ワークショップ報告書
CRDS-FY2021-WR-05

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～

2021 年 7 月～ 10 月開催

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

# エグゼクティブサマリー

本報告書は、「トラスト（Trust、信頼）」に関わる、さまざまな分野における研究動向を俯瞰するために開催したセミナーシリーズ（全15回）とワークショップの内容をまとめたものである。

人工知能（AI）技術やデジタル技術は、社会にさまざまな価値を提供し、人々の活躍の可能性を広げ、生活を豊かにしてくれる。その一方で、そのような高度で複雑な技術は、人間にとっていわばブラックボックスのようなものとなり、それを組み込んだシステムは、人々の予測や期待から外れた振る舞いをしてしまうリスクを有する。このようなリスクは、自動運転・医療のような人命にも関わる応用分野にも及ぶ。

また、デジタル社会では、他者へのなりすましやアカウントの乗っ取りといった手口による犯罪が起きている。AI技術の発展により、本物か否かを見分けるのが非常に難しいフェイク画像・音声・動画・会話文などが簡単に作れるようになってしまい、政治操作やフェイクポルノ動画などに悪用される問題も起きている。

そこで、AI技術・デジタル技術の社会受容や、それら先端技術を用いた情報システム・情報サービスの安心できる利用に関して、「トラスト」が論じられることが多くなってきた。上で述べたようなリスクを抑えるために、説明可能AI、信頼されるAI、機械学習工学など、AI応用システムの安全性・信頼性を確保するための技術開発や、デジタル署名、コンフィデンシャルコンピューティング、ブロックチェーン、フェイク検出など、真正性確保や捏造・改ざん対策のためのセキュリティー技術開発が進められている。しかし、上で述べたようなリスクを技術開発だけで完全に抑え込むことは難しい。「トラスト」は、相手・対象が期待を裏切らないと思える状態とされるが、これは技術開発だけで得られるものではなく、多様な価値観やリテラシーを持つ各個人や社会全体がどのように受け止めるかといった心理・感情面や、法律・保険などを含む制度設計・整備の状況からも大きく影響を受ける。

一方、「トラスト」に関わる研究は、人文・社会科学の分野で古くから取り組まれてきた。人文・社会科学の中でも、哲学・社会学・心理学・経済学などのさまざまな学問分野で「トラスト」に関してさまざまな捉え方がされてきた。人文・社会科学の分野では、主に人間関係における「トラスト」が論じられてきたように思えるのに対して、情報科学分野では、「トラスト」の対象として、機械・コンピューターシステムやそれらを介した先にいる相手が論じられることが多い。例えば、デジタル化されたサービスにおいて、改ざんやなりすましがなく本人・本物であることをどのようにして確認するか、コンピューターシステムやアプリケーションがユーザーの期待通りに動作するか、といった問題が「トラスト」の一面として扱われる。デジタル化の進展によって、「トラスト」に関わる要因や「トラスト」の役割も変化してきている。

そこで、さまざまな分野で取り組まれている「トラスト」研究や関連動向を把握するため、科学技術振興機構（JST）研究開発戦略センター（CRDS）では、2021年7月29日から9月1日にかけて全15回、計15名の講師にお願いして「トラスト研究俯瞰セミナーシリーズ」を開催した。さらに、その内容を総括し、トラスト研究動向の俯瞰的な整理を試みるとともに、「トラスト」に関わる分野横断的な議論を行う場として、10月1日に「トラスト研究俯瞰ワークショップ」を開催した。

本報告書では、第1章でデジタル社会におけるトラストに関わる問題意識について述べた上で、第2章でセミナーシリーズの各回の内容、第3章でワークショップの内容について、それぞれ報告する。これらセミナーシリーズとワークショップは、さまざまな分野・観点で取り組まれているトラスト研究に関して、異なる分野の研究者間で知見を共有し、分野横断の議論が行える場としても、非常に有意義なものになった。

なお、JST CRDSは、科学技術に求められる社会的・経済的なニーズを踏まえて、国として重点的に推進すべき研究領域や課題、その推進方策に関する提言を行っており、今回の俯瞰的な整理に基づき、今後、「デジタル社会における新たなトラスト形成」に関わる戦略提言をまとめていく計画である。

# 目次

**コラム一覧**

# 1 ｜ 問題意識および俯瞰の進め方



科学技術振興機構（JST）研究開発戦略センター（CRDS）は、科学技術に求められる社会的・経済的なニーズを踏まえて、国として重点的に推進すべき研究領域や課題、その推進方策に関する提言を行っている。この一環として、2021年度は「デジタル社会における新たなトラスト形成」をテーマ（戦略スコープ）とした調査・検討を進めている。

我々はこの調査・検討を、フェーズ1・フェーズ2という2ステップで進めることにした。本戦略スコープに関わる分野は非常に幅広いと考えられるため、フェーズ1では、それら幅広い分野における取り組みの俯瞰を試みる。その結果を踏まえて、フェーズ2では、戦略提言の意義・必要性を吟味した上で、研究の方向性や推進方策を中心に具体的な提言内容を検討する。

本報告書は、フェーズ1の活動の総括として、「トラスト（Trust、信頼）」に関わる研究動向の俯瞰を試みたものである。以下、本戦略スコープに着目した理由（問題意識）と、俯瞰の進め方について述べる。

**問題意識**

人工知能（AI）技術やデジタル技術は、社会にさまざまな価値を提供し、人々の活躍の可能性を広げ、生活を豊かにする。その一方で、そのような高度で複雑な技術は、人間にとっていわばブラックボックスのようなものとなり、それを組み込んだシステムは、人々の予測や期待から外れた振る舞いをしてしまうリスクを有する[1]。AI技術・デジタル技術の応用分野は拡大しており、例えば自動運転・医療のような応用において、このリスクは人命にも関わりかねないのではないかという懸念も生じている（図1-1）。

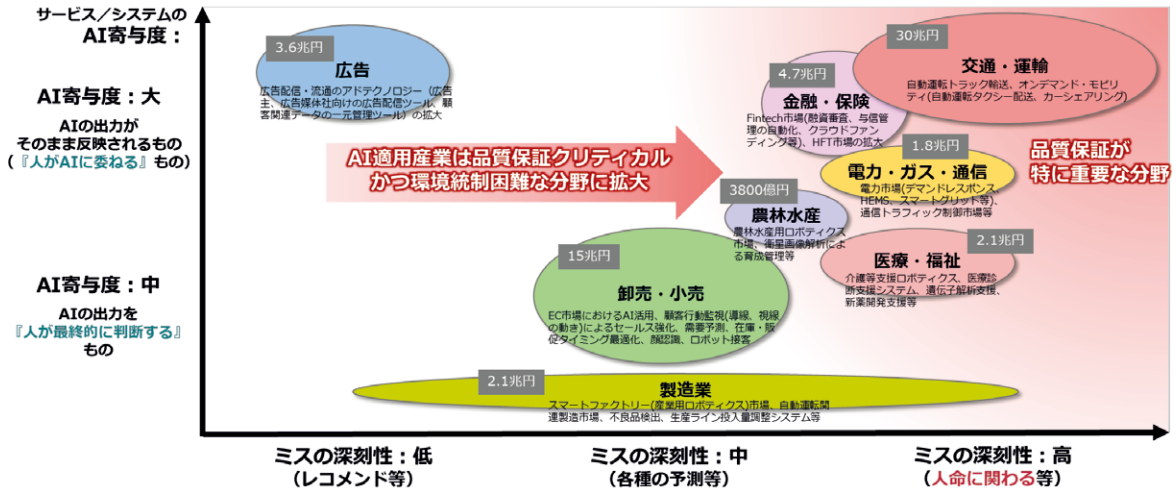


**図1-1　AI応用分野の広がり[1]**

1　図1-1は既発行報告書[1]から再掲。図中の金額は2030年のAI適用産業の予想市場規模であり、EY総合研究所のレポート[2]をもとにした。横軸の「ミスの深刻性」は、AI・機械学習が誤った判定結果を出したときに生じる問題がどれくらい深刻であるかを意味する。人命に関わるような場合は深刻性が高い。縦軸の「AI寄与度」は、問題解決のために実行されるアクションの決定にAI・機械学習がどれくらい大きく寄与するかを意味する。AI・機械学習の出力（判定結果）がそのまま反映される場合は寄与度が高く、AI・機械学習の出力（判定結果）を参考にして人間が最終的に判断する場合は寄与度が低い。また、「環境統制困難性」は、AI・機械学習を実行する際の環境条件をコントロールすることの難しさを意味する。環境条件を列挙することが難しく想定外のことがいろいろ起こり得る場合は困難性が高く、環境条件を統制することが容易であれば困難性が低い。

また、デジタル社会では、他者へのなりすましやアカウントの乗っ取りといった手口による犯罪が起きている。AI技術の発展により、本物か否かを見分けるのが非常に難しいフェイク画像・音声・動画・会話文などが簡単に作れるようになってしまい、政治操作やフェイクポルノ動画などに悪用される問題も起きている[3]。つまり、本人かどうか、本物かどうか、という疑念・不安をぬぐい切れない事態や、だまされてしまうリスクが高まっている。

このような情報の真偽見極めの困難さや先端技術の理解困難さ（ブラックボックス性）から生まれるデジタル社会のリスクが顕在化・深刻化すると、フェイクによる世論誘導や犯罪、民主主義や法制度の危機、AI応用に対する不安、科学技術に対する不安など、社会的に大きな問題を引き起こしかねない。

そこで、AI技術・デジタル技術の社会受容や、それら先端技術を用いた情報システム・情報サービスの安心できる利用に関して、「トラスト」が論じられることが多くなってきた。上で述べたようなリスクを抑えるために、説明可能AI、信頼されるAI、機械学習工学など、AI応用システムの安全性・信頼性を確保するための技術開発や、デジタル署名、コンフィデンシャルコンピューティング、ブロックチェーン、フェイク検出など、真正性確保や捏造・改ざん対策のためのセキュリティー技術開発が進められている。しかし、上で述べたようなリスクを技術開発だけで完全に抑え込むことは難しい。「トラスト」は、相手・対象が期待を裏切らないと思える状態とされるが、これは技術開発だけで得られるものではなく、多様な価値観・リテラシーを持つ各個人や社会全体がどのように受け止めるかといった心理・感情面や、法律・保険などを含む制度設計・整備の状況からも大きく影響を受ける。


**図1-2　「トラスト」に関わる環境変化**

また、「トラスト」に関わる環境変化に目を向けると（図1-2）、デジタル社会以前は、家族・隣組・職場内などの日々直接顔を合わせる身近な人たち同士の関係が「トラスト」のベースとなっていた。しかし、デジタル化の進展によって、バーチャルな人間関係が生まれて広がり、複雑な技術を用いたシステムへの依存も進み、極めて高度な「だます技術」も現れたことで、人々の間の関係や人々と社会の関係が多様化・複雑化した。その結果、旧来の方法による「トラスト」形成はもはや成り立たないような状況が広がっている。

デマやフェイク、改ざん、ブラックボックス技術といった問題は昔から起きていて、これまでは人間による判断や経験・実績の積み重ねによって、リスクがそれなりに抑え込めていたのかもしれない。しかし、デジタル技術の発展によって量や複雑さが増大するペースと、人間主体の従来型対策を強化できるペースを比較すると、前者が指数関数的に増大するのに対して後者は線形に近い（図1-3）。このような傾向も、増大するリスクへの対策や「トラスト」形成のために、新しいアプローチが必要になってきた原因と考えられる。

また、第2章は、講演の書き起こしスタイルを採っているが、文中に現れる人名について敬称を省かせていただいた点もご容赦いただきたい。文責は奥付に掲載した報告書編纂メンバーにある。

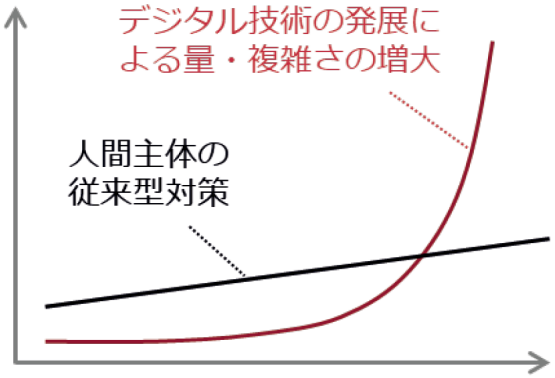


**図1-3　　従来型対策のペースを上回るリスク増大傾向**

そこで、JST CRDSでは「デジタル社会における新たなトラスト形成」を戦略スコープに設定し、技術開発だけでなく制度設計や人間の受け止め方といった面にも目を向けた調査・検討を進めることにした。

**俯瞰の進め方**

以上の問題意識に基づき、冒頭で述べたように、まず「トラスト（Trust、信頼）」に関わる研究動向の俯瞰を試みることにした。具体的には、以下で述べるような3つの活動（俯瞰セミナーシリーズ、有識者インタビュー、俯瞰ワークショップ）を実施した。いずれもクローズドな会として実施したものだが、その内容を一般公開向けに抜粋し、本報告書としてまとめた。

**❶ 俯瞰セミナーシリーズ**

「トラスト」に関わる研究は、人文・社会科学の分野で古くから取り組まれてきた。人文・社会科学の中でも、哲学・社会学・心理学・経済学などのさまざまな学問分野で「トラスト」に関してさまざまな捉え方がされてきた。人文・社会科学の分野では、主に人間関係における「トラスト」が論じられてきたように思えるのに対して、情報科学分野では、「トラスト」の対象として、機械・コンピューターシステムやそれらを介した先にいる相手が論じられることが多い。例えば、デジタル化されたサービスにおいて、改ざんやなりすましがなく本人・本物であることをどのようにして確認するか、コンピューターシステムやアプリケーションがユーザーの期待通りに動作するか、といった問題が「トラスト」の一面として扱われる。

このような「トラスト」に関する幅広い研究動向を把握するために、2021年7月29日から9月1日にかけて全15回の「トラスト研究俯瞰セミナーシリーズ」を開催した（開催概要は付録1参照）。このセミナーシリーズでは、人文・社会科学分野から情報科学分野まで幅広く、各分野での「トラスト」に関わる研究動向について俯瞰的に話していただける方や、医療・公共政策・ソーシャルメディアなどの具体的な場面で「トラスト」がどのように作用しているかについて話していただける方に講師をお願いした（表1-1）。なお、さまざまな分野における「信頼研究」を俯瞰した先行的な取り組みが書籍『信頼を考える』[4]にまとめられていることから、その編著者にもセミナーシリーズの講師をお願いした。

このセミナーシリーズ全15回の講演・質疑の概要は、本報告書の第2章（2.1～2.15）にまとめた。

**表1-1　　トラスト研究俯瞰セミナーシリーズで取り上げたトピック**

| 講演トピック | 講師 | 掲載節 |
|---|---|---|
| 人文・社会系のトラスト研究の系譜 | 小山 虎 | 2.1 |
| 社会心理学におけるトラスト | 上出 寛子 | 2.2 |
| 行動経済学・実験経済学とトラスト | 犬飼 佳吾 | 2.3 |
| 法制度とトラスト | 大屋 雄裕 | 2.4 |
| 科学技術へのトラスト | 神里 達博 | 2.5 |
| 情報科学におけるトラスト | 村山 優子 | 2.6 |
| ソフトウェア品質保証におけるトラスト | 中島 震 | 2.7 |
| 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 | 2.8 |
| 暗号プロトコルとトラスト | 佐古 和恵 | 2.9 |
| 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 | 2.10 |
| AIのトラスト | 中川 裕志 | 2.11 |
| 公共政策とトラスト | 工藤 郁子 | 2.12 |
| ソーシャルメディアにおけるトラスト問題 | 山口 真一 | 2.13 |
| 医療におけるトラスト（1） | 尾藤 誠司 | 2.14 |
| 医療におけるトラスト（2） | 山本 ベバリーアン | 2.15 |

**❷ 有識者インタビュー**

講演型のセミナーシリーズだけでなく、「トラスト」に関わる研究や事例について、個別に話をうかがう有識者インタビューも実施した（付録3参照）。2021年6月から8月にかけて、計25回実施した（計30名、うち9名はセミナーシリーズでも話していただいた）。

本報告書には、インタビュー内容をそのまま掲載することはしていないが、セミナーシリーズを補完する情報・知見が得られた。その一部は本報告書中のコラムで取り上げる。

**❸ 俯瞰ワークショップ**

JST CRDSでは、以上のようなセミナーシリーズやインタビューを通して「トラスト」に関わる研究動向を調査し、その俯瞰的整理を試みた。その妥当性や今後の方向性を議論するため、2021年10月1日に俯瞰ワークショップを開催した（開催概要は付録2参照）。このワークショップでは、JST CRDSから俯瞰的整理の内容を発表し、セミナーシリーズの講師の方々にコメンテーターとして参加していただき、内容の妥当性や今後の方向性を中心に意見をいただき議論した。その結果の概要を本報告書の第3章にまとめた。

なお、本報告書は「トラスト」に関わる研究動向の俯瞰的整理を試みたものであるが、既に述べたように非常に幅広い分野でさまざまな切り口で取り組まれているため、その全貌を把握することは必ずしも容易ではない。その意味で、本報告書は、上記❶❷❸のような活動から得られた情報・知見の範囲での俯瞰的整理に留まる点はご容赦いただきたい。

また、第2章は、講演の書き起こしスタイルを採っているが、文中に現れる人名について敬称を省かせていただいた点もご容赦いただきたい。文責は奥付に掲載した報告書編纂メンバーにある。

# 2 ｜ 俯瞰セミナーシリーズ

## 2.1 小山 虎[1]「人文・社会系のトラスト研究の系譜」

### トラスト研究の多様性

人文・社会系のトラスト研究の系譜について紹介する。2018 年に出版した『信頼を考える－リヴァイアサンから人工知能まで』[1] では、哲学の社会思想に始まり、行動科学、政治学、社会心理学、ビジネス、教育、医療、ロボット、障害者福祉、ヘイトスピーチ、ヒューマンエージェントインタラクション（HAI）における




**図 2-1-1　　『信頼を考える』[2]**

1　山口大学 時間学研究所 講師
http://www.rits.yamaguchi-u.ac.jp/?page_id=1996
大阪大学 基礎工学研究科 招へい准教授、京都大学 文学研究科 応用哲学倫理学教育研究センター（CAPE）センター員

2　『信頼を考える－リヴァイアサンから人工知能まで』[1] から表紙・目次を引用

信頼を取り上げ、信頼研究の全体像を大まかに示すことを試みた。

このようにさまざまな研究分野で信頼が扱われている。共通点もあれば、違う点もある。人文・社会系に限っても、用いられている信頼の定義には幅がある。これは信頼という現象そのものの問題なのか、それとも、我々が使う言葉の問題なのか、いろいろと議論の切り口がある。

| 分野 | 定義 |
|---|---|
| 経済学 | 「［信頼は］、1人ないし複数の行為者が**特定の行為を遂行する**という一定のレベルの**主観的確率**であり、彼らの行為をチェックすることができる以前に（あるいはチェックすることができる能力とは独立に）、その行為が自分自身の行為に影響を与える状況で形成される」(Gambetta, D (1988). Can We trust trust? In Trust: Making and Breaking Cooperative Relations: 217) |
| 経営学 | 「［信頼は］、他者の**意図や行動についてのポジティブな期待に基づきリスクを受け入れる**意図を含む心理状態である」(Rousseau, D., Sitkin, S., Burt, R., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. The Academy of Management Review, 23(3): p. 395) |
| 社会学 | 「［信頼は］欠けている情報を内的に保証された安全性に置き換えるのであり、［自分に］**利用可能な情報を超えて、行動の期待を一般化**することにより、**社会の複雑さを減少させる**」(Luhmann, N. (1979) Trust and Power: p. 93) |
| | 「［信頼は］、**他者の誠実さや愛**あるいは抽象的な原理への信念を表すような、人やシステムが一群の結果や出来事を実際にもたらすという**確信**」(Giddens, A. (1990) The Consequences of Modernity: 34) |
| 動物行動学 | 「信頼は、他者の**誠実さ**または協力への依存、あるいは少なくとも他者があなたを欺かないという**期待**、と定義される」(de Waal, F. (2009). The Age of Empathy: Nature's Lessons for a kinder Society: 167) |
| 哲学 | AがBはCすると信頼するのは、(1) AはBがCすると**期待**し、(2) このAの期待(1)が、Bが**自分の関心を叶えようという動機に基づいている**というAの信念か知識に基づいているとき (Hardin, R. (1991) Trust and Trustworthiness, Russell Sage Foundation) |
| | AがBはCすると信頼するのは、(1) AはBに重要事Cを任せ、(2) AはどのようにCを扱うのかのコントロールをある程度Bに許し、(3) AはBがCを**扱うことができると確信**しており、(4) Aは自分に対するBの**善意**に確信を持っているか、少なくともBの悪意や無関心を予期しないとき (Baier, A. (1986) Trust and antitrust, Ethics 96) |
| | AがBはCすると信頼するのは、(1) AがBの**善意に対する楽観的態度**を持ち、(2) AはBの予期される行動Cに対するBの**能力に対する楽観的態度**を持ち、(3) Aは自分が頼りにすることを認識することによって直接BがCするように**動機付けられる**と信じているとき (Jones, K. (1996) Trust as an Affective Attitude, Ethics 107) |
| | AがBはCすると信頼するのは、(1) AはCの配慮にあたってBがある**社会的規範に内的にコミットしていると期待**し、(2) Aは「BがCの配慮にあたってAによって想定されている**社会的規範を認識**し、また、その規範が何を要求しているかを理解することができる」と確信しており、(3) AはBが自分に課せられた**規範にしたがって行為することができると信じている**とき(Mullin, A. (2005) Trust, Social Norms, and Motherhood, Journal of Social Philosophy 36) |

**図2-1-2　　さまざまな分野の信頼の定義**

### 囚人のジレンマと信頼

いくつかの切り口はあり得るが、情報量の観点から、囚人のジレンマに着目して信頼研究の系譜を検討する。囚人のジレンマのような実験を用いて信頼を分析するのは、一部の分野では標準的な方法になっている。個人が合理的であるだけでは必ずしも協調行動を取るメリットがあるがどうか分からないが、相手の取り得る行動を踏まえた上で合理的に考えれば協調行動を取る方が有利な状況がある。信頼があれば、協調行動を取ることが合理的な選択になる。おそらくはそれをもって社会が形成されるだろう。反対に、信頼がなかったら人々は協調行動を取らない。それぞれが自己利益のために行動し、社会は生まれないだろう。

### 信頼研究の始まり：社会契約論

囚人のジレンマが関わる信頼研究の起源は、少なくとも、17世紀の社会契約論にまで遡ることができる。Thomas Hobbes（トマス・ホッブズ）によれば、人間というのは自身の生存を目的としている。そのため、自然状態では自己保存のために争いが起こる。これが万人の万人に対する闘争といわれるものである。

社会契約論は、簡単に言うと、社会秩序はいかにして可能なのかという問題を扱う分野を指す。これはホッブズ的秩序問題と後に呼ばれるようになる。ホッブズによると、社会契約を結べば社会は作れるわけだが、自然状態で闘争している段階から、信頼ないし協調行動を取るという途中段階がある。

注意しなければならないのは、ホッブズが現在あるような信頼研究や囚人のジレンマに関係するような信頼を中心的に研究し始めたわけではないという点である。特に、この時代までは、神や宗教の影響も強く、信頼を信仰や信用とどのように区別するかも難しい。ホッブズは信頼（Trust）と信用（Believe）や信仰（Faith）を区別していない。

18世紀には、David Hume（デイヴィッド・ヒューム）、Jean–Jacques Rousseau（ジャン＝ジャック・ルソー）、Immanuel Kant（イマヌエル・カント）といった哲学者たちが、独自の見解を出している。ヒュームは共通利益についての一般的な感覚（コンヴェンション）が、社会秩序の形成の鍵だと考える。ルソーは、ホッブズが社会秩序はどうやって生まれるのかを考えたのに対し、社会秩序はいかに維持されるのかを考えた。ホッブズが信頼に関心を持っていたのに対し、ルソーは不信を考えたと言ってよいかもしれない。カントは制度や役割に基づく相互行為の継続によって秩序が保たれると考えた。

