トラスト問題と共通基礎の両面からの取り組みを連携して進める形への変容を目指す。



これを実現するために考えられる推進方法の全体像を図4-1に示す。この図の中央の青枠内は、研究コミュニティや研究体制がどのような形に変容するのが望ましいかについての考えを示した。さらに、その変容を促進するための方策として考えるべきポイントを、青枠部分への3つの緑矢で示した。以下、その内容を順に説明する。

【方策1】分野間の知見共有・連携促進の場作り

図1-2に描いたように、全体ビジョンを共有した分野横断・総合的なトラスト研究という方向性について、研究コミュニティから幅広い共感を得られることが、活発な研究開発の出発点になる。実際、2021年に実施したセミナーシリーズとワークショップ^[1]では、幅広い分野の研究者間で活発な議論がなされ、このような機会は有意義だという意見が多数寄せられた。また、本プロポーザルの検討過程でインタビューした研究者からもその方向性や意義について賛同いただいた。

研究者個人の人脈や出会いを通してボトムアップにリーチできる範囲には限界があることから、分野間の知見共有・連携促進の場作りにはトップダウンな施策が望まれる。その施策の一例としては、上記の2021年のワークショップなど（その際はクローズドなメンバーで実施した）をより発展させた形態として、オープンに参加者を集めた公開型のセミナーシリーズ／ワークショップ／シンポジウムなどを開催することが、研究コミュニティを立ち上げ、活性化するために、より有効と考えられる。

このような場が、全体ビジョンの共有、幅広い研究事例の把握、分野横断の共同研究のタネの発見の機会となり、図1-2に描いた変容や、分野横断・学際的な研究コミュニティの活性化が期待できる。

【方策2】学際的トラスト研究の継続的活動体制と人材育成策

本プロポーザルに示したような取り組みは、さまざまな研究分野に波及するトラスト共通基礎研究としても位置付けられる。この学際的な研究を骨太化し、研究者層を厚くしていくためには、方策1のような仕掛けをトリガーとしつつも、学際的トラスト研究の継続的な活動体制を作っていくことが必要である。そのような研究活動の母体ができることで、研究人材の育成にもつながる。具体的には、学会・研究会の立ち上げ、このような学際的活動をコーディネートする人材やチームへの助成、学際的トラスト基礎研究を推進する拠点あるいはバーチャルな連携体制の構築などが、打ち手の候補として考えられる。

また、新しいトラストの仕組みの社会受容性の面では、日本社会の特性や日本人のメンタリティーを踏まえることが重要である一方、国際社会や他国との違いを最初から意識しておくことも重要である。加えて、米国IT企業がインターネット上のプラットフォームをほぼ支配している現状を考えると、「デジタル社会におけるトラスト形成」において、海外動向や海外から見た日本社会という視点から分析・考察できる人材の育成も必要と考える。

なお、分野横断・文理融合的な「総合知」に取り組むことに対する阻害要因が指摘されている。阻害要因の一例として、(1)テニユアポストが少ない、(2)「総合知」への取り組みは研究論文になりにくい一方、任期付きポストでは論文業績が必要、(3)個別分野で業績を上げながら、「総合知」に取り組まなければならないという高いハードルが現実、といったことが挙げられる。これはトラスト研究に限った問題ではなく、根本的な打ち手は簡単ではないが、方策2の中でも考慮しておく必要がある。

【方策3】ファンディングプログラムによる推進・加速

方策2によってトラスト共通基礎研究を育成する一方で、より具体的な問題解決型の目標を設定し、それをファンディングプログラムによって推進・加速することも、重要な打ち手になる。3.3節に示した研究開発課題の第3層の検討がそのベースとなり、複数のターゲット（ファンディングプログラムの候補）設定があり得る。ただし、個別の問題解決にとどまらずに4層から成る全体ビジョンを踏まえた共通化も意識することが重要である。

この方策3に関連する動きとして、図4-1の青枠内の左下パートに、既に進行している国の戦略に基づく活動体制、ファンディングによる活動、関連する政策提言活動を示した。これらは前述の表2-1からの抜粋で、表2-3に合わせ、対象真正性、内容真実性、振る舞い予想・対応可能性、ガバナンス（ルール整備・プロセス管理）という現在のスコープに沿ってまとめている。今後これらの活動について、そのスコープを拡大したり、取り組みを連携させたりすることで、本プロポーザルが描くような方向への発展が期待できる。

このようなファンディングによって、トラストを軸とした技術戦略で、わが国が国際的にリードするポジションを獲得することも狙い得る。また、これに関連して、欧米が人権や公平性などを基本に置いた理念・正義を押し出して、国際的なルールメイキングを先導している状況に対して、わが国はトラストを軸とした立ち位置が取り得るであろう。

方策1～3の時間軸

方策1は2022年から取り組んでいくとともに、方策2も模索し、1・2年程度で方策2に重点を移行させる。

方策3も方策1を通して具体的な目標設定を進めるのがよい。優先度の高い目標を中心に複数のファンディングプログラムを立ち上げることが可能であろう。競争的研究費の公募を早期に行った場合、図4-1の青枠内の左下パートに示したような情報系が中心になった研究開発に、人文・社会科学系が協力するような形態での応募が多くなると思われる。一方、方策1・2を通じた議論・連携が深まった時期になれば、人文・社

会科学系の存在感がより高まった研究開発の比率が増加すると期待される。しかし、それには長期的な時間軸を見込んでおく必要がある。わが国が総合的なトラスト研究開発で先行し、技術戦略においても国際的にリードするポジションを狙うためには、迅速な打ち手も必要であろう。まずは方策1を実行し、その中で方策3の目標とタイミングを見定めていくのがよいと考える。

図4-2には、第3章で説明した4層から成る研究開発課題の推進と3方策の関係を、時間変化も含めて示した。現状は、表2-3や図4-1に示したような活動において、トラストのある側面に限って各層の間に部分的な縦の関係ができていくのみであり、方策1によって、関係を横に広げ、分野横断の知見共有や連携を促進する。その上で、方策2は第0層・第1層を中心としたトラスト共通基礎研究を強化する。さらに、方策3は明確な目標を掲げ、課題解決に踏み込んで、第0層から第3層までを統合した取り組みを推進する。また、推進形態として、方策3は具体的な目標を掲げた有期プロジェクトが想定されるのに対して、方策2は共通基礎研究としての継続的な育成が望ましいであろう。

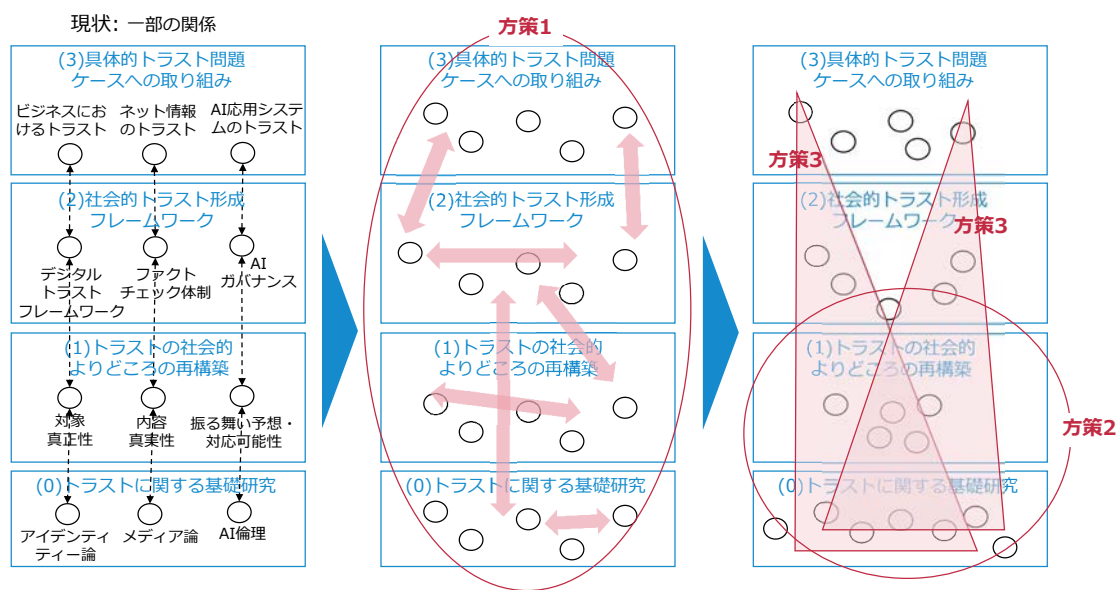


図4-2 4層の研究開発課題の推進と3施策の関係

推進体制については、AI・自律システムにフォーカスしてはいるが、英国研究・イノベーション機構（UKRI）の「TAS (Trustworthy Autonomous Systems) Programme」（付録B.2を参照）が情報系と人文・社会科学系と産業界が参画した強力なフォーメーションを組んでいることが注目される。しかし、本プロポーザルに示したような、TASと比較してさらに広いスコープで、わが国がTAS同様の強力なフォーメーションを組むことができるかは、大いにチャレンジングな課題である。方策3と併せて、強力なフォーメーション構築に向けたステップも設計していく必要がある。

また、「デジタル社会における新たなトラスト形成」は、ビジネスの発展やDX（デジタルトランスフォーメーション）を左右するものであり、産業界からの期待・関心も大きい。人々にトラストされるサービス・製品を提供していくことは、企業の使命・責任だとも言える。それと同時に、新たなトラストの仕組みが人々に活用され、社会に浸透していくかは、国の政策とともに企業の取り組みが担うところも大きい。上で言及した方策3のフォーメーションはもちろん、方策1のようなより早期のステージから学术界だけでなく産業界も参画した形で進め、産学でビジョン・方向性を共有した取り組みが望まれる。

付録

付録A 検討の経緯

国立研究開発法人科学技術振興機構（JST）研究開発戦略センター（CRDS）では、2021年度の戦略スコープ策定委員会において、戦略プロポーザルを作成すべきテーマの候補として、本テーマを選定し、検討チームを発足させた。本検討チームは2021年6月から活動を開始し、2022年8月まで検討を進めてきた。その間、有識者へのインタビューや、有識者を招いてのセミナーシリーズやワークショップを開催し、本テーマの研究開発状況の把握や研究開発課題・方向性の議論を進めてきた。以下に、その概要を示す。

A.1 俯瞰セミナーシリーズ^[1]

開催日程：2021年7月29日（木）～9月1日（水）全15回

開催形態：オンライン（Zoomミーティングを使用）

招聘講師：

回	開催日時	講演トピック	講師
1	7月29日（木） 15:00-16:30	ゼロトラストから考えるトラスト アーキテクチャー	松本 泰（セコム株式会社IS研究所 ディビジョンマネージャー）
2	7月30日（金） 10:30-12:00	医療におけるトラスト（1）	尾藤 誠司（東京医療センター 臨床研究センター 臨床疫学研究 室 室長）
3	8月4日（水） 13:00-14:30	人文・社会系のトラスト研究 の系譜	小山 虎（山口大学 時間学研究所 講師）
4	8月4日（水） 15:30-17:00	ソフトウェア品質保証における トラスト	中島 震（国立情報学研究所 名誉教授、放送大学 客員教授）
5	8月5日（木） 9:30-11:00	ヒューマンエージェントインタ ラクションと信頼工学	山田 誠二（国立情報学研究所 教授、総合研究大学院大学 教授、 東京工業大学 特定教授）
6	8月6日（金） 15:00-16:30	AIのトラスト	中川 裕志（理化学研究所AIPセンター チームリーダー）
7	8月10日（火） 15:30-17:00	情報科学におけるトラスト	村山 優子（津田塾大学 数学・計算機科学研究所 研究員）
8	8月11日（水） 10:00-11:30	暗号プロトコルとトラスト	佐古 和恵（早稲田大学基幹理工学部 教授）
9	8月12日（木） 15:00-16:30	社会心理学におけるトラスト	上出 寛子（名古屋大学 未来社会創造機構 特任准教授）
10	8月18日（水） 13:00-14:30	法制度とトラスト	大屋 雄裕（慶應義塾大学法学部 教授）
11	8月19日（木） 9:30-11:00	行動経済学・実験経済学とト ラスト	犬飼 佳吾（明治学院大学経済学部 准教授）
12	8月19日（木） 15:00-16:30	ソーシャルメディアにおけるト ラスト問題	山口 真一（国際大学 GLOCOM 准教授）
13	8月27日（金） 10:00-11:30	公共政策とトラスト	工藤 郁子（大阪大学 社会技術共創研究センター 招へい教員、 世界経済フォーラム 第四次産業革命日本センター プロジェクト 戦略責任者、東京大学 未来ビジョン研究センター 客員研究員）

14	8月27日(金) 13:00~14:30	医療におけるトラスト(2)	山本ベバリーアン(大阪大学大学院人間科学研究科 教授)
15	9月1日(水) 15:00~16:30	科学技術へのトラスト	神里達博(千葉大学大学院国際学術研究院 教授)

各回(90分)のアジェンダ:

- 1) 趣旨説明 [5分] JST CRDS(福島)
- 2) セミナー講演 [40分] 各回の講師
- 3) 質疑・議論 [45分]

聴講登録者(講師・チームメンバー以外):

- インタビューさせていただいた外部有識者 10名
- 文部科学省 10名
- JST関係者 46名
- (計66名:各回30~50名のオンライン聴講があった)

A.2 俯瞰ワークショップ^[1]

開催日程:2021年10月1日(金) 13:00~17:00(4時間)

開催形態:オンライン(Zoomミーティングを使用)

アジェンダ:

- 13:00~13:05 開催挨拶 [JST CRDS 木村康則]
- 13:05~13:35 トラスト研究動向の俯瞰的整理についての発表 [JST CRDS 福島俊一]
- 13:35~14:50 俯瞰的整理についての議論 [コメンテーターから]
 - 【論点1】トラスト研究の俯瞰図の捉え方が妥当か?
 - 【論点2】現在・今後の深刻化するトラスト問題の代表的シーンは何か?
- 14:50~15:00 今後の方向性・重要課題についての論点提示 [JST CRDS 福島]
- 15:00~16:55 総合討議 [コメンテーターから]
 - 【論点3】トラスト研究の目指すべき方向性、重要な研究課題は何か?
 - 【論点4】情報系・人文系・社会系が連携したトラスト研究を推進するための課題・方策は何か?
 - 【論点5】上記以外の問題意識・メッセージ
- 16:55~17:00 まとめ・閉会挨拶 [JST CRDS 木村・福島]