**哲学から社会学・心理学へ**

19世紀になると社会学が生まれる。1838年、フランスの社会学者（実証主義者）Auguste Comte（オーギュスト・コント）が「社会学」という用語を導入した。19世紀後半から20世紀にかけて、社会学者と呼ばれる人たちが登場してくる。その代表として、フランスのEmile Durkheim（エミール・デュルケーム）や、ドイツのMax Weber（マックス・ウェーバー）、Georg Simmel（ゲオルグ・ジンメル）などが挙げられる。

20世紀初頭のドイツ社会学はドイツ哲学（特に新カント派）からの影響が強かった。19世紀終わり頃に、新カント派の復興運動がドイツで起こった。その結果、19世紀の終わりから20世紀の頭のドイツの社会学というのは、カントを背景にして社会学を考える、あるいはカントの用語やカントの位置付けみたいなものが中で意識されることとなった。

信頼に関していうと、先ほど紹介したホッブズ的秩序問題が登場するのは、20世紀になってからである。ホッブズ的秩序問題と命名したのは、アメリカを代表する社会学者Talcott Parsons（タルコット・パーソンズ）である。ドイツで社会学を修めたパーソンズは、ホッブズからカントを経てドイツ社会学へと引き継がれた問題意識を受け継ぎ、価値や規範の共有が秩序を維持する要素であると考えた。

信頼の系譜を考える上では、心理学も重要である。今日、我々が知っているような実験心理学は、1879年にドイツの心理学者Wilhelm Wundt（ヴィルヘルム・ヴント）が実験心理学研究室を開設したのが始まりとされている。ドイツの心理学者Kurt Lewin（クルト・レヴィン）はナチスに追われてアメリカに移住し、MITでグループダイナミクス研究センターを設立し、以降、社会心理学が広まった。ヴントとレヴィンも、カントを背景にして心理学を考えた。

20世紀になるとアメリカの社会心理学者Morton Deutch（モートン・ドイッチ）が登場した。彼はレヴィンの教え子であり、「紛争（解決）」や「協調」を主な研究テーマとした。彼は「囚人のジレンマ」を用いた信頼研究を最初に実施し、社会集団内の紛争解決の要因として信頼を位置付けた。その後、信頼を主観的確率（ないし期待）として理解するのが一般的になっていった。主観的確率として信頼を考えるのは、一つの標準的な考え方であり、冒頭で言及したさまざまな分野の信頼の定義にも入っているが、これはドイッチの囚人のジレンマを用いた信頼研究とともに広がっていったと考えられる。

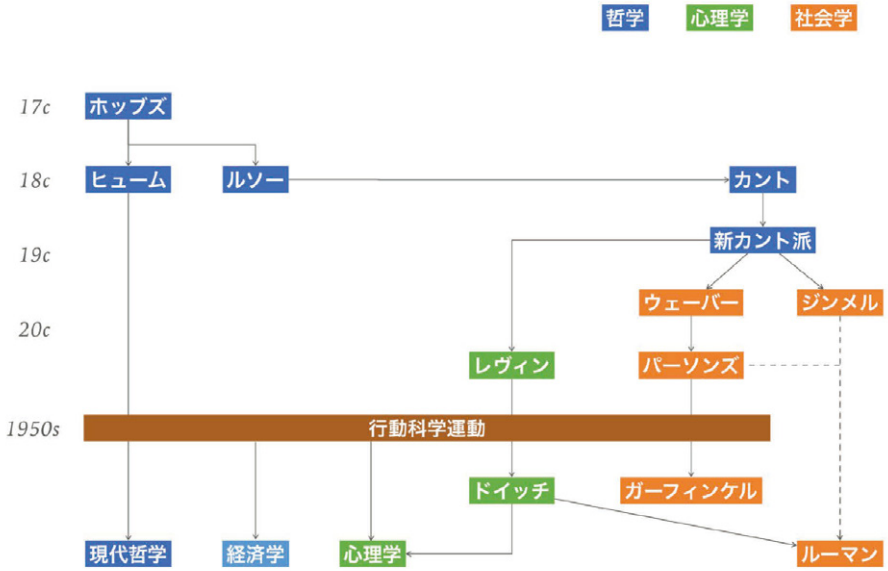


**図2-1-3　　信頼研究の系譜**[3]



**行動科学運動**

囚人のジレンマをベースにした信頼の考え方が広がっていった背景には、1950年代の行動科学運動があった。これは社会学や心理学が中心となって社会科学系の諸分野で学際研究を推進する運動を指す。パーソンズは行動科学運動に積極的に参加した社会学者の代表である。アメリカ国立科学財団（NSF）の設立時（1950年）に、心理学者、社会学者、人類学者が「行動科学」という名称で自分たちの分野も加えるように要請した。最終的に行動科学の名称としてではなかったが、NSFの中に社会科学系の分野が設置された。また、行動科学運動の広まったもう一つのきっかけとして、フォード財団が1950年代に行動科学プロジェクトを実施し、スタンフォード大学やハーバード大学、シカゴ大学に莫大に助成金を投入した。その結果生まれた組織の一例が、スタンフォード行動科学先端研究センターである。このように、行動科学運動は、現在の学際研究や異分野融合、分離横断と呼ばれる走りとなったが、1960年代には廃れた。

この時期に行動科学運動が起こった背景には、第二次世界大戦中の軍産学複合体がある。戦時中は、大学が軍から委託された学際研究を行うことが当然のように行われていた。冷戦を理由に、戦後もこの体制は継続した。太平洋戦争が始まる前、アメリカでは国防研究委員会が設立されたが、戦後、この組織は解体され、後続組織として紆余曲折の結果生まれたのがNSFである。ベトナム戦争の頃まで、第二次世界大戦中に広まった軍産学が集まって学際研究を行うという体制は続いた。行動科学運動の目的は、軍（政府）からの研究委託が途切れないようにするためであった。

囚人のジレンマは、そもそも、紛争解決に関する研究として使われていた。1950年、カリフォルニアのランド研究所で実施されていたゲーム理論の研究から囚人のジレンマ・ゲームは生まれた。ランド研究所は世界初の軍事シンクタンクであり、冷戦下の戦略研究のためにゲーム理論の研究が熱心に行われていた。1950年代のアメリカでは軍からの資金で学際的な研究をすることが盛んであった。ドイッチも海軍の委託プロジェクトを受けて、ゲーム理論の専門家との交流を通して、紛争解決や平和構築のような研究に行き着いた。ドイッチの研究は社会心理学者の間で広まるというよりは、基本的には行動科学関係者の間で、囚人のジレンマを用いた信頼研究、および「主観的確率としての信頼」として広まった。

3　『信頼を考える－リヴァイアサンから人工知能まで』[1]から図を引用

**その後の展開**

その後、行動科学運動は、さまざまな分野に影響を残すこととなった。社会学では、ドイツの社会学者Niklas Luhmann（ニクラス・ルーマン）が、ジンメルとドイッチの研究に依拠して信頼論を展開した。パーソンズ社会学の批判から始まったとされるエスノメソドロジーの分野でも、最初期には信頼研究が行われたが、相互行為についての着目するようになるにつれて、信頼という言葉はあまり使われなくなった。また、1980年代になるとイギリスの社会学者Anthony Giddens（アンソニー・ギデンズ）らによってリスク社会論が登場し、信頼が注目されるようになったが、これは別の系譜に属する。

心理学では、社会心理学者山岸俊男の信頼論は、ルーマン信頼論を批判したものだが、ルーマンが依拠している同じ社会心理学者のドイッチの議論は無視している。山岸は安心と信頼をどちらも認知バイアスとして定義しているが、ドイッチは信頼を紛争解決の要因とみなしており、両者は信頼の異なる側面に注目していると言える。現在の社会心理学では、信頼に関するさまざまな尺度が作成されている。

政治学では、1950年代から行動科学運動の影響があり、1960年代にはパーソンズのシステム理論に基づく政治文化論が生まれるが、信頼がクローズアップされることはなかった。1990年代にはアメリカの政治学者Robert Putnam（ロバート・パットナム）によってソーシャルキャピタル論が広まり、その中で信頼研究が行われるようになったが、これは別の系譜に属する。パットナムを批判するアメリカの政治哲学者Russell Hardin（ラッセル・ハーディン）の信頼論は、「主観的確率としての信頼」を発展させたものである。

他にも、経営学は行動科学の一分野であり、経営学での信頼は主に「主観的確率としての信頼」を発展させたものとして使われている。哲学では1980年代以降、「主観的確率としての信頼」を「Reliance」とし、信頼（Trust）と区別するようになった。教育思想での信頼は、協調行動や紛争解決などとは異なる種類の信頼、「無償の愛」に近い概念として使われている。医療人類学・医療社会学での信頼は「トラスト」以外にラポール（Rapport）がある。

**【主な質疑応答】**

Q：科学と社会の関係の重要性に焦点が当たるようになりトラストの重要性を皆が認めるようになってきた一方で、歴史的・理論的基盤が共有されていないと感じている。今後どのような取り組みが必要になるか？

A：どの分野も自分たちの土俵で物事を見ており、他の分野ではどうなのかという発想に至らないところに課題がある。個人としてはどれだけ親密でも、研究の話になると自分たちの分野の枠組みでしか捉えようとしないところが大きなネックになっている。これは必ずしも悪いことではないけれども、厳密さの追求だけではうまくいかない。問題を安易に考え過ぎているところがある。何らかの形で他の分野に対するリスペクトが必要。実際、行動科学運動も他の分野に対するリスペクトがあったからうまく行き、リスペクトが失われることでうまく行かなくなったという側面がある。

Q：宗教的な影響をどのように考えるか？

A：信頼はさまざまな現象と紐づけられて考えられるため、一般論で言うと、影響が大きいものもあるだろうし、そうでないものもあるということになる。そのため、何らかのシチュエーションを特定して、それに関して判断する必要がある。宗教学だと信頼ではなく信仰の話になると思われるが、現状あまり把握していない。

Q：機械や技術、システムへの信頼に関する研究の展開はどのようになっているか？

A：ロボットに関しては、コンスタントに研究が行われている。基本的には社会心理学的な方法論を使って研究を行うが、社会心理学のジャーナルやコミュニティーに向けてではなく、各自の分野で発表を行う。

Q：「安全信頼技術研究会」ではどのような議論がなされているか？

A：現在は「不信」をテーマにしており、「不信学」を作るための研究助成を受けている。行政に対する不信や、食に対する不信など。年金に対する不信も取り上げた。理工系、情報系の話よりも社会に関す

る問題を議論している。

Q：哲学がトラストの議論に貢献していくことは考えられるか？

A：難しいと感じている。哲学の研究は、いろいろな分野のトラストが「何を背景にしているか」という観点から検討を行うため、社会心理学のような実験を使うアプローチとは次元が異なる。他の分野との関係はあるが、接点を作るのは難しい。他の分野で行われているトラストの議論を、哲学のコミュニティーに向けて発表することも難しい。

Q：どうやれば信頼は上がるかという信頼の上げ方に関する研究はあるか？

A：信頼を上げる要因を特定したり、実際にそれをコントロールしたりする研究は、社会心理学者が発展させた方法論の応用として行われている。一方で、そのアプローチが、他の全然違うシステムでもうまくいくのかというと、必ずしもそうではない点などの課題もある。

Q：デジタル技術やデジタル社会の中でトラストを考えることが今回のセミナーシリーズの趣旨だと思っている。歴史的な系譜からお話しいただいたが、トラスト研究について現在起きていること、これから起きそうなこともお聞きしたい。

A：あえて言うならば、個人的にはデジタル化によって信頼について何か全く新しいことが起こるという感触はない。信頼は、結局のところ、曖昧な概念であり、今後もそのまま曖昧なまま語り続けられるので、何も変わらないであろうと考えている。今回のようにトラストはこう考えるべきという、ある程度の合意が取れるよう状況が作られるとすれば、その後、変わるかもしれない。

Q：安心は分かりにくいが、不信はすごく分かりやすい。そういうネガティブアスペクトの方向から攻めていった方が、技術であれ、社会であれ、分かりやすいのかと思う。

A：まさにおっしゃる通りで、信頼や安心は、それ自体をかっちりしたものとして捉えるのではなく、それが成り立っていない状況が顕在化することで、そうではない状態があったことが認知されるという風に考えるのがむしろ一般的だ。だからこそ、それぞれのアスペクトによって違うということが本質的ではないかと思う。信頼についてさまざまな現象があるというよりも、もし今の状況で何かが成り立っていないならば、それが成り立っている状況が信頼のある状態であると考えて、何が成り立っていないのかを考える方が研究として筋が良いと思う。

Q：組織や集団を信頼することをどのように考えればよいか？　国や会社、あるいは、AIが複数つながってネットワーク化し、さらにそこに人間という要素もAIと共同作業する形で加わる複雑な構造を持ったネットワークに私たちは囲まれている。信頼する相手が非常に複雑であるだけに十分な情報を得ることはほぼ不可能である。あるいは原理的にも時間的にも変化していくため、情報を確実なものとして得ることができない。そうすると、信頼とは、一種の思考停止だと言える。それがどのくらい確率的に危ないことなのか／危なくないことなのかということを議論するのが筋だとう思うがどうか？

A：信頼がある種の思考停止だというのは、多くの分野での共通的な見解だと思う。哲学では、信頼は不合理性の一例として考えることが結構ある。紛争解決でもそうだが、全ての情報があればというのではなく、どこかで思考停止することによって初めて信頼が可能になるところがある。思考停止しているからこそ、我々の社会はうまくいっているみたいなところもある。何が危険かというリスクの話と、どのように信頼しているかという話は、それほどストレートな関係ではないかもしれない。だからこそ、これらをリンクさせる研究は少なく、逆に安易にリンクさせる研究の中には危ないものを信頼させる、より思考停止させるにはどうすればよいかの研究もあったりするので、いかに安全なものを信頼できて、安全ではないものを信頼しないようにするにはどうすべきかが重要なテーマだ。

Q：失墜した信頼をどのように回復していくかといった研究事例があれば教えていただきたい。

A：科学技術に対する信頼回復の話として、NASAが情報公開を積極的に行うことで信頼の低下を最小限に留めたという例はある。ただし、情報公開はあらゆるケースで有効なわけではない。宇宙開発の場合はベースの期待値が高いため、情報をオープンにすることで信頼が下がらないというようなことがあ

るかもしれないが、そもそもベースの期待値が低い、今だったらAIや原発では単にオープンにすることではうまくいかないと考えられる。

Q：ジーマクレジットのような個人信用スコアが中国で大はやりだが、日本でもはやるかと聞かれた際、非常に疑問であるという話をした。なぜかというと、中国社会は不信がデフォルトになっているのに対し、日本社会では信頼がデフォルトになっている。日本での信用スコアは、「お前は信用に値しない」と言われるという意味だから、そんなサービスは誰も利用しないと考えられる。先ほどの話と同じように、不信の欠如として信頼が概念化される可能性はかなりあると思われる。信頼と不信のどちらがデフォルトかは社会の在り方や対象によって大きく異なるという印象を持っている。世界のどこかには、出会った瞬間に普通は殺し合いをするのだけど、知っている人の名前や、家族の名前を順番に全部あげていって一致する奴がいたら、それは仲間なので殺さないというような、デフォルトが殺すという社会もおそらく存在する。そのあたりが難しい問題だと思っているが、ご意見をお聞きしたい。

A：すごくややこしいところであるが、信頼とは呼ばないけれども似たような現象として、相手の行動を予測できる状況（先の例では、相手を信頼しているわけではないが、何かあったら斬りかかってくることをみんなが理解している状況）は、信頼とは明らかに呼べず、区別すべきだと考えられるが、社会の構成員がそれを自覚し、それをベースに動いているという意味では、信頼で動いているのとほとんど同じであると言える。従って、これを不信と呼ぶべきか怪しい。つまり、もう少し概念が細かく必要になってくる。過剰な信頼や過小な信頼の問題点もあるので、信頼／不信という区別を超えた概念が必要であろう。

また、信頼する／信頼しないを口に出して言うかどうかは重要になってくる。口にしないときに成立している現象と、あえて「信頼している」、「信頼していない」と口にするかどうかがより本質的な問題であるような現象もあると思われる。その際に信頼／不信で区別してしまうと見えないものが出てくる。予見可能性の問題と、例えばフレンドリー／アンフレンドリーの問題が混在している。

Q：今の議論でもあったように、用語をいくつか整理しないといけない。信頼（Trust）、信用（Believe）、信仰（Faith）など、それなりに分かっているつもりではいるが、RelyとDependがいまだに分かっていない。例えば、Googleを使って検索するけれども、本当は情報を渡したくない。でも検索エンジンについてほとんどの人たちはGoogleしか知らないので、使わざるを得ない。この状況はDependとは言うけれども、信頼（Trust）とは言わない。細かい話かもしれないが、こうした用語を細かく使い分けていくのが、信頼という言葉を乱用しないで済ませるためにどうすればよいのかということにつながるのではないか。

A：人間の言葉というものは、そんな厳密に使い分けられていないという点は大事だ。個人差や状況差が激しい。言葉にはそもそもズレがあり、手がかりでしかないという点を押さえるのは重要だろう。

哲学での信頼（Trust）とRelyの違いというのは何かというと、信頼は自分の期待が裏切られたときに、怒りのような感情的な反応を起こすものであるのに対し、Relyはそれがうまくいかなかったとしても、自分の予測していたものが外れたということで困ったりするが、そこで怒りを感じるほどではないとされている。信頼のさまざまな要素に関する心理学的な研究でも、信頼は認知的な、ある種の計算に関連する要素と、それとは別の感情に関する要素の両方あるといわれており、それに対応する話である。

## 2.2 上出 寛子[4]「社会心理学におけるトラスト」

ここでは他者や集団を対象とした対人的信頼に関しての心理学的知見を紹介する[1]。まず、社会心理学の古典的な研究から、情報源や話者に対する信頼について取り上げる。次に、社会的認知における普遍的な2つの認知次元とされる「温かさ（Warmth）」と「能力（Competence）」に関する研究の展開を紹介する。そして、世界的にも著名な山岸俊男の信頼研究の意義について論じる。

**情報源に関するスリーパー効果**

社会心理学の古典的な研究にHovland（ホブランド）らの信頼の研究がある[2]。この論文を書くまでにHovlandらは「スリーパー（潜伏工作員）効果」を発見していた。ある情報を得る際に、コミュニケーションの直後よりもある程度時間が経過した後の方が、情報提供者の意見へ説得されるという効果である。

1951年の論文では、情報源（ソース）の信頼性によってこの効果がどう異なるのかを検証している。実験は3段階に分けて実施される。まず5日前に情報源に関する一般的な信頼を聞いておく。当日の実験では、研究生が教室にやってきて、「あるプロジェクトのデータ収集のため、最近の雑誌や新聞に掲載された記事からいくつか抜粋して掲載した冊子を読み、短いクイズに答えてください」と言う。クイズでは、記事の意見に対してどの程度同意するか、すなわちどの程度説得されているかといった項目や、記事の情報源は何だったかなどについて尋ねる。そして4週間後にも同様のクイズを実施した。

冊子の内容は図2–2–1のようになっている。抗ヒスタミン剤や原子力潜水艦など、当時議論を呼んでいた4つのトピックを選択した。そして、記事の内容は同じものの、その情報源が信頼性の高いものかどうかを操作し、さまざまな組み合わせの冊子を作って学生たちに配布した。

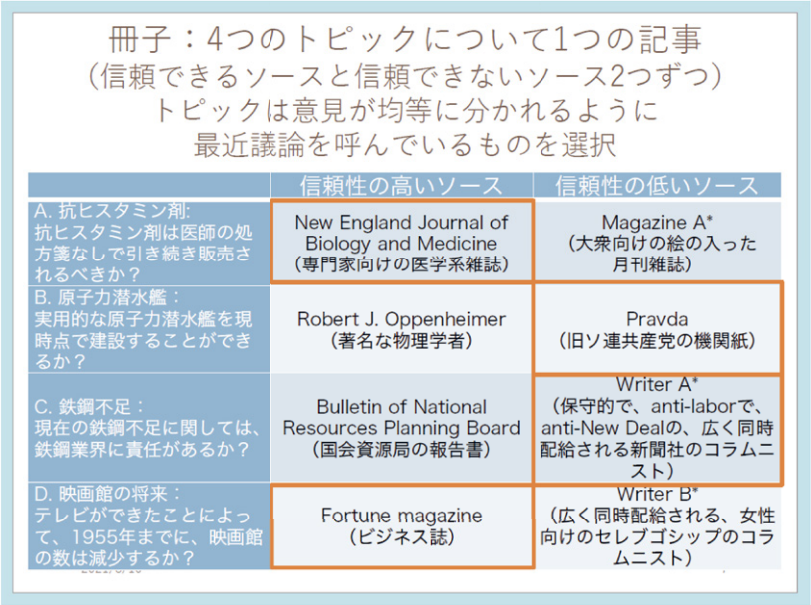


**図2–2–1　　Hovland and Weiss（1951）の実験**

4　名古屋大学未来社会創造機構 特任准教授
http://kamidehiroko.jp/

まず、情報源に対する信頼の認知に違いがあった。次に、内容が事実に基づいた正当・公平なものかどうかを尋ねると、信頼できる情報源の内容に対して、信頼できない情報源よりも正当で公平だと認識していたことが確認された。4週間後の実験でその情報源を覚えていたかどうかを尋ねると、信頼できない情報源については、当初同意していた人の76.7%が覚えているのに対して、当初反対していた人では55.3%しか覚えていなかったという結果になった。

4週間後のクイズでは、信頼できる情報源（実線）で初めは高かった同意の割合が下がり、信頼できない情報源（破線）で初めは低かった同意の割合が上がっている（図2–2–2）。これは情報源を忘れて、同意への抵抗感が薄らいだと考えられ、信頼できない情報源で「スリーパー効果」が見られる。

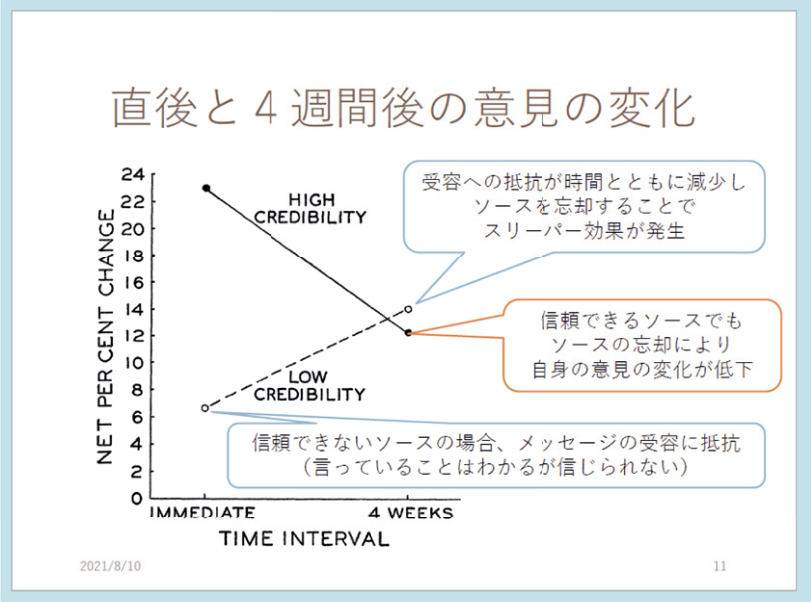


**図2–2–2　　直後と4週間後の意見の変化（Hovland and Weiss 1951）**

**不確実性の伝達方法による信頼への影響**

最近の論文でもコミュニケーションにおける信頼について検討しているものがある[3]。ここでは事実や数値に関する不確実性の伝え方が、それを受け取った人の情報やソースに対する信頼に及ぼす影響を検証している。

これまで、推定値の範囲を示して不確実性を伝えることは、正直であると認識される一方で、不確実であるという点でやはり良くない印象を与えることが指摘されてきた。そこで不確実性を伝える際に、「最小〇〇人、最大〇〇人」と数値的に不確実を伝えた場合と、「この値は多少高くなる可能性も低くなる可能性もある」と言語的に不確実性を伝えた場合を比べたところ、言語による不確実性の伝達の方が、数値や情報源への信頼をより低下させることが分かった。そのため、この論文では、不確実性の伝達は点推定値を伴う数値範囲によって行うのが適切だと提言している。

**非言語行動による話者への信頼**

非言語行動の振る舞いによって話者に対する信頼がどのように変わるのかという研究もある。 Burgoon（バーグーン）らの研究（1990）では、60人の学生がスピーチを行い、他の学生がその信頼性と説得力を評価した[4]。話者の信頼性に関しては、信頼できるという「能力」や、「社交性」、落ち着いているといった「平静さ」、正直そうであるという「特性」、饒舌であるという「活動性」の5つ、また、話者の説得力としては10項目で評価している。その後、40時間の訓練を受けた2名のコーダーが先行研究に基づいて非言語行動

を評価した。

結果は図2-2-3のようになった。「能力」と「平静さ」は、声の手がかり（特に流暢さとピッチ）が、動作的な手がかりよりも大きな役割を果たしていた。一方、正直そうであるという「特性」と「社交性」は、声の手がかりよりも動作的な手がかりの方が関連していた。「説得性」に関しては多くの非言語行動が重要であることが示されている。こうした非言語行動のセットが、「活動性」の次元を除く全ての次元で信頼性・説得性を促進していることが定量的に明らかになった。

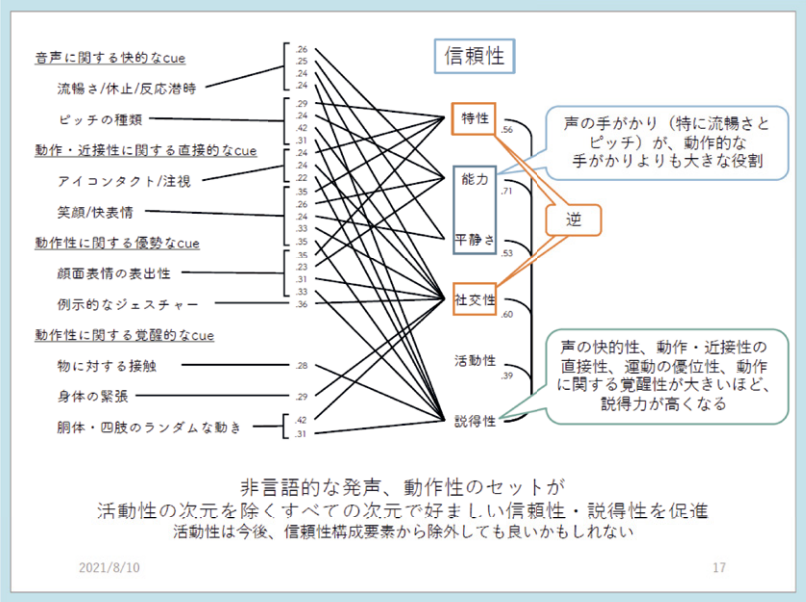


**図2-2-3　信頼性の5因子と説得性と、非言語行動の関連性（Burgoon, Birk, and Pfau 1990）**

**信頼の社会的認知：温かさ（Warmth）と能力（Competence）**

もう一つ、社会心理学の中で信頼を扱う大きな分野に社会的認知がある。これは、人や集団に対する認知である。社会的認知は進化の圧力を反映してきたとFiskeは言う[5]。社会性のある動物は、同種の生物と遭遇したときに、その相手が敵か味方かを即座に判断し、次に、相手がその意図を実行する能力を持っているかどうかを判断しなければならない。この社会的認知の普遍的な2つの次元、すなわち「温かさ（Warmth）」と「能力（Competence）」が個人や集団に対する認知で確認されている[6], [7]。

「温かさ」の次元には、親しみやすさ、親切さ、誠実さとともに、信頼性（Trustworthiness）が含まれる。これは、道徳性など、知覚された意図に関連する特性と定義されている。「能力」の次元に関しては、知性、技能、創造性、効力など、知覚された能力に関連する特性である。 Fiskeは人が自発的に行動を解釈したり他人についての印象を形成したりするとき、この2つの基本的な次元が、人が他人をどのように特徴づけるかをほぼ完全に説明していると結論づけている。

「温かさ」の判断は素早く行われる。進化的な観点から見ると、他者の善意や悪意は生存にとって重要である。実際、100ミリ秒の露光時間後に顔を判断する際、認知者は信頼性を最も正しく判断し（時間の制約がない場合と同じ判断をする）、次いで能力を正しく判断している。なお、「温かさ」の検出を優先するのは、男性よりも女性、個人主義文化よりも集団主義文化の方であるという報告もある。

この「温かさ」と「能力」の評価は、個人に対しては正の相関が、集団に対しては負の相関がある。集団に対しても、高齢者・身体障害者・精神障害者は温かさがあるが能力はない、金持ち・アジア人・ユダヤ人・女性の専門家・マイノリティーは能力があるが冷たいとイメージされている。さらにこれは、さまざまな感情

（尊敬、嫉妬、同情、軽蔑）と行動（協力、危害）につながっている。信頼に関してもこういった感情と行動が伴うということは容易に予測される。

Fiskeはコミュニケーターとしての科学者の信頼についても検討を行っている[8]。アメリカ人の職業イメージを４つのクラスターに分類すると、看護師・教師・医師などは温かさがあり有能だと評価されているが、ごみ収集人やファストフードの店員などは冷たく有能でないと評価されている。他方、研究者・エンジニア・科学者は能力が高いが温かさはあまり感じられないと考えられている。コミュニケーションには地位や専門知識だけでなく、信頼性（温かさ）も重要なので、能力があって温かくて信頼できる看護師や教師の存在に注目し、信頼できるコミュニケーションの可能性を考える重要性を指摘している。

**コンピューターエージェントに対する信頼**

この文脈で、ヒューマンコンピューターインタラクションにおける信頼と社会的認知の研究がある[9]。信頼性は「温かさ」と「能力」の両方に基づくと考えられる。この実験では、参加者がコンピューターエージェントと一緒に、Ｔ字とＵ字のみのテトリスのようなブロックパズルをする（図2–2–4）。参加者はまずエージェントにＴ字かＵ字のブロックを置くようアドバイスし、エージェントがいずれかのブロックを選択して置いた後、参加者は残った方のブロックを置く。各プレーヤーが置くと、Ｔ字は５点、Ｕ字は10点が得られ、横一列がそろって列が消えると双方にボーナススコアが入る。ここでエージェントの振る舞いを「能力」と「温かさ」について操作する。「能力」については効率的な配置をするかどうか、「温かさ」については人のアドバイスに従うか、Ｕ字ブロック（10点）ばかり取らないか（すなわち、振る舞いが利己的かどうか）、という操作をしている。

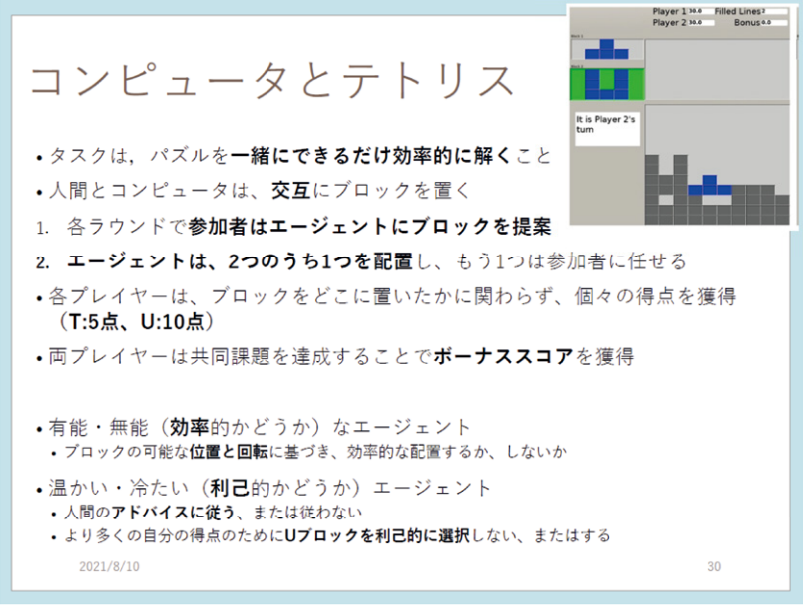


**図2–2–4　　パズルのルールとエージェントの操作（Kulms and Kopp 2018）**

そのタスクの後にいわゆる信頼ゲームをする。参加者とエージェントはそれぞれ４つのトークンを情報交換なしで自分と相手に分配するが、相手に渡したトークンは価値が２倍になる。このときにエージェントに渡したトークンを行動信頼度（Behavioral Trust）として測定した。また、エージェントに対する「温かさ」、「能力」、「信頼性」を主観的に評価した。

図2–2–5は、エージェントの効率性と利己性によって「温かさ」と「能力」がどう変わるかを示している。エー

ジェントが効率的であれば、利己的な行動によって「温かさ」は減る。一方、エージェントが非効率的であれば、利己的な行動によっても「温かさ」はそれほど変わらない。「能力」の認知に関しては、エージェントのパズルの効率性にのみ影響を受ける。

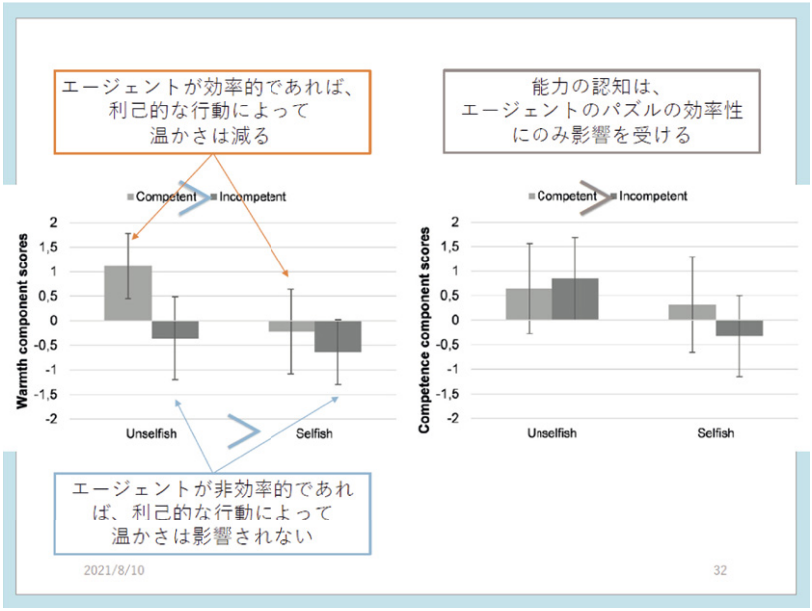


**図 2-2-5　エージェントの効率性と利己性による「温かさ」と「能力」（Kulms and Kopp 2018）**

行動信頼度であるトークンの分配については、先ほどの「温かさ」の次元と同じように利己性と効率性による正の効果は出ている（図 2-2-6）。しかし、非効率的なエージェントの場合、信頼は利己的な行動の影響を受けないが、効率的なエージェントの場合、利己的な行動によって信頼が下がるといった結果が出た。

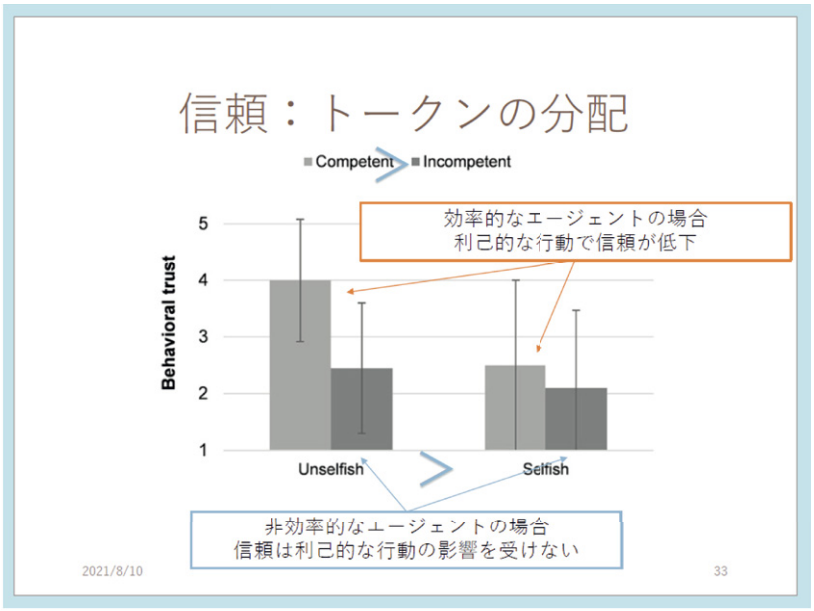


**図 2-2-6　トークンの配分（Kulms and Kopp 2018）**

Koppらは、Fiskeの言ったような人に対する「温かさ」と「能力」の認知次元はコンピューターにもやはり移行すると考える。それがコンピューターの場合はトラストにも関係する。

**信頼と安心：山岸俊男の信頼理論の意義**

世界的に著名な山岸俊男の信頼と安心の議論は、特殊な理論構築をしている[10]。山岸は、ゲームプレーヤーとしての個人を想定して、社会的な行動（他者との相互作用）を扱う。信頼（Trust）の最も広い定義は「自然的秩序および道徳的社会秩序の存在に対する期待」である（Luhmann、Barber）[11], [12]。山岸はこのうち、自然的秩序については除外し、道徳的社会秩序の存在に対する期待、すなわち「社会的ジレンマ状況における成員の協力行動の促進要因としての他者の協力性に対する期待」に注目する。

それをさらに分類すると、①社会関係や社会制度の中で出会う相手が役割を遂行する能力を持っているという期待と、②おそらく相手が責任を果たすこと、また、そのためには場合によっては自分の利益よりも他者の利益を尊重しなくてはならないという義務を果たすことに対する期待があって、いわば①が能力に対する期待、②が意図に対する期待である。これはFiskeの「能力」と「温かさ」に対応しているとも考えられる。このうち、山岸は能力でなく意図に対する期待を扱う。

意図に対する期待のうち、山岸は相手の自己利益の評価に基づくものを安心（Assurance）、相手の人間性に由来するものを信頼（Trust）、と区別する。のどの中に「針千本マシン」が入っていて、うそをつけば針千本を飲み込まないといけないという状況で、相手が自分にうそをつかないというのは、単なる安心である。そういった証拠がない不確実な状況にもかかわらず相手が自分に協力するだろうという期待、すなわち客観的な確実性がないにもかかわらず相手に対する期待を持つという認知的なバイアスを信頼として扱う。

整理すると図2–2–7のようになる。まず自然の秩序に対する期待は外す。また、相手の能力に対する期待も信頼の対象にはならない。そして、相手の自己利益に基づく相手の意図の期待は安心と呼んで、信頼としては扱わない。

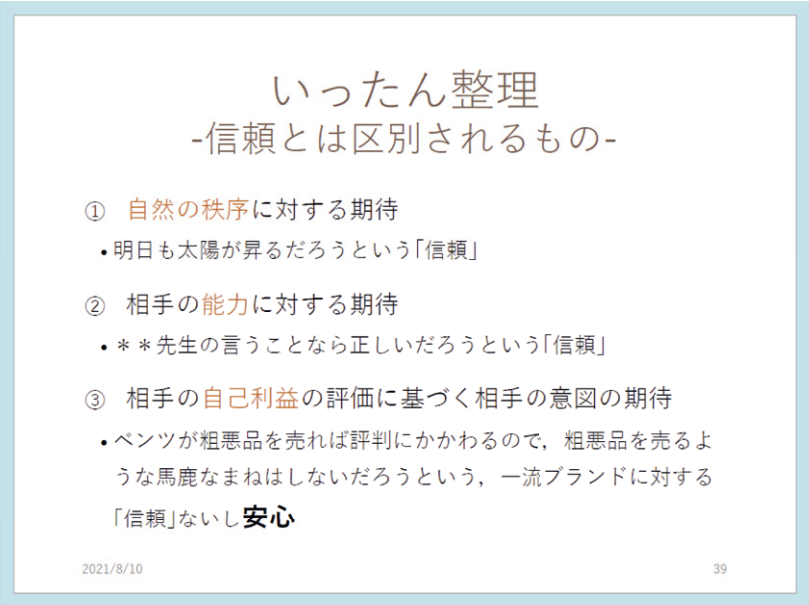


**図2–2–7　　山岸理論において信頼とは区別されるもの**

社会心理学において山岸の信頼の理論は有名なので、あたかも私たちの世界における信頼という現象についての普遍的な理論だと考えてしまうことがあるが、いま示した通り、ここでの信頼は非常に限定されている。

だからこそ、理論と実験との精緻な整合性がすばらしく、さまざまな知見が生まれている。

**内集団びいき**

例えば、山岸のグループではどれぐらい内集団びいきが確認できるのかを17か国を対象にオンラインによる信頼ゲームをして調べた[13]。参加者は、同じ国籍の相手（イングループ）、異なる国籍の相手（アウトグループ）、そして未知の国の相手（ストレンジャー）と信頼ゲームを行った。また、参加者はお互いが相手の国籍を知っていることを認識している条件（共通知識条件）と、一方的に相手は自分の国籍を知らないことを知らされている条件（一方的知識条件）も比較した。図2–2–8左にあるように、同じ国籍の相手に対しては多くを分配している。また、同図右にあるように相手が自分の国籍を知っている場合は信頼行動が高くなる。

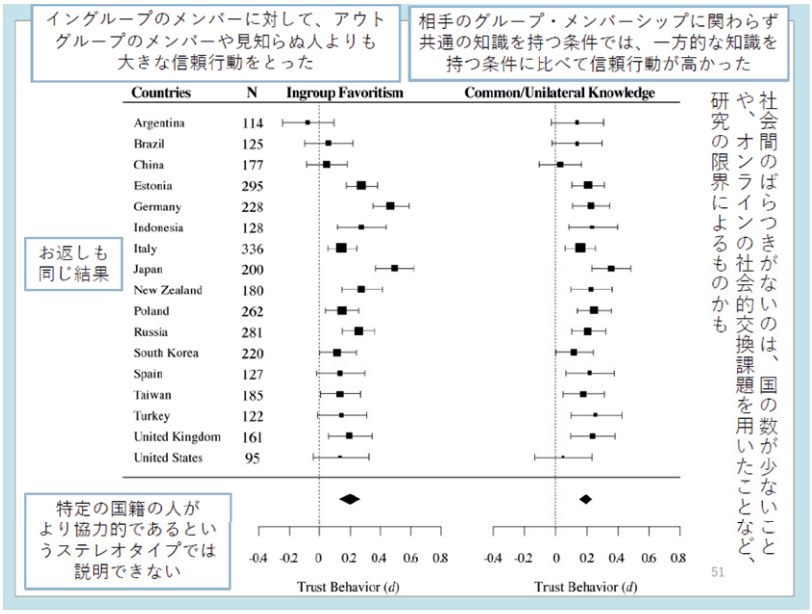


**図2–2–8　　内集団びいき（Romano et al. 2017）**

山岸理論に対する批判の一つとして、現実的に、相手の評判や情報がない不確実な状況で協力するか裏切るかという選択を迫られるという状況が、それほど一般的ではないのではないか、という指摘がある。ただしそうした場面で、見知らぬ人が別の信頼される人に似ているというだけで信頼されやすいということを検証した研究もある[14]。

参加者はどの顔のプレーヤーが信頼できるのかをまず学習する。次の信頼ゲームで一緒に遊ぶ相手を選択するときに、1回目に遊んだ人たちをモーフィング（合成）した顔を提示する。そうすると、元の信頼できるプレーヤーと似ている顔ほどこの人と一緒に信頼ゲームをやりたいという確率が上がるという結果が出た（図2–2–9）。しかも、被験者は次の信頼ゲームの相手として提示された顔が、1回目のゲームの相手プレーヤーの顔のモーフィングであるとは気づいていなかった。すなわち、過去の経験によって信頼の行動が意識的にも無意識的にも変わってくるといった一つの知見になる。

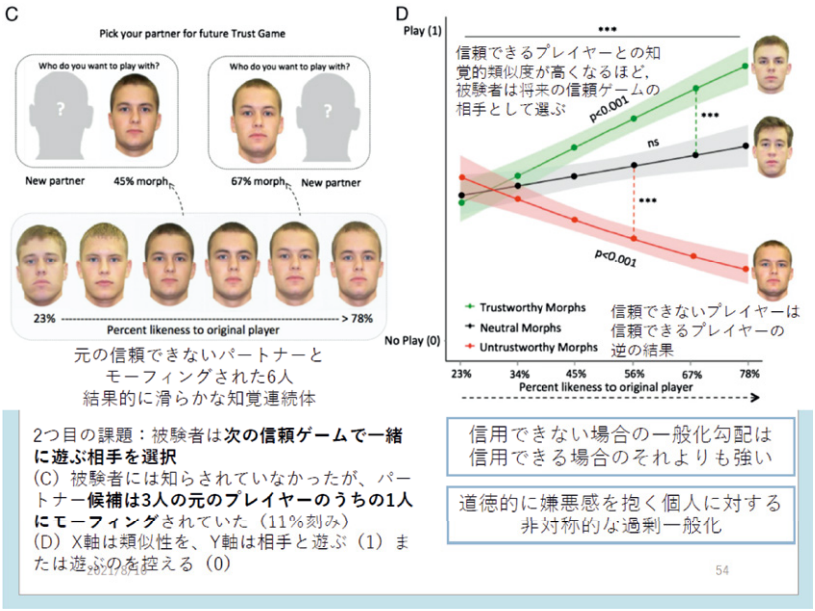


**図2–2–9　　信頼できる人に似た人を信頼する（Oriel FeldmanHall et al. 2018）**

### Partnership on AIによる文献レビュー “Human–AI Collaboration”

他にも信頼に関する研究は多く発表されている。 Partnership on AIによる2019年の文献レビューでは、7つの知見が示されている（図2–2–10）[15]。

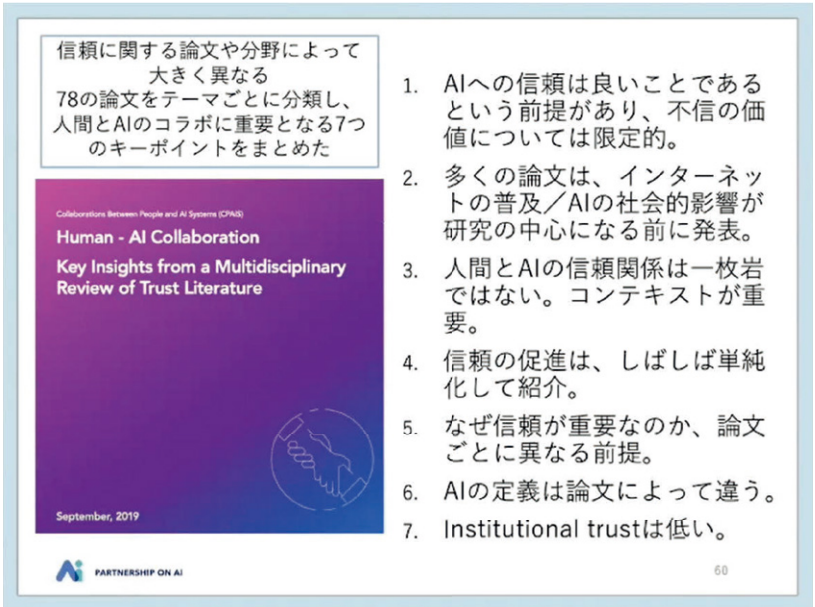


**図2–2–10　　人とAIの協働に関する研究についての知見（Partnership on AI 2019）**

**【主な質疑応答】**

Q：AIなど新しいタイプのソフトウェアに対する信頼について、「温かさ」と「能力」以外にどのような軸が考えられるか。

A：温かいけれども信頼できない可能性に関しては、予測可能性が一つの軸として入ってくるだろう。AIや

ロボットは、私たちの社会にとって新規参入者なので、それらが信頼できるという経験をどう蓄積できるかが重要だろう。

Q：AIの場合、我々には予測できないけれども我々にとって良いことをすることがある。それがAIに対しての信頼性の作り方の難しい部分であり、特にEUの人たちはAIをそのような存在として見たくないようだ。そのあたりが今後のテーマとして重要だと考えている。