コメンテーター:

(セミナー講師15名のうち都合の合わなかった2名を除く13名、50音順・敬称略)

犬飼 佳吾、大屋 雄裕、神里 達博、工藤 郁子、小山 虎、佐古 和恵、中川 裕志、
中島 震、尾藤 誠司、松本 泰、村山 優子、山口真一、山本 ベバリーアン

参加者(コメンテーター・チームメンバー以外):

- コメンテーター以外の外部有識者 3名
- 文部科学省 4名
- JST関係者 17名(計24名)

A.3 科学技術未来戦略ワークショップ^[19]

開催日程:2022年6月11日(土) 13:00~17:00(4時間)

開催形態:オンライン(Zoomミーティングを使用)

アジェンダ：

- 13:00-13:30 [30分] 開催挨拶・開催趣旨・提言骨子
木村・福島 (JST CRDS)
- 13:30-15:10 [100分] 技術開発・制度設計の状況・課題・方向性 [100分]
手塚 悟：対象真正性の視点
藤代 裕之：内容真実性の視点
杉村 領一：振る舞い予想・対応可能性の視点
稲谷 龍彦：法制度・ガバナンスの視点
- 15:10-15:20 [10分] 休憩
- 15:20-16:00 [40分] 社会・産業および人文・社会科学の視点
浦川 伸一：産業界の視点
川名 晋史：人文・社会科学の視点 (1)
内田 由紀子：人文・社会科学の視点 (2)
- 16:00-16:55 [55分] 総合討議
- 16:55-17:00 [5分] まとめ・閉会挨拶 木村・福島 (JST CRDS)

招聘参加者：

	氏名	所属・役職
発表者	手塚 悟	慶應義塾大学環境情報学部 教授
	藤代 裕之	法政大学社会学部 教授
	杉村 領一	産業技術総合研究所 情報・人間工学領域 上席イノベーションコーディネータ
	稲谷 龍彦	京都大学大学院法学研究科 教授
コメンテーター	浦川 伸一	損害保険ジャパン株式会社 取締役専務執行役員
	川名 晋史	東京工業大学リベラルアーツ研究教育院 准教授
	内田 由紀子	京都大学 人と社会の未来研究院 教授
議論参加者	犬飼 佳吾	明治学院大学経済学部 准教授
	大屋 雄裕	慶應義塾大学法学部 教授
	小山 虎	山口大学 時間学研究所 准教授
	佐古 和恵	早稲田大学基幹理工学部 教授
	中川 裕志	理化学研究所 革新知能統合研究センター 社会におけるAI利活用と法制度チーム チームリーダー
	松本 泰	セコム株式会社 IS研究所 ディビジョンマネージャー
	村山 優子	津田塾大学 数学・計算機科学研究所 特任研究員

聴講登録者（招聘者・チームメンバー以外）：

- 文部科学省 1名
- 経済産業省 1名
- NEDO 1名
- JST関係者 12名（計15名）

A.4 トラスト問題ケース分析

2021年12月には図2-1に示したトラスト問題のケースをいくつか取り上げ、複数の有識者に集まっていた
だき、各ケースの課題や対策について意見交換を行った。ご協力いただいた方々は以下の通り。

犬飼 佳吾	明治学院大学経済学部 准教授
大屋 雄裕	慶應義塾大学法学部 教授
小野 哲雄	北海道大学大学院情報科学研究院 教授
神里 達博	千葉大学大学院国際学術研究院 教授
佐古 和恵	早稲田大学基幹理工学部 教授
中川 裕志	理化学研究所AIPセンター チームリーダー
尾藤 誠司	東京医療センター 臨床研究センター 臨床疫学研究室 室長
松本 泰	セコム株式会社 IS 研究所 ディビジョンマネージャー
村山 優子	津田塾大学 数学・計算機科学研究所 研究員
山口 真一	国際大学 GLOCOM 准教授
山本 ベバリーアン	大阪大学大学院人間科学研究科 教授

A.5 有識者インタビュー（トラスト研究動向）

さまざまな分野におけるトラスト研究開発や応用分野の動向を把握することを目的として、「A.1 俯瞰セミナーシリーズ」の講師以外にも、2021年6月から9月にかけて、以下の有識者にインタビューを行った。

荒井 ひろみ	理化学研究所AIPセンター ユニットリーダー
石原 直子	リクルートワークス研究所 人事研究センター長
伊藤 孝行	京都大学大学院情報学研究科 教授
稲谷 龍彦	京都大学大学院法学研究科 教授
臼田 裕一郎	防災科学技術研究所 総合防災情報センター センター長
江川 尚志	NEC 標準化推進部 シニアエキスパート
江間 有沙	東京大学 未来ビジョン研究センター 准教授
大橋 直之	横浜みなとみらい21 企画調整課 担当課長
川島 典子	福知山公立大学地域経営学部 教授
久木 田水生	名古屋大学大学院情報学研究科 准教授
小林 正啓	花水木法律事務所 弁護士
佐倉 統	東京大学大学院情報学環 教授
積田 有平	シェアリングエコノミー協会 常任理事
福住 伸一	理化学研究所AIPセンター 研究員
森田浩史	電通国際情報サービス オープンイノベーションラボ 所長（同部門から他にも数名）

A.6 有識者インタビュー（戦略提言骨子）

2022年2月から6月にかけて、戦略提言骨子をベースとして、以下の有識者から意見をうかがうインタビューを行った。（※印は本議題の回に参加いただいたJST CRDSの特任フェロー）

相澤 彰子	国立情報学研究所 コンテンツ科学研究系 教授
有村 博紀	北海道大学大学院情報科学研究院 教授
井上 眞梨・小林 茉莉子・栗田 青陽	株式会社メルカリ mercari R4D
大沢 英一	はこだて未来大学システム情報科学部 教授
唐沢 かおり	東京大学大学院人文社会系研究科 教授
川名 晋史	東京工業大学リベラルアーツ研究教育院 准教授

國領 二郎	慶應義塾大学 総合政策学部 教授
小林 傳司	科学技術振興機構社会技術研究開発センター センター長
笹原 和俊	東京工業大学環境・社会理工学院 准教授
島岡 政基	セコム株式会社 IS 研究所 サイバーセキュリティグループ 主任研究員
中江 俊博・桑島 洋	株式会社デンソー ソフト生産革新部先端ソフト開発室
中川 裕志	理化学研究所AIPセンター チームリーダー
松尾 真一郎	ジョージタウン大学 研究教授
村山 優子	津田塾大学 数学・計算機科学研究所 研究員
森下 壮一郎	株式会社サイバーエージェント 技術本部 Data Tech Lab
山口 真一	国際大学 GLOCOM 准教授
※合原 一幸	東京大学特別教授
※浅間 一	東京大学大学院工学系研究科 教授
※喜連 川優	国立情報学研究所 所長
※後藤 厚宏	情報セキュリティ大学院大学 学長・教授
※田中 健一	三菱電機株式会社 開発本部 技術統轄
※徳田 英幸	情報通信研究機構 理事長
※西村 秀和	慶應義塾大学大学院システムデザイン・マネジメント研究科 教授
※森川 博之	東京大学大学院工学系研究科 教授
※山口 高平	慶應義塾大学理工学部 教授