A：結局良いことをやってくれていたんだという解釈に至れば良いが、高齢者の場合など、AIのようにどのような手続きでそのような結論になっているのかが不透明であれば、不安を感じる場合も出てくると思うので、ユーザーの特徴や使用される文脈に対して慎重になる必要はあると思う。

Q：今回はレビューに徹していたが、上出先生の研究やプロジェクトにこれらの知見がどのように生かされているか。

A：2009年から始めているロボットに対する安心感の研究は、今も筑波大学や大阪大学の先生との共同研究として続けている[5]。人共存型のサービスロボットに対する一般のユーザーの安心感の評価軸としては、「温かさ」のような心理的に一緒にいると癒やされるとか快適に感じるというポジティブな効果と、一方、一緒にいるとストレスがたまるし、いらいらするというネガティブな感情的要素がある。また、人の言うことをやってくれるという能力の高さと、一方で、人間のコントロールからは外れてしてほしくないといった要素が分かってきた。このように、高齢者支援や発達障害者支援のロボットに対する安心感を評価するなどしている。

名古屋大学で研究しているモビリティー技術の社会的受容は、信頼よりも枠組みが大きい。社会的受容性は1980年代にドイツで原子力の話題が盛り上がったときに議論が始まった。現在では、再生可能エネルギーなど新規の技術についての研究がはやっていて、導入する上での市場の受容性、社会・政治的受容性も一緒に考えないといけない。最近だと、ノーベル経済学賞を取ったOstrom（オストロム）が「サイエンス」に短い論文を書いている[16]。アメリカのメイン州のロブスター業が持続可能であるために、例えばその市場や社会や消費者がどうあるべきかみたいな複雑な議論をしている。そこまで複雑ではないが、名古屋大学では地域の方々と一緒になって仕組みを導入しようとしている。その中で、確かに行政や企業に対する信頼は、社会的受容の重要な視点として指摘されているし、社会的受容は人間にとっての適応性だけでなく将来の持続可能性を踏まえた議論でもあるということだと考えている。

Q：情報化やデジタル化に伴って、社会心理学の研究テーマは変わってきているか。

A：社会心理学会のプログラムでは社会問題として生じているインターネットやSNSというセクションはあるが、人工知能、ロボットはまだない。ただ、理系の先生との連携もフレキシブルにされている先生方はいる。アメリカ心理学会（APA）では、2019年に雑誌「Technology, Mind & Society（技術・心・社会）」を創刊した。また、APAのブックレット「Interconnections of Psychology and Technology」（2020年）も人と技術の相互作用を扱っている。

Q：MITのメディアラボなどが2016年から始めたモラルマシンというオンライン実験[6]では、国や性別、年齢といった属性によって回答の傾向が異なる[17]。その結果を見て、それぞれの属性に合わせた自動運転車のアルゴリズムや法制度にするといった可能性はあるか。

A：技術と人間の調和的な姿を考えるときに、一つは人々の反応に合わせていくというやり方があるが、ある程度は共通のプロトコルを作って一般の人に伝えていくことも重要である。トロッコ問題は考えること自体が問題を深刻化させる可能性があるという指摘もある。そういう状況に本当になるのか分からな

5 http://kamidehiroko.jp/other.html

6 https://www.moralmachine.net/

いし、どちらを被害者にするかはあらかじめ議論してもし尽くせない話であって、人々の傾向を自動運転に反映させることは慎重にすべきだろう。

## 2.3 犬飼 佳吾[7]「行動経済学・実験経済学とトラスト」

**社会科学および経済学研究の変遷**

社会科学は、文献研究や既存のデータに基づく研究が主流であったが、2000年代前半あたりから実験を取り入れる研究が急増した。特に、人の社会性に関する神経科学（神経経済学や社会神経科学）の研究が注目されるようになり、NatureやScienceなどの一般科学誌に多くの研究成果が発表されるようになってきた。その中でも「トラスト」は、社会性に関わる分野の一つとして取り入れられるようになった。

社会科学の潮目が変わり始めたのは、2010年代以降である。ビッグデータ全盛の時代になり、我々が持っているデータや、やり取りする情報量が劇的に増大した。特に、スマートフォンやSNS（Social Networking Service）は、ネットワークの科学との親和性が高く、それらを用いた研究が多く展開されるようになってきた。そのような中で、どのような社会にしていくべきかという社会設計としての社会実験へと、研究の流れがシフトしてきていると感じている。

経済学研究は、伝統的には、金銭的なインセンティブに基づく制度設計がなされてきた。これは、演繹的なモデルに基づいて制度設計を考えてきたものである。一方、2010年以降、先ほど少し潮目が変わってきたと述べたように、金銭的インセンティブ以外の要因に基づく制度設計も考えた方が良いのではないかという流れが出てきた。ナッジや行動インサイトといった心理的な要因など、金銭的インセンティブ以外の要因も考慮されるようになり、社会実験も多く実施されるようになってきている。

**統合人間科学としての社会科学**

このような学問の流れがある中、統合人間科学としての社会科学が重要と考えている。現在、数千や数億人単位の社会実験が多く実施されているが、その中には無駄が多く非倫理的であるとの指摘をされるものがある。また経済学の研究では、原因と結果をうまく切り分けようとする研究が多いが、これらの研究ではテクノロジーや制度が人間行動に相互に影響を与え合い、また人間行動がテクノロジーや制度にも影響するような、マイクロマクロのフィードバックの視点が多少欠けている向きがある。そのような複雑な振る舞いをする安定しない系は、経済学では比較的嫌われる傾向があるが、避けて通れないフェーズにきている。さらには、社会実験の中には、包括的な視点が欠けていると見られるものがある。社会心理学や認知科学といった分野の知見を包括的に見るような統合的な視点が必要である。最後に、ビッグデータ時代だからこそ、単に古典的な知見の検証ということではなく、新しい仕組みに対してどのように我々の心が変化し、また新しい社会としてどのようなものが出てくるかという視点が必要であると考えている。

我々のチームが念頭に置いているのは、社会科学における人間モデルの再構築である。社会の制度は、「社会におけるゲームのルール」、あるいは「人々によって考案された制約であり、人々の相互作用を形作るもの」である。外生的に制度が与えられてその中で人がどう行動するかということではなく、我々の心や行動が生み出す制度や規範を考えたいと思っている。また、生活時間、歴史・文化時間、進化時間など、さまざまなレベルでの適応合理性を念頭に置いて、人間モデルと制度の在り方を考える必要性があると考えている。

**「信頼」の要素**

「信頼」という言葉は、分野ごとの語彙やニュアンスのわずかな違いのために、包括的かつ俯瞰的な見方が困難になっている。ここではゲーム理論、とりわけ経済学や社会心理学、行動生態学の分野で扱われる語を

7 明治学院大学経済学部 准教授
https://inukailab.com/

用いて考えたい。

信頼には2つ要素がある。これは社会心理学者の山岸が提案したもの[1]であるが、一つは相手の能力に対する期待としての信頼、もう一つは相手の意図に対する期待としての信頼である（図2-3-1）。

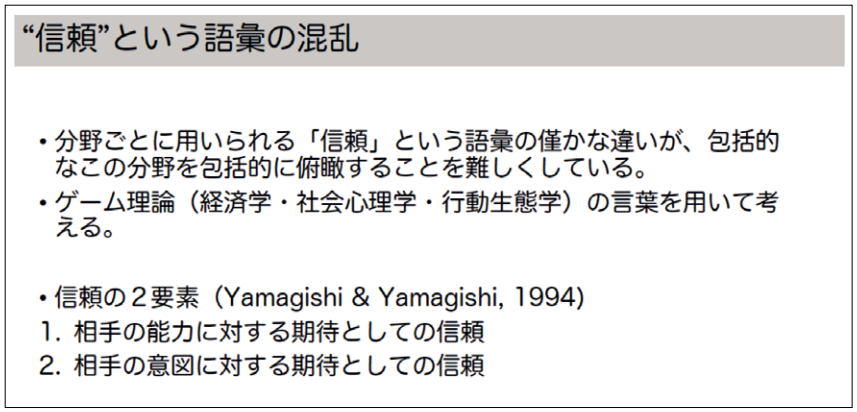


**図2-3-1　　“信頼”という語彙の混乱**

期待としての信頼については、George Akerlof（ジョージ・アカロフ）というノーベル経済学賞を取った経済学者による1970年の研究で、「レモン・マーケット」という中古車市場に関する研究[2]を紹介する。これは、中古車市場に出回る車にポンコツ車（俗語で「レモン」と呼ばれる）が多いことの理由を考えたものである。

中古車市場には、売り手と買い手がそれぞれいるが、所有者である売り手は車の状態をよく知っている一方、買い手は不完全な情報しか持っていないという構図になっている。その結果、買い手は車の状態をよく知らないため安値を付けることになり、売り手と買い手との間に売り買いのギャップが生じる。そうすると、正直な売り手は、車の状態を詳しく伝えても安値でしか買ってもらえないため市場に参画しにくくなり、結果市場にはポンコツ車だらけになるということである。これは逆選択と呼ばれる現象で、経済学では情報の非対称性とも呼ばれる現象の一つである。この問題をどう解くかの鍵として「信頼」がある。売り手の言うことを買い手が信頼できれば、良い取引ができ市場も機能するということである。

実験研究の分野では、この期待としての信頼について、信頼ゲームを用いた研究が多く行われている[3]。取引相手に何らかの形で期待を託すと、その分お返ししてもらえるという傾向が見られるというのが、信頼ゲームの一般的な研究結果である。

**信頼と安心**

「信頼」の定義として、山岸による定義[4]をもう少し詳しく紹介したい（図2-3-2）。「信頼は、相手の行動によって自分の「身」が危険にさらされる状態で、相手がそのような行動をとらないだろうと期待すること」である。要は、社会的な状況でリスクを取れるかどうかということであり、相手が「いい人」だと思うからとか、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、相手に預託してあげようという人間関係のRisk-takingを「信頼」と呼んでいる。それに対し「安心」は経済学ではアシュアランスと呼ばれる。自分を裏切ると相手自身が損をするから、裏切られるリスクを避けて罰ベースの制度を設計するという考え方である。相手が罰せられるから、相手に対して安心してお金を預けられる、といった行為がこれに該当する。

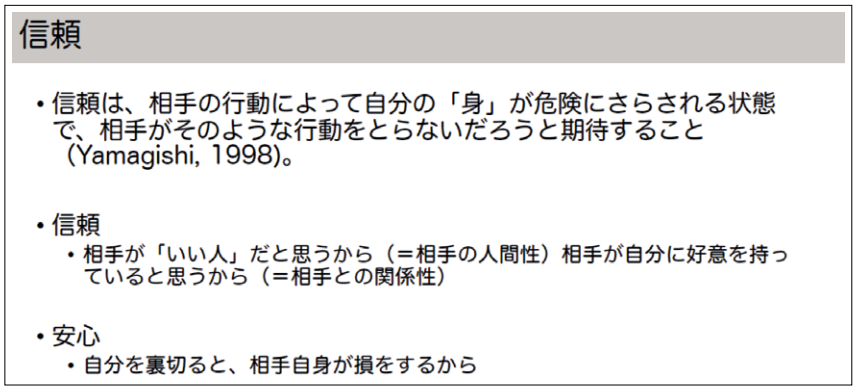


**図2-3-2　　信頼と安心**



ここで、経済学者であるAvner Greif（アブナー・グライフ）による、マグレブ商人とジェノバ商人という中世の商人の交易に関する研究を紹介する。

マグレブ商人は、一度裏切った相手とは二度と取引をしない村八分型経営で、血縁や縁故を非常に重視する商取引を行っていた。これは安心型のアシュアランスベースで、もう二度と取引しないよという罰ベースの商取引である。それに対してジェノバ商人は、積極的に新規の相手を信頼して取引をした。最終的には、ジェノバ型の商人が地中海貿易の覇権を握ることになる。

もちろん、さまざまな社会環境や自然環境なども影響するので、どちらの社会の作り方が良いと言い切ることは難しいが、新しい技術へのトラストを考えるときには、ジェノバ型の考え方が非常に重要になってくるだろう。

**社会規範や慣習の生成とガバナンスの問題、公共財**

経済学では、社会規範や慣習の生成とガバナンスは、表裏一体の問題である。何らかの制度が外生的に与えられたものであっても、内生的に出てくるものであっても、制度の存在は、個人の選好に影響を与え、さらにその好みに影響を与えられた個々人の選好がまた制度を変えるというダイナミクスがある。

社会科学の重要な問題の一つに、Thomas Hobbes（トマス・ホッブズ）の「リヴァイアサン」がある。ホッブズによると、大衆は放っておけば勝手気ままにやって無秩序状態になってしまうため、秩序を求めるように王様のような者、つまり中央集権的な制度が必要である、という。これは社会科学における重要な問いであるが、これは今日における公共財の問題と関係が深い。

公共財には、非競合性と非排除性という2つの特徴がある。非競合性は、同じ財やサービスを複数の消費者が同時に消費できることであり、非排除性は、対価を支払わず財を消費しようとする行為を実際に排除不能な性質である。純粋に非排除性のみある公共財としては、資源や共有地などがある。それに対してこの2つの特徴を併せ持つ純粋な公共財としては、NHKなどのような公共放送や、知識、イノベーションなどが挙げられる。

公共財は、社会心理学などの分野でも研究されていて、そこでは社会的ジレンマという言葉が使われる。社会的ジレンマには3つの定義がある。1つ目は、各個人が「協力」か「非協力」のうちどちらかを選択できること、2つ目は、自分の利益だけを考えれば、「協力」よりも「非協力」を選択する方が望ましい結果が得られること、3つ目は、全員が個人的に有利な「非協力」を選択した結果は、全員が「協力」を選択した場合の結果よりも悪くなること、である。例えば、環境問題はあるが自分だけ抜け駆けして守らなくてもいいよねとか、NHKの料金は自分だけ払わなくてもいいよね、といった状況である。 N=2のケースは囚人のジレンマと呼ばれるケースであり、Nが2より大きいケースは社会的ジレンマと呼ばれるものになる。社会的ジレンマでは、「非協力」によるフリーライダーの存在をどう防ぐかが鍵になっている。

この社会的ジレンマに対する伝統的な経済学の考え方として、誘引両立性がある。放っておいたらフリーライダーの増加を防ぐことはできないため、自分の利益を追求することが他人の利益となるような制度を作るべきであるという考え方である。これに対し、2000年代前半頃から、協力行動は「利他的な罰」によって説明できるという考えが出てきた。2002年のNatureの論文[5]では、我々には非常に強い互恵性（直接的な見返りがなくても他者の行為に対して何らかの形で報いる行為）があり、社会規範を逸した人に対して自らのコストを払ってまで強く罰しに行くという心の仕組みを持っていることが実験によって示され、それによって社会の規範が維持されているとされた。

**自発的な罰の効果と協力、相互監視制裁システムの在り方**

私自身が最近考えているのは、この罰ベースのシステムは、罰が人々の利他性そのものに影響を与えずに独立に機能する限りは有効だが、政府などの中央集権的でない自発的な罰は、人々の内発的な利他性や信頼性、協力率などを引き下げる（クラウディングアウト効果[8]）のではないかという点である。

これに関して、明治学院大学で公共財に関する実験を行った。1つ目が公共財に投資するかどうかのステージ、2つ目がメンバーのポイントを差し引くことができる罰ステージという、2つのステージに分かれたゲームである。この実験の結果、罰があると協力率が上がるという、罰の効果を確認することができた。一方、罰がある状況を先に経験させて、その後に罰がないフェーズに入ると、協力率が下がるという結果が得られた。

当初存在していた罰が突然なくなることで、内発的に持っていた利他性や信頼性が引き下がることが確認されたことは、非常に興味深いポイントであると考えている。昨今の新型コロナウイルス感染症においても、うまく協力しない人を罰するという行為によって、我々のトラストや協力の考え方が大きく変わる可能性があり、より具体的に調べていきたいと考えている。

もう一つ今注目しているのは、中国やさまざまな国で行われている相互監視制裁システムである。中国では、ゴミ拾いなどの善い行いによって信頼スコアが上がる。そのために人々は信頼スコアが上がるような行動を取っている。しかしこれは内発的な信頼や善行ではなく、ゲーミフィケーションされた信頼である。システムとしてうまく機能していればよいが、ひとたびそれが崩れた際に、人々のトラストの考え方が大きく変わる可能性があるだろう。

**Next Decadeの社会科学に向けて**

2020年代以降の社会科学は、人々の価値や信念、モラルなどの科学的な研究を本格的に行うフェーズに入っている。例えば、経済学がこれまで重視してきた目的合理性に加え、その目的自体の合理性や妥当性をも考える必要があるだろう。また、データ駆動型社会と功利主義的社会をどう両立させていくべきかや、機械学習のアウトプットを我々はどう信頼するか、などの観点も重要になるだろう。

では今後の社会科学研究はどうあるべきだろうか。確かに社会科学の分野内に閉じてなすべきことも多く存在するが、人間行動の基礎的理解に寄与するような研究データの収集が重要であることは言うまでもないだろう。バイオロジカルな研究であるin–vitro、実験室の実験であるin–vivo、フィールド研究であるin–situ、モデル化してシミュレーションを行うin–silicoがある。これらがうまく融合するフィードバックループを作り、社会制度設計に対する何らかの施策に生かしていきたい（図2–3–3）。

8 クラウディングアウト（Crowding Out）効果は、経済分野では例えば、政府が資金需要のために国債の増発や減税などを行ったとき、実質利子率の上昇を招いて投資の減少が起こり、結果的に民間の資金調達が圧迫されてしまうような現象（押し出してしまう現象）を意味する。


**図2-3-3　　有機的な社会設計と文理融合研究に向けて**



**【主な質疑応答】**

Q：弁護士や医者など、公的な資格を与えることによって、我々はその公的資格を安心して受け入れる。その信頼の構図は、それなりの勉強をした資格であるという側面と、何かミスをすれば資格が剥奪されるというネガティブな側面があるのではないか。

A：その通り。情報の経済学といわれる分野で多く研究されている。

Q：ジェノバ商人は、失敗しても致命傷にならないだけの経済的な豊かさがあったから、新しい取引先を積極的に開拓できたのではないだろうか。資本力のないところでは、そもそも相手を信頼して出ていくことはしないように思う。そのような観点での研究はされているか。

A：多く研究がされている。ジェノバ商人が金持ちであったことは確かである。もし裏切られたとしても、最低限のバッファーが保障されていることは非常に重要であり、その中でどの程度リスクを取ることができるかについて、多くの研究がなされている。例えば、不況期に産まれた子供は、リスクを取らない傾向があるようだ。社会的な環境と最低保障の問題含め、トラストと関係してくる点だと思う。

Q：クラウディングアウトの実験では、最初に罰を与えることで、その罰がなくなった途端に、協力率が下がるということであった。これは、罰がなくなることで、自分だけ損をしたくないから自分もやりたくない、という人間の心理的な動きではないだろうか。

A：その通りである。罰せられる人がいるから私は協力するよ、という協力のフィードバックのようなことが起きている。しかし罰がなくなると、その信念のようなものが一気に崩れる。脳科学と合わせて見ていく必要があると思っている。

Q：機械学習は非常に複雑であるため、中身を全部理解した上で行動することはほぼ不可能である。従って、ある時点で思考停止してトラストしないと、得られるリスクやゲインに対して自分の努力する資源のバランスが悪くなる。リスクがものすごく大きければ徹底的に調べるが、リスクが小さければトラストした方が経済的合理性があるだろう。そのあたりのリスク評価が重要になるのではないだろうか。

A：その通りで、先程の安心やジェノバ商人の話とも関係する部分だろう。例えば、1980年代、持ち株会社同士で互いに商取引するという日本型の経営に人気があった。どのような社会制度を設計するかは、そのときのさまざまな要因によってケース・バイ・ケースになるだろう。

Q：リスクを避けることと「責任」は関係しているか。

A：「責任」は非常に重要なポイントである。関係性が続く中で、自分に矛先が向けられるかもしれないとなると、過度に責任を逃れるような行動になるだろうし、また責任を追及する行動が、相手の次の行動へと大きく影響を与える要素になるだろう。

C：日本人は、利己的から利他的に変わってきたのではないだろうか。その変換点は東日本大震災や阪神淡路大震災のような災害であると感じる。

Q：社会シミュレーション研究で、今後強化が必要な点はどこか。

A：シミュレーションは、心理学や行動経済学などの知見をもとに、モデルに入れるパラメーターや制約条件を決めるというトップダウンのアプローチであるが、一方で、ビッグデータからボトムアップにモデルを作るというアプローチもある。この2つをどうすり合わせるかが非常に重要であり、課題である。例えば、信念や責任など、心理的な要素をモデルの中でどう捉えるかは、しっかりとした視座を持つべきである。

Q：in-vivoとin-situの研究者がそれぞれ分かれていて、接点があまりないことが背景にあるか。

A：そう思う。それぞれ独立で研究してしまっている。また社会科学自体も分野固有の問題に取り組む傾向がある。もう少し双方のつながりが必要と考えている。

Q：利他性に関わる脳の部位や、不公平感に反応するドーパミンの仕組みなど、脳科学の研究が進んでいると聞いたことがあるが、そのような脳科学とのつながりは出てきているか。

A：利他的に振る舞う行動が、自分にとっての報酬になっている可能性があることは、脳科学の研究で少しずつ明らかになってきている。猿などの動物を使った研究や、そのような動物と人間がどう違うかといった研究もされてきている。

C：有機的に連携するプロジェクトは非常に重要だが、なかなか難しい。制度設計では、エビデンス・ベースド・ポリシー・メーキングという言葉がキーワードになっていて、何らかの実証がなされていないと制度設計に至らないと感じている。モデルシミュレーション化だけでなく、実際のフィールドデータでの実証に結びつくようになると良い。

C：経済学はモデル化が初めにあって、それに合わせる傾向があった。今は経済学も逆の流れになってきて、フィールドのデータを集めてそれをモデル化しようとする動きになってきている。ただやはりプラットフォームは必要である。フィールドのデータを持っている人、モデル化の能力がある人、その先にin-vivo/in-vitroの研究ができる人が集まって議論できるような場所は必要である。そのときに、「トラスト」が一つのキーワードになるだろう。

C：慣習がトラストを作っているように感じている。罰の話があったが、最初は罰だったとしても、それが制度化して慣習化すると、世の中の人はそれをトラストするのではないだろうか。例えば、ハンコにはセキュリティーはないが、人々の慣習がトラストを作っている。その慣習を変えることが難しいということが、デジタル社会の阻害になっている。

## 2.4 大屋 雄裕[9]「法制度とトラスト」

**トラストをどう訳すか？**

2.1節でもトラストをどう訳すか、「信頼」なのか「信用」なのかといった議論が出たが、我々からすると重要なのは「信託」である。「信託」とは、①自己の財産を、②第三者に委託し、③自己または他者のために運用・管理させる制度、と説明される。

例えば不動産を、もともとの持ち主である委託者Aさんから受託者Xさんに移転することを考える（図2–4–1の右側）。ただし、その移転の条件は、不動産から得られた利益を、第三者であるBさんのために活用するものとする。つまりXさんはAさんから受け取った不動産を運用し、その利益を受益者たるBさんに与えねばならない。これが信託という制度である。我が国においては、おおむね信託銀行によって担われている業務である。



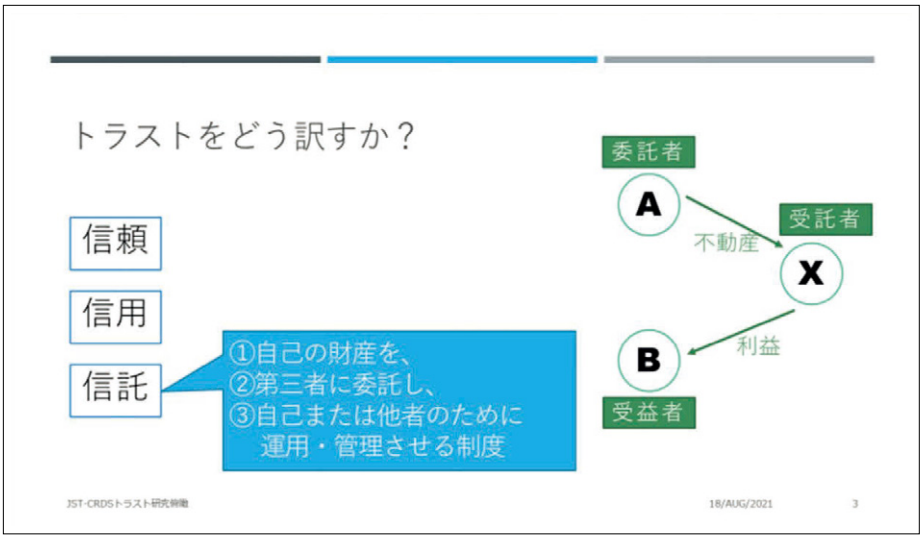


**図2–4–1　　トラストをどう訳すか？**

このようなトラストや信託というものが、どこから生まれてきたかについて話しながら、中身について検討していきたい。

**法体系の分類と、信託との関係**

まずは法律学のバックグラウンドの初歩的なところから簡単に説明する。世界の各国が持っている法のうち、いわゆる先進国の法体系というのは大きく大陸法と英米法の2つに分類される。

大陸法というのは、もともとは古代ローマの法に由来し、それが中世イタリアで再発見され、フランスやドイツといった大陸ヨーロッパ諸国に継受されていった。日本も特にドイツから法律の多くを学んだため、一般的には大陸法諸国に分類されている。

9　慶應義塾大学法学部法律学科教授
https://www.k-ris.keio.ac.jp/html/100000819_ja.html

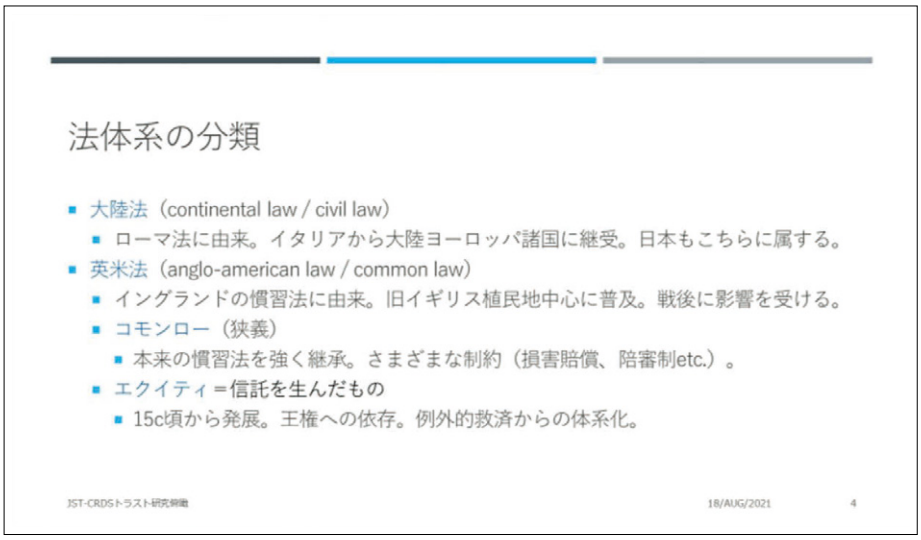


**図 2–4–2　　法体系の分類**

これに対し、大陸から海を渡った向こう側の島国であるイングランド[10]においては、地場の慣習が残り続けた。その後の歴史的な経緯もあり、イングランドにおいては慣習法をもとにした判例法体系というものが成長した。法の世界から言うと、もともと田舎の少数者の慣習法だったが、さまざまな歴史の偶然により外交政治的にはこちらがすごく強くなってしまった。このイングランドの慣習法が旧イギリス植民地を中心に普及することになる。

さらに、イギリス植民地から独立したアメリカもこの英米法を継受したので、政治経済的には巨大パワーがこの英米法によって統治されている状態になっている。日本も第二次世界大戦後にアメリカに占領統治された経緯があり、その際に行われた法的改革のかなりの部分が英米法に基づいているため、ある種の混交状態にあるということになる。

ところで、この英米法のことを英語ではアングロアメリカンロー、それから、コモンローと言ったりする。コモンローとは要するに、地場の慣習だからみんなが共有しているもの、というようなイメージである。しかしコモンロー体系をよくよく見ると、本来の慣習法を強く継承した狭義のコモンローと、エクイティー（衡平法）という15世紀頃から発展した別の法体系から成っていると説明される。なぜそんなことになったかというと、コモンローはもともとの慣習であるため、さまざまな制約が組み込まれていたことによる。

具体的に言うと、コモンローにおいては、救済手段や訴訟で勝ったときに相手から取れるものが金銭賠償に限定されていた。従って、「先祖伝来のよろい」を奪われたと訴え出ても、よろいの価値分のお金しかもらえないみたいな制約があり、それだけでは救済にならないといった問題があった。

またコモンローでは陪審員制が保障されていたので、加害者が有力者の血縁者だったような場合に、周りの市民がグルになって守ろうとして無罪判決が出てしまう、といった問題もあった。

そこで、こういったコモンロー上の制約を打ち破るため、古くは12世紀ぐらいから国王権力に依存して例外的な救済を認めてもらうという風習が始まった。それが徐々に体系化されて新たな法体系へと結実するが、

**10**　現在「イギリス」と呼ばれる国家（グレートブリテン及び北部アイルランド連合王国）は、歴史的にはイングランド・ウェールズ・スコットランド・アイルランドという異なる国・地域として発展してきた。英米法、特にコモンローはこのうちイングランドの慣習を基礎として成立しており、イングランドへの統合が早かったウェールズ（13世紀末）も基本的に同一の法体系に従っていたが、1694年まで独立していたアイルランド、1707年に法的に統合されたスコットランドにおいてはこれと異なる法体系が発展していた。このため、イングランド＋ウェールズで発展した法体系のことを「イングランド法」、スコットランド・アイルランドを統合して発展した国家のことを「イギリス」と表記している。

その非常に大きな契機となったエクイティー法体系の中心を成しているとされるのが、実は「信託」である。

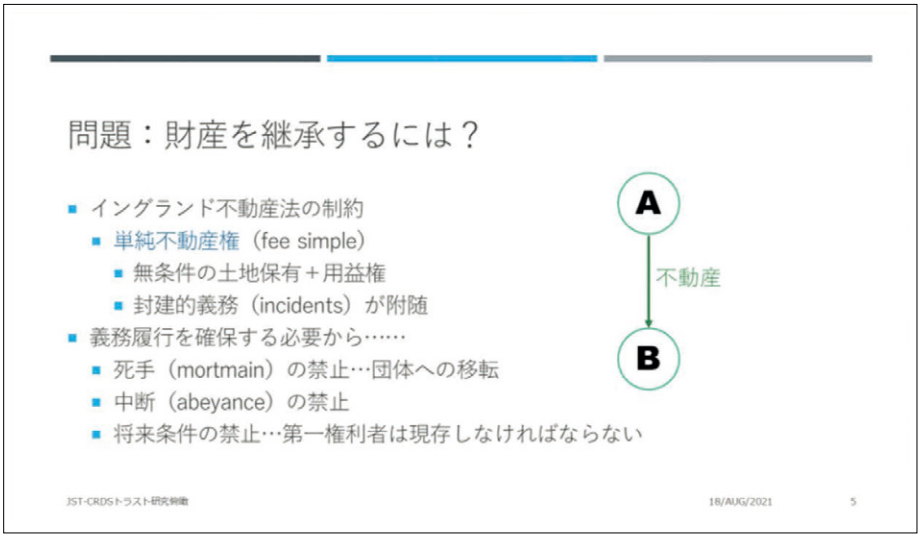


**図 2–4–3　　財産を継承するためには？**



**信託が必要とされる背景**

信託が必要とされる背景は、12–13 世紀の封建時代におけるイングランド不動産法に由来する。封建時代においては、国王が全ての土地の権利を持っていて、それを家臣に分与していくという発想がある。その分与のときに、単純不動産権（Fee Simple）と呼ばれるものが与えられた権利の典型と想定されている。

単純不動産権とは、無条件に与えられた土地を保有し、それを使用して収益ができること（用益権）。さらにそれを領民に耕させて年貢を取ることができること、といった権利であると観念されている。その一方で、国王から土地支配を許容された代償として、封建的義務がくっついている。具体的には、軍事奉仕の義務、結婚する際には結婚を認めてもらったお礼を払う、といったものである。単純不動産権は、このような負担つきの権利と観念されており、日本でいうと御恩と奉公に当たるようなものと考えられる。

負担つき権利であることから、イングランド不動産法には複数の禁止規定が存在し、不動産を勝手に処分することができなかった。

死手の禁止とは、例えば教会・大学・法人といった団体に不動産を譲ってはいけないことを指す。理由は、団体は軍事奉仕できないからである。中断の禁止とは、例えば子供に不動産を譲りたいが、軍事奉仕できないからその間中断、といったことを認めないということである。将来条件の禁止とは、将来生まれてくる子供のように、譲渡する時点で現存しない相手には譲渡できないということである。理由は、残念ながら生まれなかったときなどに所有者不在となり、中断が発生するためである。

以上の条件によりコモンローにおいては、例えば不動産の持ち主Ａさんが明日をも知れぬ命になり、自分の子供Ｂさんに譲渡したいが未成年あるいは軍事奉仕を今行えない状態にある場合、ＡさんからＢさんへ直接譲渡することができない。

そこでユース（Use）というものを設定することで解決しようとした。つまり第三者であり成人していて権利能力があるＸさん（＝典型的にはＡさんの信頼できる友達あるいは親族）に、その不動産を譲渡するという方法である。その不動産から得られた利益を自らの子供Ｂさんのために使ってほしい、あるいは将来Ｂさんが成人したらＢさんに再度譲渡してほしい、返してくれということを込みでＸさんに不動産を譲渡する。

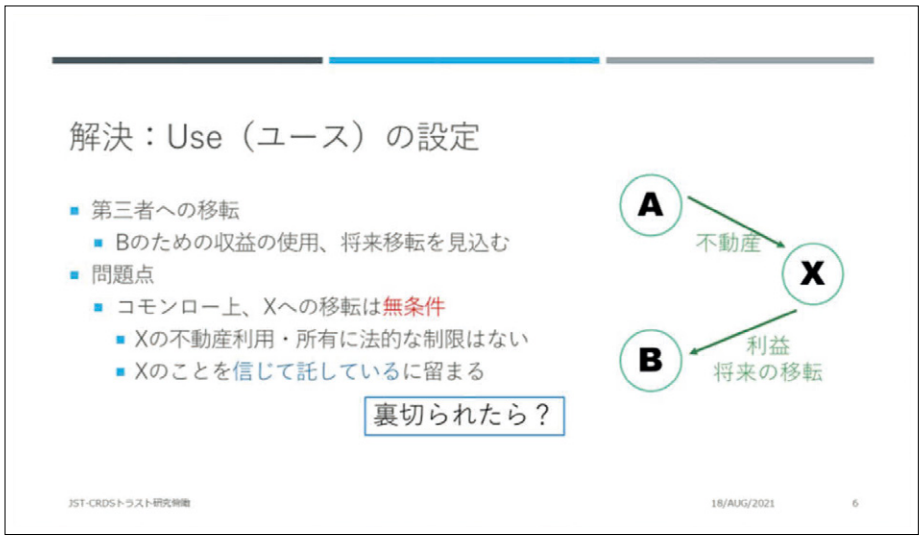


**図 2-4-4　　ユースの設定**

その際に、Ｘさんが無条件の使用収益ができる権利を手にしないと、単純不動産権は成立しない。すなわちＡさんがＸさんに不動産を譲るときは、法律上は無条件でなくてはならないので、Ｘさんのことを信じて、Ａさんは自らの財産を託す。それが信託、ユースが出来上がった状況である。

しかし問題は、信じて託したにもかかわらず、Ｘさんが裏切った場合の処置である。Ｘさんは無条件で100%の単純不動産権を所有しており、ＡさんもＢさんも手が出せない。そこで発生したのがエクイティーである。

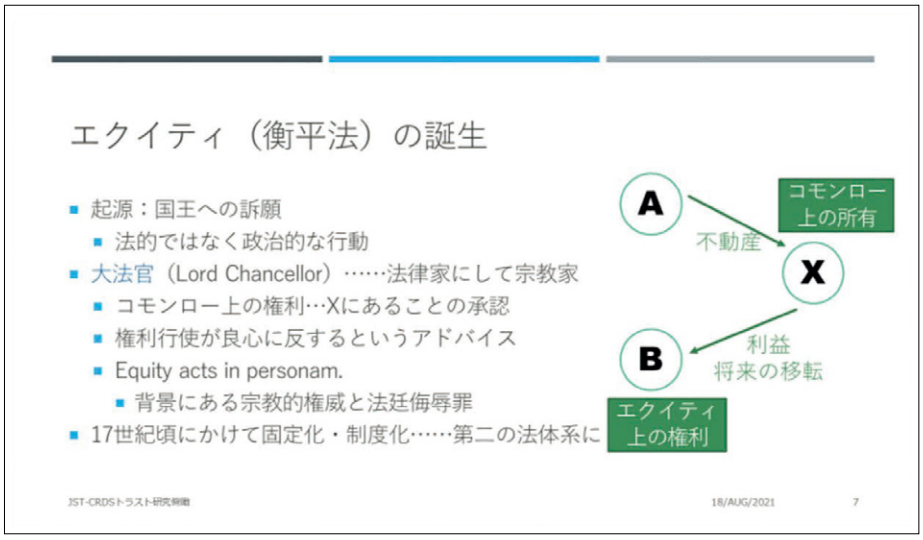


**図 2-4-5　　エクイティー（衡平法）の誕生**

**エクイティー（衡平法）と信認関係**

エクイティーは、もともとは法ではない。法の外側で、何とか国王の権力を用いて「私をお救いいただけませんか」といって訴願に出ると、それは国王から大法官に任される。法律家にして宗教家でもある大法官は、法律家としてはＸさんが持つコモンロー上の権利を認めざるを得ないが、宗教家として「あなたの良心にかけ

てどうするか」をXさんに問いかける。もしこの時代に偉い宗教家の言ったことを拒むと、宗教的権威を汚したことになり、例えば教会から破門されて死後の救いが得られないことになってしまう。さらに大法官は法廷を抱える法律家でもあるので、そのアドバイスに従わないことは法廷侮辱罪（刑事罰を含む）を構成し、監獄に入れられて死に至る可能性もある。従ってXさんは、持っているはずの権利を使わないことに同意せざるを得なくなっていく。

このような「国王に訴願したら例外的救済が受けられた」といった事例が世の中に広まり、固定化・制度化して17世紀頃にかけて成立したのが、第二の法体系としてのエクイティーである。エクイティーはルールとしてではなく、「あなたの人間としての生き方の問題ですよ」といった普遍的な条件として働きかけるため、Equity acts in personam.（エクイティーは個人的に働く）という法のことわざもある。

つまり英米法においては、コモンロー上認められる所有権としてXさんのところにある権利と、エクイティー上の権利（使用収益する権利、得られた利益を受け取る権利）が分裂したものとして観念されるようになった。このようにして生み出された制度から、さらに抽象化された関係としての信認（Fiduciary）というものが、英米法では観念されていくようになる。



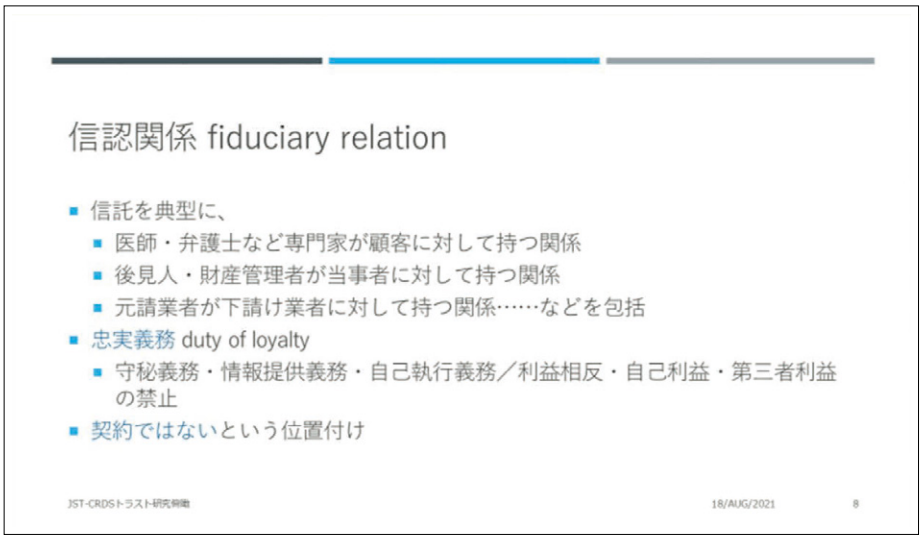


**図2-4-6　信認関係 fiduciary relation**

信託とは、死んでいく人から信じて託されたものであり、それをどのように扱う義務があるか、どのように扱うことが良心に反さないかといったことを述べた。これを典型として、医師や弁護士などの専門家が顧客に対して持つ関係、後見人や財産管理者が当事者に対して持つ関係、元請業者が下請業者に対して持つ関係、のように、第三者のために財産などを保管し、あるいは運用する関係を「信認」と位置付け、信じられた側に忠実義務が課されていった。

忠実義務を具体的に言うと、守秘義務、情報提供義務、自己執行義務（＝自分でやらなくてはならず、第三者に丸投げしてはいけない）。あるいは、利益相反する場合に受けてはいけない、自らの利益をむさぼるためにその信認関係を利用してはいけない、信託された受益者以外のためになるような第三者利益を発生させてはいけない（第三者利益の禁止）、等々である。

注意すべきこととして、「第三者の利益を図ることで受益者を傷つけてはいけない」は利益相反の禁止であり、上で述べた「第三者利益の禁止」とは異なる。「第三者利益の禁止」は、たとえ受益者に対して一切関係ない、受益者の受け取るお金が目減りするようなことが一切なかったとしても、受益者以外の第三者に利益が発生するようなことをしてはいけないという原理である。

一例を挙げると、弁護士がクライアントから依頼を受ける際、訴訟費用などのために受け取る仮払金は、英米法において信託のような関係になる。これは弁護士本人の財産とは完全に分離して管理しなければならないので、弁護士は利子のつかない銀行口座に入れていたが、銀行はある種の公益のためであれば利子のつく口座を作れる、と言い出した。利子を公益のために使ってもクライアントの不利益にはならないとの考えであったが、アメリカにおいては第三者利益の禁止に抵触するということで大きな問題になった。たとえ本人や受益者に不利益がなくても、それ以外の人の利益を勝手に考えてはいけないという点まで、忠実義務で配慮された例である。

ここで興味深いのは、信託とか信認は契約ではない点である。契約というのは、双方当事者が対等な関係であるのに対して、信認というのはそうではない。言い換えると、契約はコモンローの世界であるが、信託とか信認はエクイティーの世界という形で区別されている。

**中世・近代からの社会変化と、信認関係の現代的意義**

トラストをウィキペディアのドイツ語版、フランス語版で各々引いてみると、どちらも「トラストというのはコモンロー上の制度で、こういう受託者のために第三者の下に財産を移転する契約」のように説明されており、大陸法には存在しない考え方であることが示されている。

実は日本にはトラストの考え方が存在しており、1905年に担保付社債信託法ができ、1922年からは一般法としての信託法が制定されている。これは日露戦争において、日本は莫大な戦費調達のためにロンドンで社債を発行したことによる。ロンドンで発行する債権の担保を、信託制度を用いて設定しないと引受手が納得しなかったためである。その結果、本来大陸法国である日本に例外的に信託法理が入り、その後信託銀行を中心に活用されてきた。

ちなみにフランスに信託法が入ったのは2007年ぐらい、ドイツはいまだに一般法がないとのことである。ではなぜそのようなことになったのか？

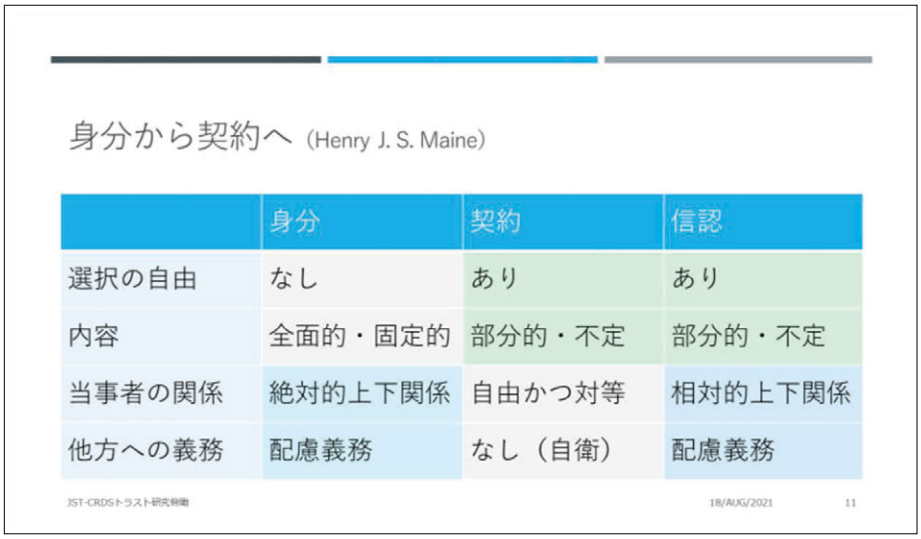


**図2-4-7　　身分から契約へ**

イギリスの法制史家であるHenry Maine（ヘンリー・メイン）が唱えたテーゼにおいては、中世は身分の社会、近代は契約の社会と説明されている。

身分とは何かというと、要するに生まれに伴って選択の余地がなく、全面的・固定的な上下関係が発生するものである。その上下関係においては、もちろん下が上に対して納税・軍事奉仕などさまざまな義務を果た

す必要があるが、逆に上も下に対して結婚相手の紹介や、残された子供の面倒を見るといった配慮を行う義務がある。相手に対して配慮をする義務というのが、必然的に組み込まれているのが身分法である。

ところが、近代に移行したときに契約の社会に変わった。契約とは何かというと、自由かつ自発的に対等な当事者同士が、特定の目的のために勝手に結ぶものである。一般論として言うと、契約というのは自己利益の最大化のために結べばよいので、その結果、相手がどうなるかをあまり考える必要はない。相手は相手なりに考えて、これが良いと思って合意したから放っておけばよいというのが契約法の世界である。

先ほど述べた信認関係というのは、身分法と契約法あるいは中世と近代のハイブリッド的／中間的なモデルであると整理される。つまり信認関係においては、誰を信じて託すかを選択する自由があり、何をどう託すかも決められる。しかし、託したら上下関係に入り、ただし、上の者は下の者に対して配慮する必要がある。