(所属等はインタビューやワークショップの実施時のもの)

付録B 国内外の状況

B.1 トラスト研究の変遷

2.1.2 節で述べたように、トラストに関する研究は幅広い分野でさまざまな切り口から取り組まれてきている。特に人文・社会科学分野では長い研究の歴史がある。付録Aに示したように、2021年に俯瞰セミナーシリーズと俯瞰ワークショップを開催して、幅広い分野での取り組みを俯瞰し、その変遷と現在の状況を報告書^[1]にまとめた。同報告書からその概観図と簡単な説明を以下に再掲する。

まず、人文・社会科学分野におけるトラスト研究の変遷を図5-1にまとめた。古くは17・18世紀頃から哲学・社会学の分野で社会秩序問題という面から捉えた研究の流れがある。20世紀半ば頃には、心理学の立場から「囚人のジレンマ」問題を含む社会集団内の紛争解決要因としてトラストを位置付けた研究が進んだ。

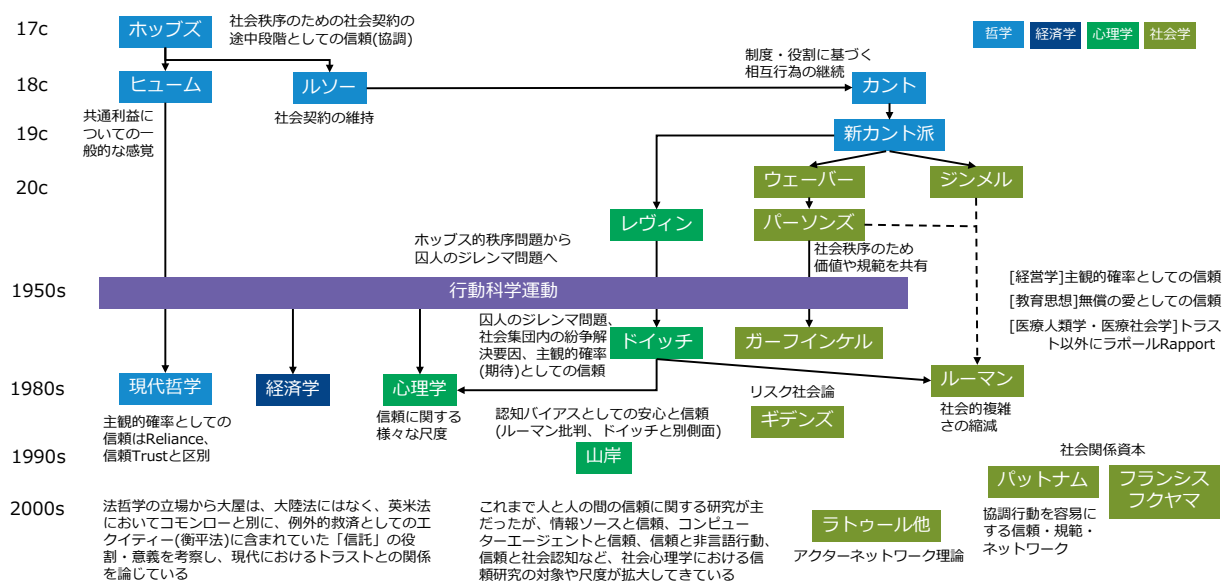


図5-1 トラスト研究の変遷(1) 人文・社会科学分野

その後、1980年頃以降は、トラストがさまざまな役割の中で捉えられるようになってきた。Niklas Luhmann（ニクラス・ルーマン）がトラストの役割を「社会的複雑さの縮減」と位置付けたのは有名である。また、Robert Putnam（ロバート・パットナム）やFrancis Fukuyama（フランシスフクヤマ）はトラストを社会関係資本（Social Capital）として位置付けている。Anthony Giddens（アンソニー・ギデنز）のリスク社会論や、近年のBruno Latour（ブリュノ・ラトゥール）らによるアクターネットワーク理論との関係も深い。国内では、社会心理学の立場から、認知バイアスとしての安心と信頼（トラスト）を対比して論じた山岸俊男の研究が知られている。

人文・社会科学分野では、これまで人と人の間のトラストに関する研究が主だったが、近年は、情報ソースとトラスト、コンピューターエージェントとトラスト、トラストと非言語行動、トラストと社会認知など、社会心理学におけるトラスト研究の対象や尺度が拡大してきている。

次に、情報科学分野におけるトラスト研究の変遷を図5-2にまとめた。情報科学分野の研究として、「トラスト」という言葉を使っでの取り組みが始まったのは1990年代以降である。

情報科学技術分野でのトラストと言ったとき、セキュリティやプライバシーの研究に関わる取り組みが一

つ大きな流れになっている。デジタルトラスト、トラストサービスといった取り組みは、デジタルアイデンティティや認証基盤などの仕組みをベースに個人・組織などが本人・本物であることを保証しようとするものである。また、その下位のコンピューティング層では、CPUなどのデバイスが正しく動作することを保証しようという Device Integrity の取り組みがある。これらは、コンピューターとネットワークを用いたさまざまなデジタルサービスにおけるトラストを支える基盤技術となっている。そこに関わる新しいパラダイム、特に改ざんや盗聴を防止するための新技術として、近年、ブロックチェーン技術や量子暗号（量子鍵配送）技術も注目されている。

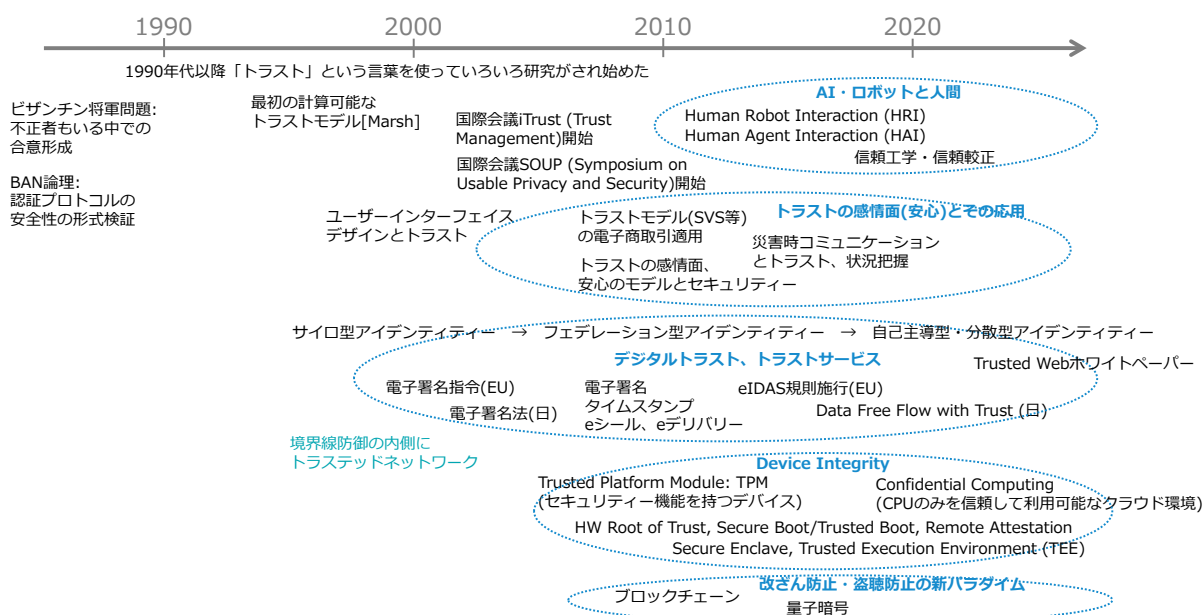


図5-2 トラスト研究の変遷 (2) 情報科学分野

一方、アプリケーション層では、電子商取引、災害時コミュニケーションなどにおいて、トラストの感情面を考慮した取り組みがある。また、AIやロボットと人間との間のトラスト関係が、近年注目されるトピックとなっている。