ここで法において信頼がどのように確保されてきたかについて、歴史を振り返って考えると、古来においては当事者の相互承認であった。小さな社会であれば、お互いのことを見て情報の蓄積が働くということである。

古代ローマ法の世界において、貿易などを安全に行うためにどうするかという問題が生じたが、彼らはお互いに信頼する、信頼のネットワークを作ることによる解決を選んだ。取引をして、その取引の条件を裏切らなかったという情報を共同体でシェアすることによって、その範囲内で信用取引をすることを許すという関係であり、裏切ったらその共同体からたたき出される。当時、例えば地中海貿易に関与できる商人の数はごく少ないため、こういう共同体の信頼のネットワークで対応することができた。

しかし、近代に入り社会が大きくなると、顔見知りの共同体ではなく、国民国家と呼ばれる一定の領域を持ち、巨大な人口を抱えた組織を作らなければいけなくなった。そこで例えばパスポートのように、国家の手によって信頼すべき人物に保証を与える制度が導入された。

また国家は、信頼すべからざる人物を特定し排除する、といった権力も行使することで、社会の中に普遍的信頼をもたらした。つまり札付きではなく、その辺を歩いている人間は全部信頼してよいという感覚を、市民にもたらしたのが近代社会である。

イングランドにおいては、そのような近代社会への転換というのがあまり起きなかった。フランスやドイツは、何らかの革命を通じた近代社会への転換点というものが結構あり、法制度の切替えを行っている。これに対しイングランドは、少なくとも法的には大体1225年ぐらいから継続性がある。中世封建制の時代から延々と積み重ねて近代へ緩やかにシフトし、先ほどの契約的な社会への全面転換が起きずに、中間形態としての信認法理みたいなものを抱え込んだまま現代社会に来たと言える。

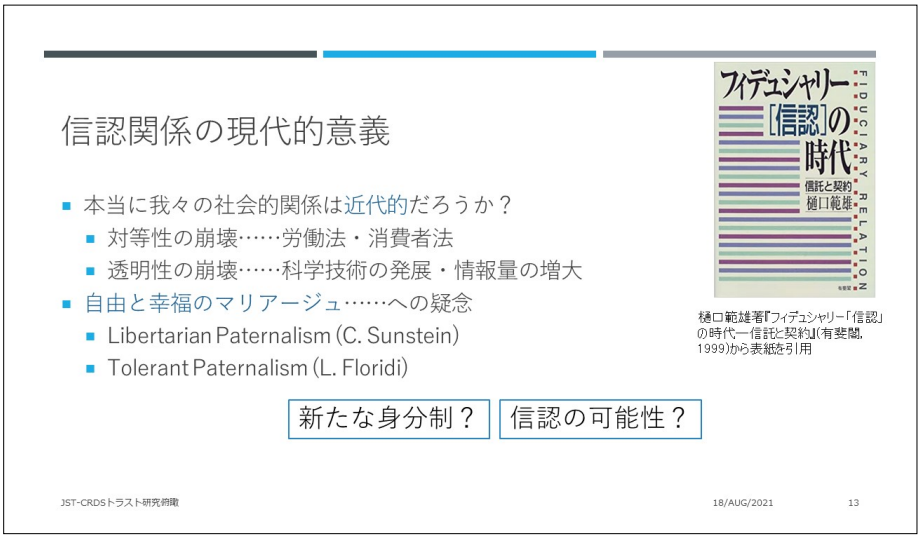


**図2-4-8　　信認関係の現代的意義**

ところが、実はそれによって現代的には大きな意義を持つことになったと言える。

つまり自由かつ対等な市民が、自己利益を実現するために契約を結ぶ社会が近代であると述べたが、現代はそうなってはいない。典型的には労働法や消費者法において、全ての市民が対等であるという前提を、もはや放棄せざるを得なくなっている。また本セミナーで重要な点として、科学技術の発展や情報量の増大によって、我々はもう社会の状況が分からなくなっている。

自由と幸福のマリアージュというのは私の表現であるが、人々に自由を与えておけば自分に配慮して自己利益を最大化し、全ての人間が最大限に幸福になり、従って社会全体の幸福も最大化されるという信念はもはや失われている。

その一方で、一定の義務や責任を課すための法理としての信認というものの重要性が見えてくる、と考えることもできる。つまり信認というのはヨーロッパ的な見方からすれば前近代的なもの、あるいは中世から近代への移行段階に例外的に発生したものだと片づけられるかもしれない。しかし、実は現代においてこそハイブリッドモデルとしての信認、信じて託すけれども、誰を信じるかは私が選び、託されたものをどうするかについては責任がある、という議論が生き返るのではないかと思われる。

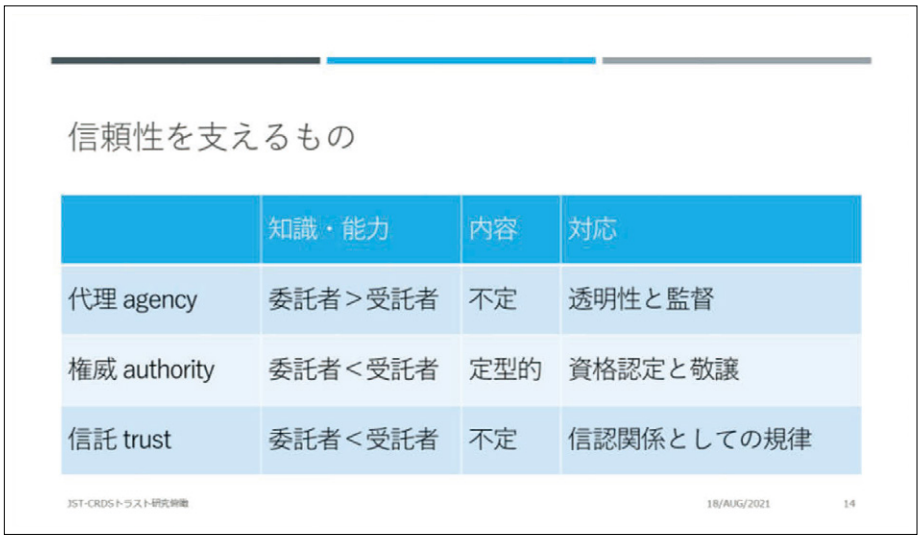


**図2-4-9　　信頼性を支えるもの**

**信頼性を支えるもの**

以上のような観点から、誰かに何かを頼む、仕事を委託するとしても、委託者と受託者の関係によって3通りぐらいの分類ができると考え、図2-4-9を昨年の人工知能学会で報告した。

特に委託者より受託者の知識や能力が高く、不定の内容を持つ場合、かつ、あらかじめ国家によるライセンスなどにより統制することができないものについては、信認関係として、いわば現代における信託として扱っていくことができるのではないかと考えている。これが法制度から見た場合のトラスト、現代の情報化社会の中でトラストの背景にある信認関係が持つ意味であろう。

**【主な質疑応答】**

Q：日本国憲法にも前文と基本的人権に「信託」と記載されているとのことだが、その背景を教えてほしい。

A：GHQの草案が元になっており、きちんと立法経緯を確認したわけではないが、例えば我々は国際社会を信じて安全を託すことになるので無権利といった考え方（第9条）に見られる。ただし、法理自体は入っているが、実態についてはいろいろな指摘がある。

Q：「第三者利益の禁止」については、今後見直す可能性はあると思うか？　また見直した場合に何が問題になると予想されるのか？

A：基本的には、受託者は受益者のことしか考えてはいけない。さらに言うと、預かった信託財産を銀行が自行口座で管理したら自己利益違反となった例もある（後に自行口座でもよいとの法律もできたので難しいが）。忠実義務とは、第三者から見て明らかにロイヤリティーが損なわれていないように行為する義務と理解されている。

本日典型例として挙げたように、親Aさんが亡くなって子供Bさんが無権利者となり、かつBさんが例えば重度精神障害者で能動的に権利能力を持たないケースもある。それでも本人のベストインタレストを実現していることが、社会的に確認できなければならないと理解されているようである。

Q：大陸法、英米法の2つがありきとのことだったが、今の世界情勢も踏まえてそこをゼロから見直す可能性は？

A：この2つがハイブリッド化する機会はあったが、結局はそこまで至らなかった。これまでの各々の継続性や安定性を配慮すると、ハイブリッド化やゼロから見直すのは難しいのではないか。

Q：プラットフォームビジネスにおいて信認関係は成立しているだろうか？　例えば市民がGAFAに個人情報を提供しているが、法律に基づく契約というよりは信認のように見える。さらに言うと、デジタル遺産を誰に託すかといった問題にも関わる。

A：プラットフォーム屋さんは分からないが、法律屋さんはクラウドコンピューティングについて考えている。人様からコンピューティングの力を預かり、情報を預かって使うことは信認モデルであり、分離管理しなければいけないと当然考える。

そして信認モデルは原則で言うと無報酬であり、預かったものから自己利益を得てはいけないはず。得られるのは、法律上もう明確に規定されているか、信託契約で明示的な報酬が書いてあるときだけ。そこがおそらく、英米の議会や取り締まり当局において問題視されているのではないか。

Q：中国においては法律云々よりも、顔認証システムによる監視とか、ジーマクレジットのような個人信用度のスコアリングによって、信認関係が作りやすくなったという見方もできそうだが、そのあたりはどうか？

A：専門家によると、中国は近代において大陸諸国が築いたような中央権力による普遍的な信頼構築に失敗し、代わりに人的なつながりを広げてネットワーク化することで信頼を作ったといわれている。もしネットワークから外れると、例えば図書館から本を借りるのにデポジットしか信用できないということになり、スコアリングが浸透した。つまりベースラインが低信頼だったという背景がある。

もし日本でも同様にスコアリングを使う可能性があるかと訊かれたら、個人的には否定的である。なぜならば、日本はもともと高信頼なのでスコアリングを使う意味がなく、むしろネガティブな指標にしかならないと思うからである。欧米も同様ではないか。

Q：人工知能学会で説明された、信頼性を支える「代理」「権威」「信託」の3つは法的な面を言っているのか、それともAI技術を使う社会で言えることなのか？

A：もともと法的にこの3タイプぐらいのシステムが抽出できるという考え方がある。日本で言うと、「代理」はそのまま代理、「権威」は医師・弁護士などの認定システム、「信託」は信託銀行のようなシステム。この話がAIの管理・監督についても参考になると思って整理した。

AIでの具体例を言うと、「代理」は例えばAIによる保育園の入所マッチング（ルールが決まっている）。「権威」は例えばレントゲン写真の読影（ただし、医師よりもAIの方が高精度という認定がある場合）。「信託」はそれ以外のリコメンデーションAIなど、人間には判断の中味がよく分からず、かつAIもお墨付きではないにもかかわらず、信じて託す場合。その場合「信認関係としての規律」に従って、受益者が不利益を訴え、後日受託者が責任を取る、といった対応がなされる。

Q：トラストは片方向の関係性だと思っていたが、信認は双方向と考えてよいか？

A：心理的な信頼は一方的のものと考えられるが、信頼が裏切られることもある。そこで信頼を担保する社会的装置としてのエクイティーがあって、初めて制度として信頼が成り立つと理解するのが良いと思う。

Q：信認関係がビジネスになる例はあるか？　忠実義務を負うのは重たいことであるし、自己利益も持てない。となると、例えば大きなインフラを持っているプラットフォーマーなどに限定されるのではないか？

A：最初に中世で発生したときは無報酬だったし、日本法でも委任は無報酬が前提だった。その後どうなったかと言うと、委任契約に書いていなければ無報酬だが、書けばもらえるようになった。「これだけもらう」を事前に明示することが、ビジネス化に必要であると思う。過去にビジネス化できなかった事例では、そこができていなかった。

Q：例えばAIが自動運転する車における、AIと乗車している人の関係は「信託」か？

A：「信託」よりは「権威」に近いと思う。例えば電車の乗客は、運転手に命を預けていることになっている。運転できるという国家資格が与えられているところに、乗客がそのサービスをクレジットして利用することの源泉があると思う。

自動運転車についても、少なくとも現在各国で想定されているのは型式認定だと思われる。このタイプの自動運転車であれば路上に出してよろしい、これは駄目というのを国が認定することになり、どちらかと言うと権威のシステムになると考えられる。

## 2.5 神里 達博[11]「科学技術へのトラスト」

専門である科学技術社会論あるいは科学史の立場から、科学技術に対する信頼、あるいは、損なわれてしまった信頼の回復に向けた、さまざまな過去の努力について紹介する。その際に、まず、科学と技術は違うものであり、それをどう考えるのかというところから始める。また、日本における信頼の問題の特質についても触れる。

**科学技術への懐疑**

科学技術に対する信頼が疑いを持たれた時代がある。そもそも、科学と技術とは別ものである。科学は知識であって、技術は人類共通のあらゆる文明の中から生まれてきたものである。科学と技術とが結びついたのは産業革命というイメージがあるが、実際に産業革命では職人が知恵と試行錯誤でいろいろなものを発明していた。水力紡績機を発明したアークライトは理髪師だった。そこにはいわゆる科学の知識は全く影響していない。

ただ、逆に、職人の作ったものを科学者が観察して、一体何でこれはこのようにうまくいくのだろうと考えることで科学が発展したケースはある。蒸気機関は熱力学の成立を促したが、基本的には別々である。

Francis Bacon（フランシス・ベーコン）は、技術のための科学を400年以上前に言っていたが、彼の頭の中だけにあったもので、現実にはその時代には科学と技術は結びついていない。結びつき始めたのは19世紀後半、本格的な結合は20世紀のアメリカである。

ナイロンを発明したデュポンのように、20世紀前半のアメリカでは企業内に研究所がたくさんできた。プラグマティズムの国アメリカで科学と技術が本格的に結びつき、実学として使われるようになった。マンハッタン計画は日本にとっては悲劇だが、アメリカにとってはすばらしい、輝かしい歴史である。核開発に科学と技術

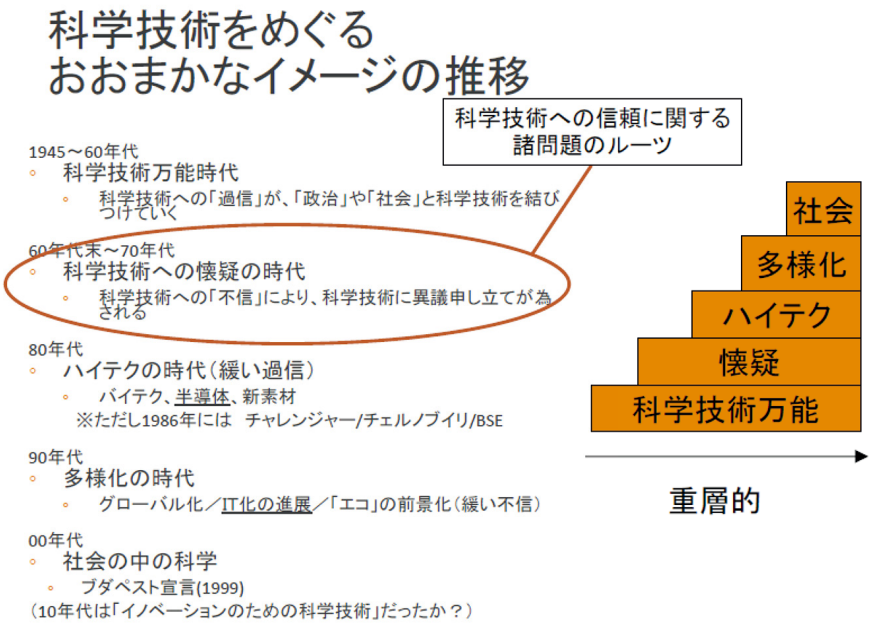


**図 2-5-1　　科学技術をめぐる大まかなイメージの推移**

11　千葉大学国際学術研究院教授（総合国際学位プログラム長）、大阪大学 客員教授、朝日新聞客員論説委員
https://researchmap.jp/kamisato

が結びついた形で動員され、アメリカは第二次大戦で一人勝ちした。要するに、科学と技術の結合によって国家全体を強大なものにしたということで、ファシズムを倒したアメリカの価値観が世界を席巻した。科学と技術とが結びついた科学技術はすばらしいという考え方が常識となった。

科学技術がありとあらゆる問題を解決していくという信頼と自信に満ちた時代、「科学万能時代」が、戦前戦後1945年からやってくる。しかも、それは政治や産業とも結びついたため、全てが科学技術の下にあるという時代が60年代半ばまで続く（図2–5–1）。

60年代から70年代にかけて「科学技術への懐疑の時代」がやってくる。本日はここに注目したい。最初に科学技術の信頼の問題が生じる時代である。科学技術に対する信頼の最初の揺らぎというのは実はこの時代にある。このときに考えられたことは実は何度も形を変えて世の中に現れてくる。この年代に懐疑の時代がやってきた背景には後ほど述べるようにベトナム戦争がある。

80年代に入るとベトナム戦争も終わり、レーガン大統領が登場し、「ハイテクの時代」になる。この頃からアメリカはプロパテント、知的財産で国力を養う。バイテク、半導体、新素材に光が当たり、科学技術に対する信頼が少し回復するというか、緩い過信の時代になる。ただ、その陰ではスペースシャトルのチャレンジャー号の墜落事故、チェルノブイリ原発事故、狂牛病BSEの発生といった、また別の形で科学技術に基づいた社会に対する不信を作るような出来事が起こる。

90年代になり冷戦が終わる頃には、さまざまな形でこれらの問題が表面化する。同時に、グローバル化とIT化が進展したのも90年代で、地球温暖化問題が世界的なアジェンダになった。70年代に出てきたエコロジーの問題が再度前景化する。90年代は「多様化の時代」と名付けたが、やや不信感が強い時代かもしれない。

00年代は新しい世紀だが、直前の99年にブダペスト宣言があった。科学というものが独立してあるのではなくて、社会の中のある重要なアクター、まさに「社会の中の科学」であるということがユネスコなどによって宣言された。

10年代をどう理解すべきかと、まだはっきりしたことは分からないが、「イノベーションのための科学技術」という時代と私は理解している。いろいろな制度の変更もあり、世界的な大競争時代の中で、科学技術があくまでイノベーションのためにあるものだという方向に進んでいった時代なのかもしれない。興味深いのは、科学技術をめぐる大まかなイメージというのは、科学技術万能の時代、懐疑の時代、ハイテク、多様化、社会の中の科学という形で、新しいイメージが出てくるが、そのたびに昔の考え方が消えてしまうのではなくて、科学技術万能というイメージもあるし、科学技術に対する懐疑が根強いという部分もある。重層的に重なることで、科学技術とは何かということが徐々に曖昧になる。歴史的にはそういうことが起こっていると私は理解している。

**懐疑の時代**

60年代後半から70年代前半にかけては最初に科学技術に対する信頼が問題になった時代である。ベトナム戦争が始まった。アメリカは月に人を送り込めるのに、貧困、犯罪、麻薬、戦争といった地上の問題は解決できていない。果たして、科学技術の発展は私たちを幸せにしてくれるのかということを一般の人たちが感じるようになった。

さまざまな反応があった。科学技術文明自体をもう終わりにさせようという、過激な、いわゆる反科学主義者も出てきた。また、かつては科学技術政策は意識されていなかったが、政治の側が、研究開発を管理、修正して、方向性を与えていくことが大事だといわれるようになり、以前のように手放しで科学技術の発展が尊重される時代は終わった。

大きな要因はベトナム戦争である。大量の枯葉剤をベトナムの農村部に散布した。アメリカにも反戦運動やエコロジーはあったが、もともとは関係ない思想だった。それが、枯葉剤、二重胎児の問題で、反戦とエコロジーのイデオロギーとが結びついた。環境問題は、ベトナム戦争以前から、「沈黙の春（Silent Spring）」

（Rachel Carson）によって、社会的に認識されて、ケネディ政権の中で大きな政策に結びついた（DDT禁止）。また、ローマクラブの「成長の限界」というレポートも大変なインパクトをもたらした。

カウンターカルチャーが興隆し先進国で大学紛争が起こる中で、大学に対して大学改革を求めた。軍事研究と大学の普通の学問を分けるべきだという議論が出て、学部とは別の研究所が作られ、学部では軍事研究をしなくなった大学もある。文系に対しても、カントやアリストテレスの研究もいいが、目の前で起こっているベトナム戦争に対して何も語らないのか、それは何の意味を持つのだというような突き上げが、学生や若手の教員から出てくる。その結果、応用倫理学という分野が生まれ、生命倫理や環境倫理を扱い、ジェンダー研究も広がった。科学技術社会論もこの時代に生まれた。

中東戦争の影響で石油ショックが起こった。日本も大きな影響を受け、産業構造が変わってしまった。60年代後半から70年代は、それまでの文明が当たり前に進んでいくという成長のイメージが、最初にくじかれ、反省を迫られた時代であった。



**1970年台に現れた新しい「スタイル」**

この時代において、いろいろな考え方が出現する。例えば、カウンターカルチャーの影響から、「ソフトエネルギーパス」という考え方が出た。 Amory Lovins（エモリー・ロビンス）は非常に早い段階でこの問題に注目し、エネルギー問題の解決には、必要とされるエネルギーを疑う必要がある、バターを電気のこぎりで切るようなものが多いでのはないか、と指摘した。巨大な発電所ではなく、必要な電気を必要な場所で小さく作って、小さく消費すればいい。福島原発事故以降に、分散型とか再生可能エネルギーが日本でも注目を集めているが、そういった議論が最初に提起されたというのがこの時代である。また、「スモールイズビューティフル」は、巨大化に対して警鐘を鳴らした。科学技術がどうあるべきかを改めて考え直そうという、シューマッハーのような経済学者が出てきた。

主流派の科学者のリアクションには、代表的なものとして「テクノロジーアセスメント」が挙げられる。1969年にアメリカの科学アカデミーが出した報告書がベースである。科学技術を大規模に導入する際に、その社会的影響を予測し、マイナスの影響を最小化する。これ自体難しいわけだが、1972年にアメリカの立法府に技術評価局（OTA）ができ、最初は手探りの作業が続くが徐々に方法論も熟してきた。エネルギーから医療、環境などさまざまなテーマが扱われた。多くの利害関係者の意見を聞き、さまざまな選択肢とその結果の予想を併記した読み応えのある報告書が年間に50本程度作られた。行政が国の研究開発を推進するのに対して、議会はチェックする装置であるので、議会にこういう機関があるのは妥当である。のちの政権で廃止されたが会計検査院がその機能を担った。

「トランスサイエンス」という考え方も1970年代に出現した。核物理学者のAlvin Weinberg（アービン・ワインバーグ）が、Minervaという雑誌に、「科学に問うことはできるけれども、しかし科学には答えることができないタイプの問題」をトランスサイエンスと名付けようと提唱した。例えば、不確実性が非常に高い問題、健康リスクや環境リスク、あるいは確率が低いけれども、いったん起こってしまったら被害が大きくなる問題、つまり、原発事故や地球温暖化である。また、価値観が絡む問題、生命倫理や、エネルギーの問題などもある。原子力発電の社会的受容の問題が、この考え方が生まれた背景にはあったが、トランスサイエンスという考え方はさまざまな分野に影響を与えるキーワードになった。モダンのモデルでは、科学は事実を明らかにして、政策はその価値判断ごとに決定をする。政府の審議会は、専門家に事実を出してもらい、行政が整理して議会に提出して政策として確定させていくというプロセスを取る。けれど、安全性の問題や倫理的な問題は、最初から両者が混じり合った領域があり、やり取りをしながら物事が決まっていく。科学は事実、政治は価値という切り分けができないため、議会制民主主義や制度の枠を超えてしまう。

信頼との関係では、近年トランスサイエンスの問題が科学技術の信頼に傷がつく原因となることが多い。科学技術と政治の重なった領域は適切なインターフェイスがまだできていない。今まさにコロナでは、典型的なトランスサイエンス的な問題が毎日起こっている。日本政府がコロナ問題で信頼を落としてしまったのは、トラ