さらに、トラスト研究の一面として、科学技術のリスク面に目を向けた取り組みがあり、その変遷を図5-3にまとめた。

まず、ITシステムのリスクに関する考え方として、1980～1990年代はコンピューター安全という面からハードウェアや基盤ソフトウェアの信頼性が着目された。2000年代になると、それにアプリケーションまで含めた信頼性・安全性を考えるようになり、ソフトウェアディペンダビリティという見方がされるようになった。さらに近年は、AIやCPS（Cyber-Physical Systems）のTrustworthinessとして、信頼性・安全性に加えて、回復性・プライバシー・セキュリティなども併せて論じられるようになった。

一方、1975年のアシロマ会議に始まり、科学技術のELSI（Ethical, Legal and Social Issues: 倫理的・法的・社会的課題）面の議論が活発に行われるようになり、近年、ITシステム関連では特にAI ELSIが重要課題になっている。「信頼されるAI」「Trustworthy AI」といった表現が用いられ、AI社会原則・AI倫理指針が国・国際レベルで掲げられるようになった。国際標準化活動においてもAIのTrustworthinessが取り上げられている。

また、デマや炎上は古くから存在したが、インターネットやソーシャルメディアの普及・発展に伴い、フェ

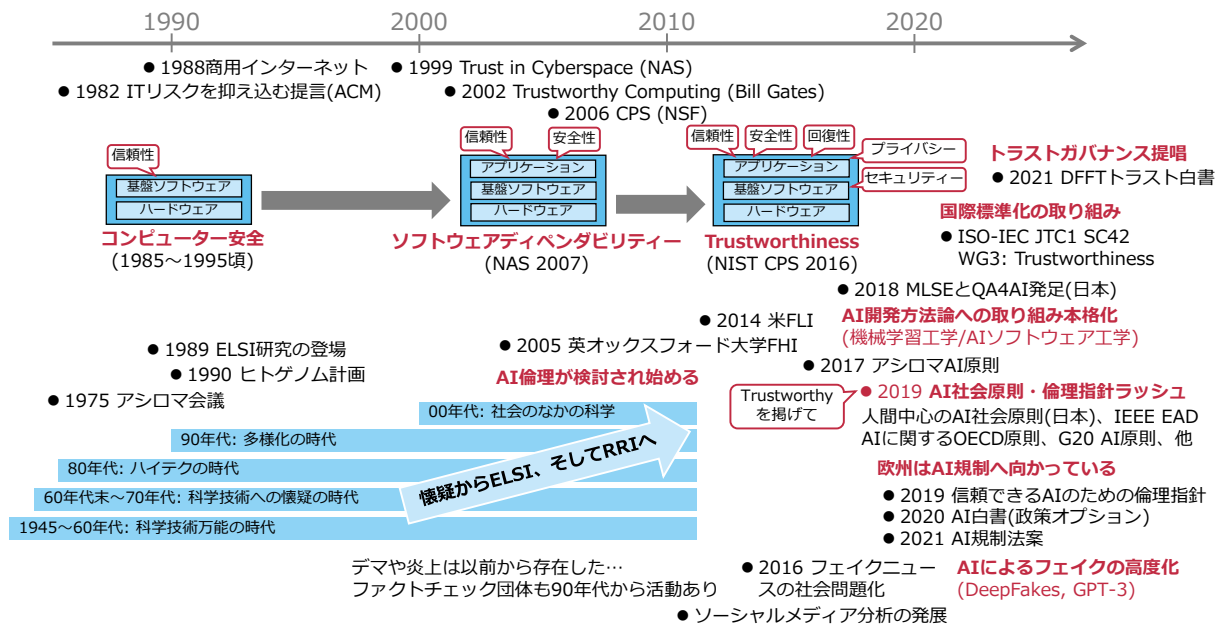


図5-3 トラスト研究の変遷 (2) 情報科学分野

イクニュースが社会問題化し、ディープフェイクで生成されたフェイク動画などのAIによるフェイクの高度化に懸念も高まっている。

B.2 トラスト関連政策提言・プログラム事例

上で述べたような変遷を経て現在に至ったトラスト研究の状況を2.1.2節や2.1.3節で述べた。ここでは、現在の具体的な取り組み事例として、表2-1「注目されるトラスト関連政策提言・プログラム事例」に挙げたものの中からいくつかを取り上げ、補足説明を加える。

「A. デジタルトラスト」関連では、2019年1月のダボス会議（世界経済フォーラム年次総会）において日本政府は「Data Free Flow with Trust」（DFFT：信頼性のある自由なデータ流通）というコンセプトを提唱した。これは「プライバシーやセキュリティ・知的財産権に関する信頼を確保しながら、ビジネスや社会課題の解決に有益なデータが国境を意識することなく自由に行き来する、国際的に自由なデータ流通の促進を目指す」というものである。日本政府における具体的な検討は、デジタルガバメント閣僚会議データ戦略タスクフォースやデジタル庁データ推進戦略ワーキンググループなどで進められている。また、民間企業を中心とした利用者視点での検討母体としてデジタルトラスト協議会も立ち上がっている。DFFTが重視するのは「自由で開かれたデータ流通」と「データの安全・安心」であり、まずは対象真正性に重点を置いたトラストサービスの国際相互連携が欧州・米国・日本の間で推進されている。

「A. デジタルトラスト」関連でもう一つ注目の取り組みは、内閣官房デジタル市場競争本部 Trusted Web 推進協議会である。そのコンセプトや設計方針・ユースケース検討などは「Trusted Web ホワイトペーパー」（2021年3月にVer.1、2022年7月にVer.2案を公開）にまとめられている。Trusted Webでは「やり取りされるデータが信頼できるか、データをやり取りする相手方を信頼できるか、提供したデータの相手方における取り扱いを信頼できるか」といった点をペインポイントと捉えて、インターネットやWebの良さを活かし、その上に重ね合わせるオーバーレイアプローチを採る。特定サービスに過度に依存せず、ユーザー（自然人または法人）自身が自らに関連するデータをコントロールすることを可能とし、データのやり取りにおける合意形成の仕組みを取り入れ、その合意の履行のトレースを可能としつつ、検証できるケース（図2-5の内側の円の範囲）を拡大することによって、トラストの向上を目指している。

「B. フェイク対策」関連の国の政策として、早い段階から推進しているのは米国の国防高等研究計画局（DARPA）である。まず2016年から2020年にかけて「Media Forensics（MediFor）プログラム」が実施され、さらに、その後継として2021年から「Semantic Forensics（SemaFor）プログラム」が実施されている。MediForでは、ピクセル単位での分析や、物理法則・他情報源と組み合わせた分析によって、画像・音声データの改変・不整合を検知するための技術開発が行われた。SemaForでは、それをさらに発展させ、情報源に信頼性があるか、改変の意図に悪意があるかといった意味的な処理まで試みる。なお、フェイクニュースの真偽判定に関わる民間・NPO活動である「ファクトチェック」については、付録Cに記載した。

「C. 信頼されるAI」関連では、統合イノベーション戦略推進会議から2019年6月に公表された「AI戦略2019」（その後のアップデートとして、2021年6月に「AI戦略2021」、2022年4月に「AI戦略2022」が公表されている）において、主要な研究開発課題として「Trusted Quality AI」が掲げられた。これを受けた文部科学省の戦略目標の2020年戦略目標として「信頼されるAI」が設定され、ファンディングプログラムとして、JSTではCREST「信頼されるAIシステムを支える基盤技術」、さがけ「信頼されるAIの基盤技術」が開始された（2020年度～）。NEDO「次世代人工知能・ロボット中核技術開発事業」においても、「人工知能の信頼性に関する技術開発事業」（2019年度実施、2020年度からは「人と共に進化する次世代人工知能に関する技術開発事業」）に発展・移行）が実施された。

「C. 信頼されるAI」関連の海外の政策では、英国研究・イノベーション機構（UKRI）の「Trustworthy Autonomous Systems Programme」（TASプログラム）が注目される。これはトラストできる自律システムに関する学際的研究プログラムで、ELSI/RRIの諸課題を包括している。資金規模は約52億円で、その体制は、Verifiability（検証可能性）、Governance and Regulation（ガバナンスと規制）、Trust（信頼）、Security（安全保障・人権等も含むセキュリティ）、Resilience（レジリエンス）、Functionality（機能性）という6つのノードとハブで構成され、主要大学と多数の産業界からの参加がある。自動運転、医療・介護、防衛・安全保障、AI倫理とガバナンスなどにわたり、12件のプロジェクトが採択されている。

「D. AIガバナンス」関連では、2019年にAI社会原則・AI倫理指針が国・国際レベルで議論・策定された後、「原則から実践へ」とフェーズが移行し、2021年度には、AIガバナンス関連の提言やガイドラインの策定が活発化した。特に国内では、経済産業省から「Governance Innovation Ver.2：アジャイル・ガバナンスのデザインと実装に向けて」（2021年7月）、「AI原則実践のためのガバナンス・ガイドライン」（2021年7月Ver.1、2022年1月Ver.1.1）が公開された。また、世界経済フォーラムから2021年3月に公開された「Rebuilding Trust and Governance: Towards DFFT」白書は、DFFTに向けたトラストの再構築をガバナンスの面から論じている点が特に注目される。

「D. AIガバナンス」関連の海外の政策として、2021年4月欧州委員会から出されたAI規制法案「Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts」は、かなり踏み込んだ規制案である。この中では、AIをリスクの大きさによって、(a) 容認できないリスク、(b) ハイリスク、(c) 限定的なリスク、(d) 最小限のリスク／リスクなし、という4段階に分類し、(a) に該当するAIは使用禁止とされ、(b) は事前に適合性評価、(c) は透明性の確保が必要とされる。

付録C 用語解説

トラストと関連の深い用語

これまでのトラスト研究を俯瞰的にまとめた報告書^[1]に示したように、「トラスト（Trust：信頼）」についてさまざまな定義がなされているが、ある程度共通な理解として、本プロポーザルでは「トラストは相手が期待を裏切らないと思える状態」と定義した（詳細は2.1.4節、図2-4）。ここでは、「トラスト」と類似する用語や関連する用語について、「トラスト」との違いや関係を述べる。ただし、これらの用語自体も分野や研究者によっては異なる捉え方がされているので、あくまで定義や解釈の一例である。

まず、情報システム分野では、システムの「信頼性」について「Reliability」「Dependability」「Trustworthiness」といった用語が用いられる。日本工業規格「JIS Z 8115: 2019」では、「Reliability」は「アイテムが、与えられた条件の下で、与えられた期間、故障せずに、要求通りに遂行できる能力」、「Dependability」は「アイテムが、要求されたときに、その要求通りに遂行するための能力」と定義されている。ここで「アイテム」はシステムやその構成要素などを指す。「Dependability」（総合信頼性）は、「Reliability」よりも上位の概念とされ、「Reliability」（信頼性）だけでなく「Availability」（可用性）、「Maintainability」（保全性）なども含む。また、米国のNIST（国立標準技術研究所）の「NIST Framework for Cyber-Physical Systems」（2017年）では、これらに関わる品質特性を「Trustworthiness」とし、「Safety」（安全性）、「Reliability」（信頼性）、「Privacy」（プライバシー）、「Security」（セキュリティ）、「Resilience」（回復性）などの要素からなるとしている。また、国際標準化機構（ISO）と国際電気標準会議（IEC）の合同による「ISO/IEC TR 24028: 2020」では「Trustworthiness」を「ステークホルダーの期待に検証可能な形で答える能力」と定義し、「Reliability、Availability、Resilience、Security、Privacy、Safety、Accountability、Transparency、Integrity、Authenticity、Quality、Usability」など、さまざまな品質特性を含むとしている。このように情報システム関連の「信頼性」については、故障しないというレベルの「Reliability」から、故障以外の事態も含めて要求通りに動くという「Dependability」に広がり、さらに「Trustworthiness」として、プライバシー、セキュリティ、回復性、透明性など、より広い（社会的な）要請も満たすことが求められるようになってきている。

さらに、「信用」（Credit）は、「デジタル大辞泉」によると「確かなものと信じて受け入れること」「それまでの行為・業績などから、信頼できると判断すること」とされている。「信頼」（Trust）と比べると、過去・実績の評価という面が強い。

また、「信頼」と関係の深い用語として「安心」がある。これらの関係について、山岸俊男^[15]は「信頼は、相手の内面にある人間性や自分に対する感情などの判断に基づいてなされる、相手の意図についての期待」であるのに対して、「安心は、自分を搾取する行動をとる誘因が相手に存在していないと判断することから生まれる」と区別している。すなわち、「信頼」は、相手が「いい人」だと思えるからとか、相手が自分に好意を持っていると思うからといった理由で生まれ、「安心」は、自分を裏切ると相手が損をするからといった理由で生まれるという考えである。一方、村山優子^[20]は、「信頼」の感情的な部分を「安心」と捉えている。本プロポーザルでは、第1章などにおいて、トラストの効果について「リスクがあるとしても、相手をトラストできると、安心して迅速に行動・意思決定ができる」と述べたが、この記述における「安心」は村山らの捉え方に沿っている。