ンスサイエンスの処理に失敗したための信頼の毀損だという言い方もできる。

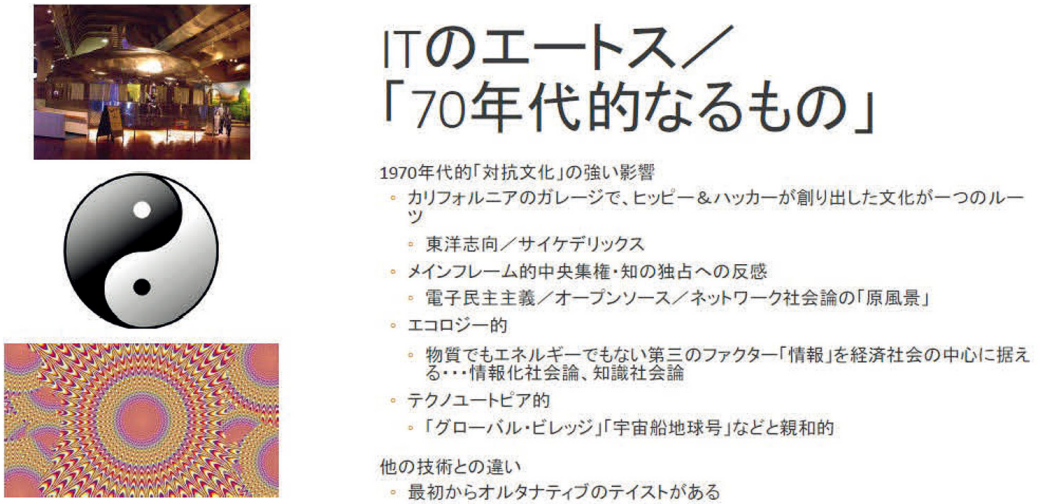


**図 2-5-2　　ITのエートス／「70年代的なるもの」**

ITの話をしておきたい。コンピューター自体は1940年代からあった。今、私たちが使っているノートパソコンやインターネットを支える思想というのは、1970年代の対抗文化の影響が強いと言える。カリフォルニアのガレージでヒッピー＆ハッカーが作り出した。若い頃のスティーブ・ジョブズらが考えたことがルーツになっている。昔のAppleは陰陽マーク（図2-5-2の左中央）を使っていた。ヒッピーたちの東洋思考から出ている。メインフレーム的な中央集権に対して、分散されたシステムで、みんなが自由にいろんなことをやるというイメージ、これが実は70年代に生まれた。

エコロジーが70年代に最初に前景化する。物質でもエネルギーでもないものを経済の中心に据えるということは非常にエコである。理屈で考えると。情報自体が社会経済の中心になれば、エネルギーやものを消費しないのではというイメージもあった。情報化社会や知識社会論もこのあたりにルーツがあると言える。一方で、科学技術によって世界は最適化されていくというテクノユートピア的な思想も同時にあったりして、反科学なのか、親科学なのか、複雑な文化が70年代に生まれた。

その影響を現代も多くの人が受けているのではと私からは見える。他の技術と違い、ITは最初からオルタナティブのテイストがある。今DXを推進しようとしているが、思想的な背景が違う技術であるということが、現代の日本においてうまくいかない理由の一つであろう。原子力や応用化学やある種の医学など、社会との大きな緊張感、摩擦を生んだ分野に比べ、ITは、社会と決定的な対立を起こしたことがない分野であるといえる。

**ELSIという考え方の登場**

ELSI（Ethical, Legal, and Social Implications / Issues）について触れる。元をたどれば1970年代のアシロマ会議に行き着く。このまま分子生物学が発展することで人類にとって悪い影響があるのではという不安を感じた科学者がいた。 DNAの構造を解明したJames Watson（ジェームズ・ワトソン）などだ。アメリカ科学アカデミー（NAS）で規制をすべきだという警告的なレターをサイエンス誌に書いたところ世界から反応があった。1975年にNASの主催でカリフォルニアのアシロマに150人の専門家や法律家、ジャーナリストが集まって国際会議を開催し、組み換えDNA実験のガイドラインを作成した。法的な拘束力はないが、アメリカ国立衛生研究所（NIH）はガイドラインに従わないものには助成金を出さないことにしたため、事実上の規制になった。その後、日本も含め各国も同様のガイドラインを制定した。

科学者は自分の研究に世の中が邪魔をするのをうるさいと思うのが普通と思うが、なぜ当時の分子生物学者は自らの行動を規制しようと考えたのだろう。いろいろ理由がある。分子生物学に参加した核物理学者は、核兵器の悲劇に自分が加担してしまったことに倫理的な感情を持っていた。そういう仕事はしたくないと思って移ってきたのに、また世の中に対して悪い影響を与えたら何のために移ってきたのか。また、当時、ナチス優生学がまだリアルなものとして記憶されていた。生命科学が政治的に使われることに対する恐怖である。さらに、参加した法律家が、裁判で負けたら莫大な賠償金を取られるため、あらかじめ自己規制をかけておいた方が良いと助言した。これが結構効いたといわれている。

もう一つの原因は、1972年に発覚した「タスキギー梅毒実験」である。アラバマ州のタスキギーという村で戦前からアメリカ政府のお金で臨床研究が行われていた。梅毒の罹患者とそうでない人を治療せずに観察をするという、事実上の人体実験であった。しかも、その対象者が全部黒人だったということでアメリカは大騒ぎになった。国家研究規制法という法律ができ、NIHの予算は事前届と承認が必要になった。

そんなことが1970年代の前半にはいくつかあったため、歴史的には例のない、科学者が自らを規制したのである。当時のマスコミは「放火犯が消防団を組織しようとしている」として評価しなかった。ただ、「科学研究は無制限に行われてよいのか」、「科学技術は誰がコントロールすべきなのか」といった問題に対しての議論に大きな影響を与えた。科学技術の倫理だとかリスクについて考える場合に引用されるのはアシロマ会議である。

アシロマ以降の15年で遺伝子組み換え操作のリスクはさほどでないと判明したところで、1990年にヒトの全遺伝子を解読しようするヒトゲノム計画が動き出す。人間のDNAを読むと、情報に関する倫理の問題、プライバシーとか個人情報の問題が当然起こる。当時ヒトゲノム計画のトップだったワトソンは、何かしないと社会的な信頼がまた壊れてしまうと思い、記者会見において、全研究予算の3%を倫理的、法的、社会的なインプリケーションの研究に充当すべきだと述べた。ここにELSIプロジェクトがアメリカの研究の世界に立ち現れる。おそらく自然科学の研究者には人文・社会系の研究の予算規模がわからなかったのではと私は思うが、3%（その後5%に上がる）の予算規模は生命倫理の分野としてはあり得ないほど高額であった。現在も続いている。科学研究の一律何%を人文・社会系に充当するルール自体が驚きである。これにより科学技術に対する社会的な研究に安定して予算がつくようになり、研究者の層が厚くなり発展していった。

ELSIは、科学者が科学技術に対する信頼が毀損してしまいそうだと考えたときに、人文・社会系にお金を入れることによって、悪く言えばアリバイ、よく言えばバランス感覚でもって、研究を正当化する仕組みを作ることができた。自らの内側に他者を入れることによって科学技術に対する信頼を担保する。発想はアメリカ的だが、有効である。最近日本政府もELSIプログラムを立ち上げている。遅いがやらないよりいい。

本日は時間もないのでRRI（Responsible Research and Innovation）[12]の詳細は省くが、簡単に言えば、研究開発の上流にテクノロジーアセスメントとかELSIの機能を入れてしまおうということである。社会的な信頼が最初から醸成される形で新しいテクノロジーを作って、社会に実装していこうというのがRRIの発想である。

懐疑の時代からELSIが生まれ、そこからRRIまできたが、科学技術の暴走にブレーキをかける役割が弱まったという指摘もある。自然科学者の側に人文・社会系の研究者が取り込まれてしまったという言い方をして、敬遠する研究者もいる。一方で、社会の中の科学技術やイノベーションの前景化は間違いなくある。科学技術と社会は共進化していくものだろう。いろいろなインターフェイスを考えていくことで科学技術に対する信頼を確保しながら発展させていく。そういう仕組みがこれからはより大事になってくると思う。

**12** RRI（責任ある／応答的な研究・イノベーション）新しい技術に対して、社会はどのような期待やニーズ、あるいは懸念や課題を抱いているのか。これを、技術そのものの特性や事業化の在り方、関連する法制度に反映させることにより、イノベーションを社会・人間・自然にとってより良いもの、責任あるものにすること（Owen et al. 2013）

**専門家の倫理**

最後に、科学や技術に対する信頼形成の基盤となる専門家の倫理について述べる。まず歴史から振り返る。ヨーロッパでは、聖職者、医者、法律家は、3つのプロフェッション、「Liberal Professions」といわれる。


**図2-5-3　　Professionalの倫理と責任**

精神的あるいは肉体的、社会的に弱い立場にある人を守る仕事（Profession）であり、商売という意味の仕事（Trade）ではない。キリスト教的な文脈の中で言えば、神にそういう立派な仕事をしろと言われて（召喚Vacation）、分かりましたと答える（表明Profess）ということになる。もともと倫理的基盤が非常に強い仕事が、この聖職者・医者・法律家である。プロフェッサーというのも同じ語源で、キリスト教的な文脈の中で位置付けられた仕事なのかもしれない。大学もギルドから発展してキリスト教的な影響の中で生まれた。

技術者のルーツはどうか。職人階層は、もともと主にアングロサクソン、ゲルマン系では徒弟制度という形でギルドなどの中で職人の後継者ができていくという仕組みとして今日まで続いている。技術は職人によって担われてきたので職人階層の役割は非常に大きい。もう一つの系統の技術者がある。フランスの絶対王政期に作られたアンジェニール、天才という意味のラテン語から来ているが、国家のために技術的な専門家として育成されたある種のエリートである。エコール・ポリテクニークという高等教育機関という形で現在も残っている。つまり、職人の流れとエリートとしてのエンジニアの流れが、ヨーロッパ、アメリカの技術者のイメージにはある。アメリカでは日本と比べて技術者の地位が高い分野が多いのは、フランスの制度的な影響も大きい国家であるためだと思う。

では、科学者はどうか。もともと科学者は3つのプロフェッションに近い倫理基盤を持っていた。ある種の宗教的な背景があったが、徐々に抜けていく。現在では信仰と関係なく科学が行われており、職人やエンジニアのような倫理的なベースが科学者にはない。科学者自身の不正が一時多かった。倫理基盤がないことが一つの原因と考えられる。

科学者自体が実は変容してしまったとよくいわれる。戦前、アメリカの社会学者のRobert Merton（ロバート・マートン）が、公有主義、普遍主義、無私性、独創性、懐疑主義（CUDOS）を守るのが科学者の倫理的態度だと提示した。その後、John Ziman（ジョン・ザイマン）という物理学者が、PLACE（Proprietary, Local, Authoritarian, Commissioned, Expert work）という言葉で、現代の科学者は、実際には、真理の追究などはしておらず、謙虚でもないし、独立性がなくて、政府や企業から委託を受け、懐疑主義というよりは専門家として答えるべきことを答えるという役割を果たしている、科学者の科学者像がマートンの時代と

はだいぶ違うものになってしまった、と言った。

**日本における倫理／信頼の基盤**

日本の倫理／信頼の基盤の問題について話す。日本の「職（しき）」は、ヨーロッパよりも広い範囲で仕事を担う。ヨーロッパでは、職人的な、技術的なことが職人の対象の範囲内になるのに対し、日本では、政治や行政、軍事、宗教などにおいて、いろいろなものに「職」という概念がある。つまり、役割のようなもので、その役割を果たすのが大人の仕事のありようである。天皇すらもある種の「職」を担う人として理解でき、そこには貫徹した論理があるとする歴史学者もいる。古典にはさまざまな「職」、仕事が出てくる。いろいろな役割を果たす人がたくさんいて、その集団として社会ができている。日本の歴史を見ると、農業を行う職人としての農民には独自の展開があり、「貴族vs.職人」というよりは、「農民vs.職人（武士も含まれる）」という対立構造があるかもしれない。日本で「技（わざ）」の本質を考えるという上で「職」という概念はたいへん重要である。

近年、日本では企業で不正事件が多く起きている。毎年のように一流企業において長年不正をしたということが表に出る。日本の産業技術に対する信頼を毀損する問題である。本質的な原因は何か。西欧社会では、医者、法律家、神父にはキリスト教文化の倫理的基盤があり、職人階層はギルド的なものから、またエンジニアは国家主義的なところから倫理基盤があると申し上げた。日本には、「自由と責任」や「社会と個人」というタイプの考え方が希薄であると昔からいわれている。加藤周一は、西洋から入ってきたものが、歴史的に似たようなものあれば受け入れられるが、ない場合にはうまく消化できないと言っている。「平等」は、日本の中にアーキタイプ（範型／原型）として集団主義があったので理解できたが、「自由や責任」は理解が難しい。確かに、自由業というと、何か勝手なイメージの方が強くて、それが責任と深く結びついた概念であるということは理解されにくい。阿部謹也も、日本には社会がなくて、世間があるだけだと言っていた。

日本にも宗教的な倫理基盤がなくはないが、現代においては強いとは言えない。職人階層の品質に対する意地もあったと思うが今はどうか。日本で倫理的基盤として強かったのは、所属組織による自治あるいは相互監視だったのではないか。しかし、1990年代以降の新自由主義的な改革の波によって共同体的な性質が希薄になった。いわば、ゲマインシャフト的な傾向があった企業がゲゼルシャフトに徐々にシフトしている。規制緩和がなされ事前規制から事後監督や事後監視になっていくがうまく機能していない。今テレワークが進んでいるが、それが一段落してから、大量の企業内の不正が出てくるのではないかと懸念している。倫理基盤がないことでさまざまな問題が起こっている可能性がある。

**科学技術への信頼のために考えるべきこと**

まとめに代えて、科学技術への信頼のために私たちが考えるべきことを述べる。「科学技術を第三者が見ている」ことは信頼確保のために重要である。科学技術に利害関係がない、独立に科学技術のことを考える知性を確保するべきであるということ。そのためには知の在り方自体を見直すことも必要だと思う。

イノベーションとのバランスも重要である。イノベーションシフトが2010年代以降に強くなった。日本政府も法律改正をして人文・社会系を動員しようと言っているが、完全に従属してしまいアリバイとして使われるようになると、ELSIの精神とは無縁なものになる。そうなると、人文・社会系の研究者の中に、政府の理工系の仕事を手伝う人と、そこから縁を切って大所高所から語る人に分かれてしまう。今回のセミナーシリーズのように、人文・社会系の研究者自身が問題自体を発見するとか、問題自体のリフレーミングに参画していくことが大事であろう。ただ、そういう文化が日本の人文・社会系の研究者にはないため、どうやってそういう人を育てていくのかというのが重要になる。

ジャパナイゼーションという言葉を聞く。いろんな意味で日本化していくことを言う。リーマンショックのときにもいわれた。日本で起こった金融危機のようなものが世界でも起こった。それ以外にもさまざまな日本化が進んでいるように思える。日本が微妙なタイミングで近代化したということが大きい。科学と技術で言えば、

科学と技術が19世紀の後半にドイツ、プロシアで合体しかけていたときに、日本は西洋近代と真正面からぶつかった。長い歴史的な背景は捨て、科学と技術が最初から結合した形で受け入れた。であるからこそ世界最初の工学部が東京大学の中に作られた。

小宮山宏が「課題先進国」ということを言い出した。日本から始まる問題は結構多い。近代工業化社会として過剰適応したからこそ、最初に近代工業社会の問題にぶつかった。私たちは、3.11にしても少子高齢化にしても、最初に向き合わなければいけない。従って、科学技術に限る話ではないが、科学技術への信頼を立て直して組み立て直す実験場に、日本はなり得るのではないか。国家にも世間にも宗教にも頼らない倫理基盤というものは可能か。非常に奥の深い問いだが、科学技術に対する信頼というのは、人とシステムへの信頼も大きいので、その基盤をどう作るのか。個人的には、もともと日本の職人階層の中にあったプライドとか、プロ集団の誇りのようなものをうまく使って、相互チェックの形を作ることで信頼を再構築できないかと感じている。

**【主な質疑応答】**

Q：いくつか質問がある。まず、日本のプロ集団の価値観は本来正しいのか。また、ITが社会との決定的な対立を起こしていないのはITが成長していなかったからではないか。最後に「コンピューターサイエンス」は科学なのだろうか。

A：プロ集団の倫理観は、職人的なプライドを復活させることである。伝統工芸の技術に対する向き合い方や品質に対する責任など信頼を保つための誇りを想定している。
ITは原子力や応用化学に比べる今までは大きな問題は起こしてない。科学技術の社会に対する影響のタイプが違うと思う。 ITがそもそも単なる技術なのか、情報とは何なのかと考えていくと、もしかしたらITはテクノロジーですらないのかもしれない。浅い言い方をすると、ITはハッピーな技術に見えている。原子力は大変な厳しい技術だが、GoogleやAppleは未来的で明るくてハッピーなもの、社会と親和的でリベラル、そんなイメージがある。今後も社会と決定的な対立を見せないにもかかわらず、社会に大きな影響を与える可能性もある。姿が見えない分だけ危ない。
コンピューターサイエンスについては詳しくないが同じように、これは単なる科学なのだろうか、とも思う。一方で、全く別ものとして発展してきた科学と技術が20世紀に結びついた。サイエンスアンドテクノロジーというよりはサイエンスベースドテクノロジーになった。科学と技術がお互いに混じり合っていく中で、純粋な科学も純粋な技術も徐々になくなっていく。分けることに意味がなくなっている時代にある。ジャパナイゼーションの一つである。

C：日本では、浮世絵や鳥獣戯画など面白いものの作り手は一種のオタクだったのではないか。興味本位に突っ走るが、自分たち社会を形成し切磋琢磨もあった。江戸時代には日本は孤立してある種の文化を発展させたが、似たようなことが今日でも起こっている。その延長で考えると、普遍性と離れて何かができるタイプの、科学的なあるいは技術的な構造を持った国なのかと思う。

A：先ほどの「職」という概念から言うと、結局この国では役割が非常に重要で、役割が与えられれば、つまりオタクならオタクという役割が与えられると、その中では内的な自由が確保されるという、一種の、オートノミーが出てくる。浮世絵師も職としてラベリングされ社会の中で一度確定すると内的自由がある。その仕組みあるからこそ、アニメとかオタクとかいろいろなものが発展してきたとも言える。一方で、そのラベリングがうまくいかない場合に、差別などの社会的な問題を起こす。今後、研究が必要な問題である。

Q：「職」というのは非常に面白い概念だが、お金がなければ何もできない。西洋のプロフェッションの考え方だと国家からお金が出るが、日本ではそうでない部分、例えばアニメにお金を落とすことは国はやらない。本当にお金を落とすべきところに落としていないような構造になっているのでは心配している。

A：産業政策と関わるとその分野は衰退することがある。伝統的には日本の場合は芸事という形でマネタ

イズすることが多い。芸事のお師匠さんになって教える。すると、なぜかお弟子さんが来て月謝を払ってくれる。そういう仕組みというのは意外といろんなところにあった。最近YouTubeとか見ながら、あれは一種の芸事かなと思う。

Q：以前別のパネルで、最近は、専門家の集団が、専門家であり素人であるという集団になってきたと伺った。そのことは専門家に対するトラストに影響しているのか。

A：世界的な傾向である。分業が進むことで、誰もがある種の専門家であり、それ以外の他の分野に関しては素人である。そうなると、ほとんどの人が自分の専門とそれ以外の分野に関する素人という組合せになってしまって、その間を架橋する方法が難しくなる。私としては架橋するものこそが教養だろうと考えている。専門の階段を専門家までは行かないにしてもゼロイチでなくつなぐような知の仕組みが新しい時代の教養であり、これからの大学の重要な役割ではないかと思う。私は、千葉大の国際教養学部に所属して教養概念の再構築を目指しているが、実際にやるのはかなり大変だ。

Q：マスコミなどで、専門家であり素人である人が自分の専門でないところで専門家っぽく言うことが不信につながるのではないか。

A：非常に難しい問題である。医師は独立性が高い仕事で一人一人いろいろなことが言える。それ自体はいいことだが意見が違いすぎて、特にコロナの初期はみんな困った。本当の専門家は誰かというのは実は難しい問題で、メディアも分からないからテレビに出ている人を連れてくる。専門性とは無関係の場合もある。

これに対する取り組みもある。例えばイギリスでは、あるNPOが専門家と呼ばれる人たちに質問状を送って、回答をウェブサイトに公開し、議論を重ねると自ずと偽物と本物があぶり出される、という取り組みを行っている。すぐにパッと答えを出すのは、逆に、民主的で自由な社会を壊してしまう。政府がこの人が専門家だと決めるのは良くない。昔の日本は東大教授を連れてくれば一応問題は解決したが今は無理。新しいタイプの権威の調達方法、権威の確認方法が必要と思う。

Q：1970年代はソフトウェアやインターネットが飛躍した年代だがお話に出てこなかった。どう思われているか。通信工学の学者と違い、インターネット系の会議では、みなTシャツと短パンで発表していた。まさにカウンターカルチャーだった。

A：技術としては陸軍のARPANETという形で研究が進んでいた、90年代以降に民間商用にオープンになったということでインターネットが実際に社会の中に出てきた。分散型自体がヒッピーの文化の影響である。冷戦期に東海岸にあった軍事関係の研究所がカリフォルニアに移ることによって、国家主義的なものとカリフォルニア的な自由なものが混じり合った。 Appleにもそういう文化のルーツが見えてくる。当時の若い人が実際に社会の中で活躍し浸透していくのが90年代以降と理解している。

## 2.6 村山 優子[13]「情報科学におけるトラスト」

**情報科学におけるトラスト研究の変遷（1）1990年より前**

2005年1月から2008年3月まで、JSTの戦略的国際科学技術協力推進事業において、Carl Hauser（当時アメリカのワシントン州立大学（WSU）准教授）と、トラストと安心に関する共同研究を行った。アメリカの大学の図書館の蔵書は素晴らしく、毎年夏休みに滞在するたびに、関連研究調査を存分に行えたことを思い出す。当時の調査結果も活用しつつ、情報科学におけるトラスト研究の変遷を表2–6–1に沿って紹介する。

情報科学の中で、トラストについて本格的に研究がされ始めたのは、1990年代以降である。

1990年以前の関連研究として、ビザンチン将軍問題（Byzantine Generals Problem）[1], [2]が挙げられる。不正者もいる中での合意形成問題で、現在も仮想通貨システムなどにおいても使われている考え方だと思う。囚人のジレンマ問題にも似たところがあり、相手をどの程度信用するかが問題となる。ロンドンに留学中、フランスの研究所、INRIA（the French Institute for Research in Computer Science and Automation）からの研究者Christian Huitemaから、当時の私の研究に対するフィードバックとして、ビザンチン将軍問題を教えていただいたことを思い出す。なお、当該問題についての論文[1]の著者、Leslie LamportはLaTeXの開発者だが、著名な分散システムの研究者である。

1990年以前でもう一つ重要な研究として、BANロジック[3], [4], [5]が挙げられる。ある二者間での認証プロトコルに基づくメッセージ交換において、受信者が、その内容から、いかなる知識あるいはBeliefを得たかを検討し、認証プロトコルの分析を行った初めての試みである。最初1989年にDECのレポート[3]として発表されたときは、独自の記法が用いられていたが、翌年にACM Transactionsに採録された論文[4]では、それらの独自記法が受け入れられず、文章での説明に置き換えられたため、読みにくくなってしまった。論文の筆頭著者のMichael Burrowsは、私がUCL（University College London）に留学中同じ講義を受けていたことがある。その後、彼はケンブリッジの大学院に行き、Roger Needham（ケンブリッジ大学教授、hは発音しなくてニーダムと読む）の下でPh.Dを取得、アメリカのDigital Equipment Corporation（DEC）の研究所、the Systems Research Center（SRC）に入り、検索エンジンAltaVistaを開発した。論文のもう一人の共著者であるAbadiは数学者で、ロジックを使って認証プロトコルの安全性を形式的に証明する仕事をした。