上述の「Trustworthiness」が確保されるための要素として「セキュリティ」が挙げられているが、「セキュリティ」自体は情報システムや情報サービスを支える重要な技術分野として活発に研究開発が進められている。JST CRDSでは「研究開発の俯瞰報告書」^[21]を公表しているが、その中では、近年「セキュリティ」に「トラスト」の観点も合わせて研究開発が広がりつつあると捉えて、「セキュリティ・トラスト」の研究開発動向を俯瞰している。同報告書では、2つの関係を次のように捉えている。すなわち、情報システムや情報

サービスは進歩・発展を続けており、我々の生活に欠かせない存在になっているが、これら情報システムや情報サービスの安全性を確保するのが「セキュリティ」の役目であり、それらを安心して利用できるよう信頼を確保するのが「トラスト」の役目である。

トラストの3側面

本プロポーザルでは「トラストの3側面」(対象真正性、内容真実性、振る舞い予想・対応可能性)という概念を導入し、その「社会的よりどころ」の再構築と多面的・複合的な検証という方向性を示した。この概念に関して、本文中で定義・説明しているが、コラムや脚注などで補足説明の追加も行っており、ここではそれらを簡単に整理する。

「トラストの3側面」は、相手(Trustee)の何をトラストするかに関する分類である。表1-1にまとめた通り、「本人・本物であるか?」という(A)対象真正性、「内容が事実・真実であるか?」という(B)内容真実性、「対象の振る舞いに対して想定・対応できるか?」という(C)振る舞い予想・対応可能性という3側面である。表1-1の脚注1に書いたように、(C)は多くの場合、「対象の振る舞いが想定通り(予想通り・期待通り)か?」と解釈してよい。ただし、想定から外れたとしても、何らかの補償・賠償が約束されているならば大丈夫と思えることがあるので、やや広く「想定・対応できるか?」という表現にした。特にAI応用システムの場合、コラム4「AI応用システムの振る舞い予想・対応可能性」に書いたように、必ずしも想定通りの振る舞いを保証することはできないので、この「対応できるか?」が意味を持ってくる。

2.1.4節で述べたように、「トラストは相手が期待を裏切らないと思える状態」という定義では、(C)の側面(相手の振る舞い)を中心に置いているが、そもそもその相手が本人・本物だという(A)の側面は大前提である。また、相手が人や組織やシステムという場合は、その振る舞いが期待を裏切らないかを考えるが、相手が情報という場合もあり、その場合は、その内容が事実・真実だという(B)の側面を期待する。これは別な言い方をすると、相手が情報の場合、それを使って行動・意思決定をするのは自分であり、それによって自分が招いた結果が期待通りのものかが、内容真実性に左右されるということである。なお、対象が情報の場合、コラム1「情報に関する対象真正性と内容真実性」でそれらの関係についても言及した。

さらに、コラム3「内容真実性：事実や真実は定まるものなのか?」に書いたように、内容真実性は、100%間違いがないと保証できる絶対的な事実・真実が定められるという性質のものではない(ある種の社会的合意と捉えるのが妥当)ということ踏まえておくべきである。

上記2点以外に、いくつかの専門用語についても、簡単な解説を以下に示す。

深層学習

深層学習(Deep Learning)は、多層構造のニューラルネットワークを用いた機械学習技術である。教師あり学習(正解ラベル付きのデータからモデルを学習)において、大量データを用いることで、画像認識などの識別タスク、異常・故障などの検知、時系列変化の予測などで、人の判断を上回る精度が示されるようになり、第3次AIブームを牽引している。さらに、試行錯誤を通してより良い意思決定方策を見つける深層強化学習(Deep Reinforcement Learning)、正解ラベルを与える必要のない自己教師あり学習(Self-supervised Learning)、識別結果だけでなく生成過程までモデル化する深層生成モデル(Deep Generative Model)などの技術発展が続き、応用分野は自然言語処理、機械翻訳、ロボット制御、画像・文章生成などへも広がった。しかし、大量のデータや計算資源が必要なことに加えて、精度が高くとも理由が説明されないこと(ブラックボックス性)、確率的で想定外の振る舞いが起こり得ること、敵対的サンプル(Adversarial Examples)やデータ汚染(Data Poisoning)などの脆弱性が指摘されていることなど、トラストに関わる懸念が生じている。この懸念に対処するため、「信頼されるAI」のための技術開発も活発化した。

総合知

第6期科学技術・イノベーション基本計画（2021年3月26日閣議決定）では、「人文・社会科学の知と自然科学の知の融合による人間や社会の総合的理解と課題解決に貢献する「総合知」と記載されている。「総合知」がなぜ必要かについては、世界の研究や技術開発の目的の軸足が、「持続可能性と強靱性」、「国民の安全と安心の確保」に加えて、「一人ひとりが多様な幸せ（Well-being）を実現できる社会」に移りつつあり、わが国の科学技術やイノベーションが世界と伍していくためには、あらゆる分野の知見を総合的に活用して社会の諸課題への的確な対応を図ることが不可欠だとされている。

さらに、内閣府総合科学技術・イノベーション会議有識者議員懇談会での検討を経て、2022年3月に公表された「「総合知」の基本的考え方及び戦略的に推進する方策 中間とりまとめ」では、「多様な「知」が集い、新たな価値を創出する「知の活力」を生むこと」が「総合知」であり、「総合知」を推し進めることが、科学技術・イノベーションの力を高めるとされている。

パーソナルAIエージェント

パーソナルAIエージェント（Personal AI Agent: PAIA）は、そのユーザーについて、個人データの収集・管理・保護をおこなう代理ソフトウェアである。ユーザーに関するさまざまなデータから学習したり、ユーザーから明示的な指示を受けたりすることで、データの管理・保護などにおけるユーザーの方針や嗜好を獲得する。その方針・嗜好に基づいて、さまざまなアプリケーションに対して、ユーザーの個人データをどこまで開示するか代理判断まで行えることが目指されている。

ヒューマンエージェントインタラクション

ヒューマンエージェントインタラクション（Human-Agent Interaction: HAI）は、人とさまざまなエージェントとの間のインタラクション（人とAI・ロボットとのインタラクションや、デジタル技術を介した人と人の間のインタラクション）を対象とし、その設計方法論の確立と工学的応用を目的としている。エージェントの外見・動きなどの非言語情報や人間の心理・認知なども含めた総合的なインタラクションを扱う学際的研究分野である。

ファクトチェック

ファクトチェック（Fact Checking）とは、フェイクニュースに対して、発信された情報が客観的事実に基づくものなのかを調査し、その情報の正確さを評価・公表する取り組みを意味する。ファクトチェックを行う団体として比較的早期に立ち上がった米国のSnopes（1994年～）やPolitiFact（2007年～）がよく知られている。この動きは世界的に広がっており、日本では2017年にFactCheck Initiative Japan（FIJ）が発足した。2015年には、こういった活動をつなぐInternational Fact-Checking Network（IFCN）も組織されている。

付録D 参考文献

- [1] 科学技術振興機構研究開発戦略センター,「俯瞰セミナー&ワークショップ報告書:トラスト研究の潮流 ~人文・社会科学から人工知能、医療まで~, CRDS-FY2021-WR-05 (2022年2月).
<https://www.jst.go.jp/crds/report/CRDS-FY2021-WR-05.html>
- [2] 世界経済フォーラム,「Rebuilding Trust and Governance: Towards Data Free Flow with Trust (DFFT)」白書 (2021年3月).
<https://jp.weforum.org/whitepapers/rebuilding-trust-and-governance-towards-data-free-flow-with-trust-dfft>
- [3] Trusted Web 推進協議会,「Trusted Web ホワイトペーパー Ver.1」(2021年3月).
chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.kantei.go.jp/jp/singi/digitalmarket/trusted_web/pdf/documents_210331-2.pdf
- [4] ニクラス・ルーマン (著), 大庭健・正村俊之 (訳),『信頼: 社会的な複雑性の縮減メカニズム』(勁草書房, 1990年).
- [5] 工藤郁子,「人々の「眠り」と「目覚め」、社会の信頼 再構築を」, 朝日新聞デジタル「にじいろの議」(2020年8月12日). <https://www.asahi.com/articles/DA3S14584996.html>
- [6] レイチェル・ボツツマン (著), 関美和 (訳),『TRUST: 世界最先端の企業はいかに〈信頼〉を攻略したか』(日経BP, 2018年).
- [7] 日本ネットワークセキュリティ協会 PKI 相互運用技術WG・電子署名WG,「デジタル社会におけるトラスト」セミナー講演集, PKI & TRUST Days online 2021 (2021年4月15日・16日).
<https://www.jnsa.org/seminar/pki-day/2021/index.html>
- [8] 科学技術振興機構研究開発戦略センター,「公開ワークショップ報告書: 意思決定のための情報科学 ~情報氾濫・フェイク・分断に立ち向かうことは可能か~, CRDS-FY2019-WR-02 (2020年2月).
<https://www.jst.go.jp/crds/report/CRDS-FY2019-WR-02.html>
- [9] 山口真一,『ソーシャルメディア解体全書: フェイクニュース・ネット炎上・情報の偏り』(勁草書房, 2022年).
- [10] 科学技術振興機構研究開発戦略センター,「戦略プロポーザル: AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」, CRDS-FY2018-SP-03 (2018年12月).
<https://www.jst.go.jp/crds/report/CRDS-FY2018-SP-03.html>
- [11] 石川冬樹・丸山宏・柿沼太一・竹内広宜・土橋昌・中川裕志・原聡・堀内新吾・鷺崎弘宜,『機械学習工学』(講談社, 2022年).
- [12] 経済産業省 AI 原則の実践の在り方に関する検討会「AI 原則実践のためのガバナンス・ガイドライン Ver. 1.1」(2022年1月).
<https://www.meti.go.jp/press/2021/01/20220125001/20220124003.html>
- [13] 小山虎 (編著),『信頼を考える: リヴァイアサンから人工知能まで』(勁草書房, 2018年).
- [14] Piotr Pietrzak and Josu Takala, “Digital trust – a systematic literature review”, *Forum Scientiae Oeconomia*, Vol.9, No.3 (2021). DOI: 10.23762/FSO_VOL9_NO3_4
- [15] 山岸俊男,『信頼の構造: こころと社会の進化ゲーム』(東京大学出版会, 1998年).
- [16] Roger C. Mayer, James H. Davis and F. David Schoorman, “An Integrative Model of Organizational Trust”, *Academy of Management Review*, Vol. 20, No. 3, pp. 709-734 (1995). DOI: 10.2307/258792
- [17] 科学技術振興機構研究開発戦略センター,「戦略プロポーザル: 複雑社会における意思決定・合意形

- 成を支える情報科学技術」, CRDS-FY2017-SP-03 (2018年3月).
<https://www.jst.go.jp/crds/report/CRDS-FY2017-SP-03.html>
- [18] AIプロダクト品質保証 (QA4AI) コンソーシアム, 「AIプロダクト品質保証ガイドライン」(2022年7月31日時点の最新版は2022.07版). <https://www.qa4ai.jp/download/>
- [19] 科学技術振興機構研究開発戦略センター, 「科学技術未来戦略ワークショップ報告書:トラスト研究戦略 ~デジタル化における新たなトラスト形成~」, CRDS-FY2022-WR-05 (2022年8月).
<https://www.jst.go.jp/crds/report/CRDS-FY2022-WR-05.html> (予定)
- [20] 村山優子・藤原康宏, 「トラストの感情としての安心およびその要因について」, 日本信頼性学会誌, Vol. 31, No. 1, pp. 41-46 (2009年). DOI: 10.11348/reaishinrai.31.1_41
- [21] 科学技術振興機構研究開発戦略センター, 「研究開発の俯瞰報告書:システム・情報科学技術分野 (2021年)」, CRDS-FY2020-FR-02 (2021年3月).
<https://www.jst.go.jp/crds/report/CRDS-FY2020-FR-02.html>

なお、URLは2022年7月31日時点のものである。

作成メンバー

総括責任者	木村 康則	上席フェロー	CRDS システム・情報科学技術ユニット
リーダー	福島 俊一	フェロー	CRDS システム・情報科学技術ユニット
メンバー	井上 真梨	フェロー	CRDS システム・情報科学技術ユニット (2021年9月まで)
	加納 寛之	フェロー	CRDS 科学技術イノベーション政策ユニット
	上村 健	フェロー	社会技術研究開発センター 企画運営室
	住田 朋久	フェロー	CRDS 企画運営室
	高島 洋典	フェロー	CRDS システム・情報科学技術ユニット
	戸田 智美	フェロー	CRDS ライフサイエンス・臨床医学ユニット
	花田 文子	フェロー	CRDS 企画運営室 (2021年10月から)
	福井 章人	フェロー	CRDS システム・情報科学技術ユニット (2021年10月から)
	的場 正憲	フェロー	CRDS システム・情報科学技術ユニット
	宮藺 侑也	フェロー	CRDS ライフサイエンス・臨床医学ユニット (2022年3月まで)
	茂木 強	フェロー	CRDS システム・情報科学技術ユニット
	山本 里枝子	フェロー	CRDS 企画運営室 (2021年10月から)
	若山 正人	上席フェロー	CRDS システム・情報科学技術ユニット

戦略プロポーザル

CRDS-FY2022-SP-03

デジタル社会における新たなトラスト形成

STRATEGIC PROPOSAL

New Trust Formation in the Digital Society

令和 4 年 9 月 September 2022

ISBN 978-4-88890-814-6

国立研究開発法人科学技術振興機構 研究開発戦略センター

Center for Research and Development Strategy, Japan Science and Technology Agency

〒102-0076 東京都千代田区五番町7 K's 五番町

電話 03-5214-7481

E-mail crds@jst.go.jp

<https://www.jst.go.jp/crds/>

本書は著作権法等によって著作権が保護された著作物です。

著作権法で認められた場合を除き、本書の全部又は一部を許可無く複写・複製することを禁じます。

引用を行う際は、必ず出典を記述願います。

This publication is protected by copyright law and international treaties.

No part of this publication may be copied or reproduced in any form or by any means without permission of JST, except to the extent permitted by applicable law.

Any quotations must be appropriately acknowledged.

If you wish to copy, reproduce, display or otherwise use this publication, please contact crds@jst.go.jp.

FOR THE FUTURE OF
SCIENCE AND
SOCIETY



CRDS

<https://www.jst.go.jp/crds/>