ここではトラストではなく、Belief（信念）について言及している。ロンドン留学中、オックスフォード大学からAIの研究チームに加わった哲学の研究者から「この世の中にはTruthは存在せず、全てBeliefである」といわれたことがある。BeliefとTrustの関係はいまだに解明できずにいる。

**情報科学におけるトラスト研究の変遷（2）1990年以降**

1990年以降になると、さまざまな情報科学の分野でトラストが取り上げられるようになった。

セキュリティー研究の流れで有名なのは、L. Jean Campの「Design for Trust」という2003年の本[7]で、トラストの概念をコンピューターセキュリティーの観点から提言したものである。 CampはもともとCMUのDoug Tygarらの下でPh.Dを取得した。Tygarは、1999年に発表した「Why Johnny Can't Encrypt」という論文[6]で、セキュアな技術もインターフェイスが悪いと使われないという、トラストとユーザーインターフェイスの課題を提起している。このような流れから、CMUのCyLabというセキュリティーの研究所から、SOUPS（Symposium on Usable Privacy and Security）という国際会議が始まった。

**13** 津田塾大学 数学・計算機科学研究所 特任研究員（岩手県立大学 名誉教授）
https://researchmap.jp/read0124223　https://www.ipsj.or.jp/award/4-murayama.html

**表2-6-1　　情報科学におけるトラスト研究**

| |
|---|
| **1990年より前** |
| ビザンチン将軍問題：ビザンチン故障など [1], [2]<br>　分散処理環境における合意形成の研究、不正者もいる中での情報処理 |
| BAN論理：（Belief vs. Truth）[3], [4], [5]<br>　認証プロトコルの安全性の形式的証明、認証プロトコルのメッセージ交換の受信者のBeliefを検証 |
| **1990年以降** |
| セキュリティーからのトラスト（CMU 関連）：<br>　PGPのインターフェイスを例に、トラストとユーザーインターフェイスの考察 [6]<br>　　Secureな技術もインターフェイスが悪いと使われない、Carnegie Mellon University（CMU）→UC Berkeley<br>　トラストの概念をComputer Securityの観点から提言 [7]、CMUで学び、Doug Tygarの下で電子商取引で学位<br>　なお、CMU CyLabからセキュリティーとユーザビリティーの国際会議Symposium on Usable Privacy and Security（SOUPS）は始まった |
| ヒューマンインターフェイス分野から<br>　トラストの定義。好感の持てる反応は顧客の意思決定に影響 [8]<br>　　ロンドン大学University College London →ドイツの大学<br>　People trust people, not technology [9]、value sensitive design（VSD）（an approach to account for human values in the design of information systems）を推進 |
| AIの分野 [10]<br>　トラストを−1から+1の範囲で定量化し、初めての計算可能なトラストモデルを提案<br>　トラストの国際会議iTrustを開催、その後、IFIP WG11.11（Trust Management：TM）を設立し、毎年国大会議TMを開催。現在の研究：Computational RegretやComputational Forgiveness |
| セキュリティーと最近のAI分野 [11]<br>　ロボットの敵意について。著者はケンブリッジのセキュリティー研究者でRoger Needhamの門下生。以前、セキュリティー心理学を提唱。今回は、状況把握やトラストに言及。 |
| 電子商取引におけるトラスト [36]<br>　推薦エージェントの利用者のトラストの感情部分に言及 [12]<br>　ホームページのレイアウトや画像などのデザイン要素が、電子商取引の売り手と買い手の間のトラストに影響 [13]<br>　主要価値類似性（SVSモデル）の電子商取引への応用 [14], [35] |
| セキュリティー分野のトラストや安心<br>　トラストモデル：Operating Systemのセキュリティー、情報のAssuranceなどの研究 [15]<br>　　トラストモデルのサーベイ論文 [34]<br>　安心の研究 [36], [31], [32] |
| 災害対応におけるトラスト<br>　東日本大震災における経験から、トラストの必要性 [16]<br>　SNSにおける誤情報の対策としてCritical Thinkingを提唱 [17]<br>　　今現在のTwitterへの投稿をリアルタイムに分析、発生している災害に関する問題・トラブルを自動的に抽出、誤報対策：矛盾する投稿をどちらも提示 [18] |
| 人間工学（Ergonomics）分野：<br>　状況把握（Situation Awareness）<br>　　各個人の状況把握の3段階モデルを提唱、複数人での状況把握（team SA）[19]<br>　　相互の信用（Belief）に基づくチームの状況把握のための認知モデルを提案 [21]<br>　災害対応における状況把握の共有のために必要なこと [20]<br>　　技術ができること：離れた場所にいる意思決定者に同じ情報を届ける<br>　　その情報をまとめ、構造化し、理解するための共通手法がある<br>　　重要な意思決定する際、制度、文化、体験の基盤を共有し、理解した意味合いを知識とする |

ヒューマンインターフェイス分野では、今はGoogleにいるRiegelsbergerがUCLでM. Angela Sasseの下にいた頃に、好感の持てる反応は顧客の意思決定に影響するということをCHI 2003で発表した[8]。電子商取引についても取り上げてている。 UCLのコンピューターサイエンス学科は、当初はインターネット中心の

研究が進められていたが、Sasseがヒューマンインターフェイスのグループを独立させて、今はドイツの大学に異動している。 ACMのCHIはヒューマンインターフェイス関連で世界トップの国際会議だが、コンピューターサイエンスだけでなく、社会科学や心理学の研究者も多数参加している。また、Batya FriedmanがCACM（Communication of ACM）に発表した論文[9]では、「People trust people, not technology」、すなわち、トラストは人間同士のものだと主張している。このFriedmanは、ヒューマンインターフェイス関連でValue Sensitive Designを推進していた。

次にAI分野で、私の知っているところでは、Stephen Marshが最初に計算可能なトラストモデル[10]を提案したといわれている。トラストを−1から+1の範囲で定量化したのだが、複雑で分かりにくい。著者のMarshとは、カナダのモントリオールで開催されたCHI2006併設のトラストワークショップで出会った。Marshはイギリス人だが、スコットランドのスターリング大学でPh.Dを取得、今はカナダの大学で教えている。Marshの研究仲間が中心となりiTrustという国際会議を立ち上げ、その後、それを引き継いだトラストマネジメントの国際会議が情報処理国際連合（IFIP）のトラストマネジメント技術グループ、WG11.11 on trust managementの年次会議として、毎年開催されている。彼のバックグラウンドは実はAIで、現在は、後悔（Regret）や許し（Forgiveness）を情報科学の観点から研究している。

Marshの論文[10]が発表されたのは1994年だったが、その後も活発に進められ、セキュリティーと最近のAI分野というと、今年2021年にRoss Andersonが発表した論文[11]がロボットの敵意を取り上げていて大変面白い。これについては今日の講演の後半に状況把握のトピックでも触れたい。彼はセキュリティーの研究者だが、本論文では、人間、ロボット、機械学習システムの間のさまざまな懸念課題について取り上げた。この論文をもとに、2021年5月のIEEEの基調講演を行うはずだったのだが、IEEE側との意思疎通がうまくいかず、発表せず、自分のホームページで論文を公開した。 AndersonはSEC2007で、セキュリティー心理学についても基調講演を行った。私が情報処理学会で「セキュリティ心理学とトラスト研究会」を立ち上げる際に、役立った。

電子商取引の分野もトラストに関わる研究が活発な分野である。顧客の購入につながるなど、結果や効果が解りやすいので、トラストのアプリケーションとして取り扱いやすいとられる。私の2013年の論文[36]では、電子商取引を中心に関連研究の調査を実施した。私はトラストの感情的な部分を安心と捉えている。トラストの感情的な部分を電子商取引で考える研究[12]を行ったのがSherrie（Xiao）Komiakで、論文を書いた2003年当時は大学院生だったが、現在は、カナダの大学の教員である。他に、Stephensは、ホームページのレイアウトや画像などのデザイン要素が電子商取引の売り手と買い手のトラストに影響するという論文[13]を2004年に発表した。それから、1999年のプロジェクトレポート[35]も詳細なトラストのレポートとして参考になる。国内ではNIIの小林・岡田[14]が、トラストの中の一つのモデルである主要価値類似性（SVS）モデルを電子商取引に利用している。

セキュリティー分野のトラストや安心については、CACMに掲載された「Trust beyond security」[15]の著者のLance J. Hoffmanは必ずしもセキュリティー研究中心の方ではないようだが、トラストについてのモデルを提示した。また、フィンランドのヘルシンキ大学のPradip Lamsalのテクニカルレポート[34]は、トラストモデルを詳細に調査しており、広く参照されている。

私が取り組んだ安心の研究[36], [31], [32]では、先述の電子商取引のトラストの関連研究調査の他、定量的調査・アンケート調査を実施した。しかし、その結果を用いてシステムに適用するのは難しい。例えば、「あなたは幸せですか」というような問いには答えにくい。その一方で、不幸な理由は挙げやすいので、否定的な観点から進めた方が解りやすかったかもしれない。 NTTの研究所では、山本らが不安の研究に取り組んでいた[37]。

その後、東日本大震災をきっかけに災害対応をいろいろやり始めた。最初は情報機器のインターネット接続の手伝いといったものだったが、そのとき対応者と支援者の間でよく意見の相違や衝突が起きたため、災害時コミュニケーションにおけるトラストの必要性を感じ、研究を始めた[16]。 HICSS（Hawaii

International Conference of System Sciences）という学会は、もともとはハワイ大学コンピューターサイエンスの研究者らが立ち上げたが、途中から同大のビジネススクールが主導権を取り、経営管理やビジネス系の参加者が大幅に増えた。幅広い分野の研究者が参加し、学際的な論文が発表されるようになった。我々の論文もそのような中で査読され、貴重なフィードバックを得た。

HICSSで出会った研究者、田中優子（現：名古屋工業大学准教授）は、一時期アメリカのニュージャージー州のStevens Institute of Technologyで研究員をされ、その時期に同僚の坂本らと共著の論文[17]を発表した。 SNSにおける誤情報の対策としてクリティカルシンキングを提唱したもので、HICSSで優秀論文賞を取得した。誤情報対策としては、NICTの大竹らのチームで開発したシステムDISAANAがよく知られている。Twitterへの投稿をリアルタイム分析して、発生している災害に関する問題やトラブルをAI技術で自動的に抽出して提示してくれるシステムで、誤情報対策の機能としては、ある投稿と、それを否定したり矛盾したりする投稿の両方を提示し、読者の判断に任せるようにしている。

最後に人間工学（Ergonomics）分野は、トラスト研究そのものではないかもしれないが、関連研究を挙げたい。私は災害対応の研究の中で、状況把握（Situation Awareness: SA）という言葉を使った。本の執筆の機会に、改めてこれについて調べたところ、人間工学では昔から一つの研究領域として取り組まれていたことを知った。1995年の論文[19]の著者であるEndsleyはアメリカ軍で、爆撃機で今自分がどこにいて、ターゲットはどこだという状況把握という軍事目的から状況把握という分野が立ち上がった。これが軍事目的以外のいろいろな分野で取り組まれるようになった。この論文では、第1段階として情報を集め、第2段階としてその意味合いを理解したら、第3段階で近い将来を予測するという、基本的な3段階モデルが提案された。さらに、複数人でそういう状況把握の形成をしようというteam SAというものも考えている。災害対応についても研究されており、このHICSSでの発表[20]は、また軍関係なのだが、重要な意思決定をする際、制度、文化、体験などの基盤を共有して初めて、情報を同じように理解することができると言っている。日本でも東大の菅野・古田らの論文[21]において、team SA、チームの状況把握のための認知モデルとして、信用（Belief）が一つの要因だと述べられている。



**安心な状態にする技術**

従来のセキュリティー研究者には「セキュリティー技術を提供さえすればユーザーは安心する」という思い込みがあったように思う。しかし、英語を母国語とする人たちが、例えば「I fell secure at home.」と言うときの「セキュア（Secure）」は、日本語では「安全」というよりも「安心」を意味していて、セキュリティー技術は本来、安心な状態にする技術なのではないかと思った。それで、安心とは何だろうと考えるようになった。

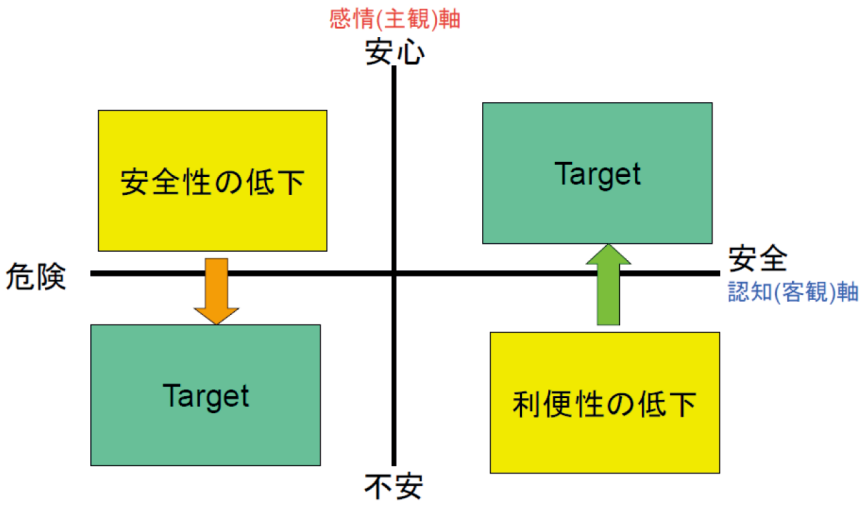


**図2-6-1　　安心と安全**

図2–6–1の横軸は認知軸、客観的に計れる安全性を意味していて、左に行くほど危険な状態ということになる。縦軸は感情軸、ユーザーがどう感じるかという主観軸で、上に行くほど安心で、下に行くほど不安な状態ということになる。 Targetとした右上は、安全でユーザーが安心している状態を意味していて、セキュリティー研究者が最も望む状態と言える。それに次いでもう一つ考えられるTargetは左下で、危険なシステムに対してユーザーは不安を感じている状態なので、使うことが避けられることになる。

一方、右下の「利便性の低下」というケースは、安全ではあるが、人々が不安に思っているので使われないという状態である。システムは安全でもインターフェイスが悪いと使われないと思うが、ここでは不安になると使わないという言い方をしている。また、左上の「安全性の低下」というケースは、危険なシステムを安心して使ってしまうという状態である。この図では、移行すべきだという矢印を上下方向に描いたが、移行のさせ方としては横方向や斜め方向もあると思う。

トラスト（Trust）と類似・関連した用語として、セキュリティー（Security）、セーフティー（Safety）、信頼性（Reliability）がある。それらの関係を図2–6–2に表現してみた。セキュリティーは、もともと不正者がいるという前提で、故意による心的な脅威に対するものとして使われる。それに対して、セーフティーはどちらかというと故意でない身体的な脅威を指すことが多い。例えば、アメリカのホテルなどで「For Your Security」は「大事なものは金庫の中に入れましょう」、「For Your Safety」は「火事のときはこうやって逃げましょう」というような関係になる。ただし、ナショナルセキュリティーやテロの脅威など、故意かどうかや、セキュリティーとセーフティーのどちらかで分けにくいものもある。

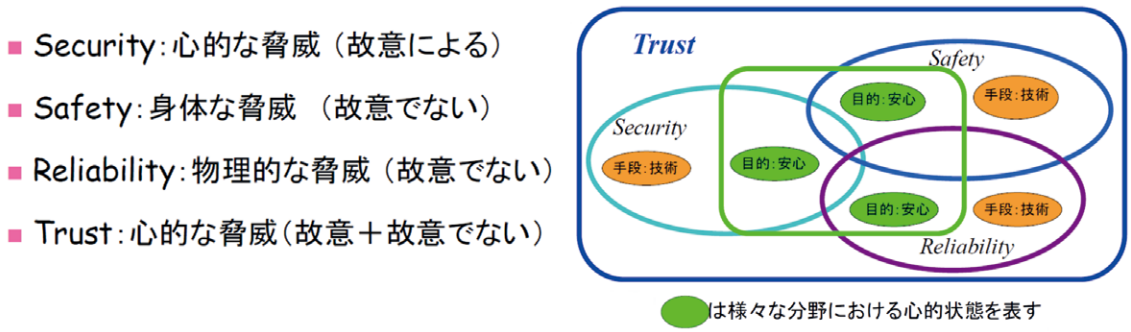


**図2–6–2　　安心やトラストに関連する概念**

それから、信頼性は、もともとハードウェアの故障率から出てきたものなので故意ではない。ただ、対象とするシステムがハードウェアだけでなくソフトウェアにも広がり、さらにネットワーク化されるようになってくると、人間によるバグや、外部からのアタックなども起き、故意の脅威も含むようになってきた。すると、システム系の技術者はディペンダビリティー（Dependability〉という言葉を使うようになってきた。トラストもこれとほぼ同じものを指すと思われ、心的な脅威に対するもので、故意のものも故意でないものも含むと思う。システム系ではディペンダビリティーが使われるが、人間寄りになるとトラストが使われるように感じている。

図2–6–2の右側に示したように、技術的な手段という面では、セキュリティー、セーフティー、リライアビリティーなどの概念が使い分けられているが、それらの目的という面では、いずれも安心のためということになる。それらの概念を全部含むものとして、トラストあるいはディペンダビリティーがある。また、トラストのモデルとしてCamp[7]やHoffman[15]が提案しているものでは、セキュリティー、セーフティー、信頼性に加えて、ユーザビリティー（Usability）、アベイラビリティー（Availability）、プライバシー（Privacy）も要素として含めている。

それで私は、それらの要素のそれぞれに、客観的に測れる部分と感情的な部分があって、感情的な部分を

まとめて「安心」というのではないかと考えた。それらの要素を全部含む全体は、トラストあるいはディペンダビリティーというわけなので、つまりは、トラスト（あるいはディペンダビリティー）の感情的な部分（Emotional part of Trust/Dependability）が「安心」と呼ぶことができると思っている。このアイデアは国際会合、国内会合、どこで発表してもこれまでのところ反対されたことがない。

**トラストとは**

前半の説明でも名前の出てきたRiegelsbergerら[8]は、「トラストはリスクが存在するときのみ必要」で、「リスクはさまざまな分野により定義や意味が異なる」と述べている。リスクは不利な結果（an adverse outcome）の可能性を表し、同義語として「不確実さ（Uncertainty）」があるといわれる。

日本における信頼研究で有名な山岸俊男[22]も、安心とトラストの差異について、安心な環境というのは、自分から搾取する要因が相手に存在しないということで、社会的不確実性の低い、つまり、相手が自分を裏切らないというような環境ではトラストは必要ない、と述べている。また、トラストが必要な環境とは、自分から搾取する要因が相手に存在する、つまり、目の前の人が今は友達だけれども、明日は裏切るかもしれないというような、社会的不確実性の高い環境としている。

社会心理学・組織心理学の分野で、吉川肇子（慶應義塾大学教授）ら[23]は「知識のある安心」「知識のない安心」として、専門家は技術的なリスクを減少させる努力だけでなく、情報提供や教育も行うべきと言っている。同じく社会心理学の分野では、Deutsch[24, 25]が、個人間のトラストを定義し、その後、他者に期待することをトラストにおけるConfidenceとして紹介している。また、社会科学者のGambettaら[26]は、他者の行動が自分に対して好意的かどうかという個人の主観的な確率からトラストを定義している。

さらに、Solvic[27]はトラストの非対称性原理ということを言っている。信頼を得るには肯定的実績の積み重ねが必要だが、信頼の失墜は一瞬で、いちど失うと再構築するのはとても難しい。構築と失墜のバランスが取れないので非対称性といわれる。信頼を崩す出来事は伝えやすいとか、否定的な事柄は肯定的なものよりも信頼評価への影響が大きいし、一般化されやすく、危険性を主張するための論拠に使われやすいとか、いろいろな要因がある。

それから、トラスト構築の3要素として、能力（Competence）、誠実（Integrity）、善意（Benevolence）がよくいわれる。5要素でいわれることがあっても、この3つは必ず入っている。

また、SVSモデル[28, 29, 14]がある。SVSはSalient Value Similarityで、中谷内一也（同志社大学教授）は「主要価値類似性」と訳された。相手が自分と主要価値、つまり、基本的な問題の捉え方が共通していると、相手の言っていることを信じやすいというものである。小林哲郎（香港城市大学准教授）と岡田仁志（国立情報学研究所准教授）の論文[14]では、SVSモデルを電子商取引に利用している。

トラストの感情部分、つまり、「安心」部分に関しては、Lewisら[30]やXiaoら[12]の論文がある。トラストは、認知的（Cognitive）な部分と感情的（Emotional）な部分があるとされ、トラストの感情的な部分を電子商取引で考えるということもされている。

精緻化見込みモデルがよく知られている。例えば、何かものを買いたいとする。そのとき、なぜそれを買いたいのか、買おうとするものについてよく知っているか、といったものが全部分かっているなら自分で決められる。図2–6–3の左側のYESの中心ルートで、相手の意見や情報をしっかり吟味して、受け入れるかを判断することになる。しかし、必ずしもそうでないときは、右側の周辺ルートによる処理になる。この場合は、相手からの意見や情報は顧みられず、相手の信頼性、つまり、相手をどれくらい信じられるかによって判断する傾向が強くなる、というモデルである。

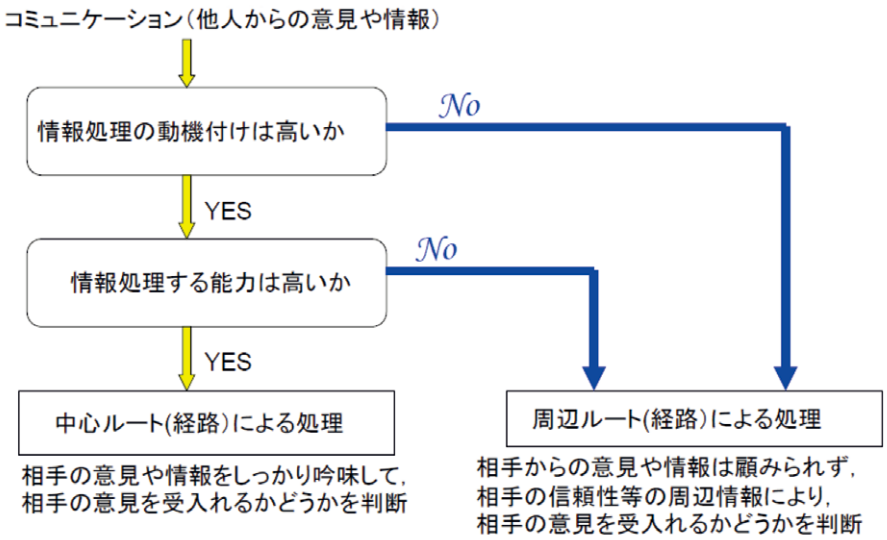


**図2-6-3　　精緻化見込みモデル**

トラスト分野における新たな研究課題について挙げてみる。英語では、不信（Distrust）がトラスト（Trust）の対義語ではない。 CHIのトラストのワークショップに参加した際、に、「Trustの反対語はDistrustでいいのかな？」と聞いたら、「Absence of trustだ」と他の参加者から言われた。トラストは「あるか／ないか」ということである。

それから、先ほど言ったように、トラストには認知的なもの（Cognitive Trust）と感情的なもの（Emotional Trust）がある。安心や不信（Distrust）は、この感情的な部分だと思う。あと、過信（Over Confidence）や、過失とトラストについては、信頼性分野、安全システム学などで研究されている。ミストラスト（Mistrust）という言葉もあるが、これは定義が曖昧になっている。

**災害対応におけるトラスト、および、状況把握**

自分が取り組んだ災害情報処理とトラストについても話したい。災害時のIT支援では、需要が必ずしも認識されないという問題がある。

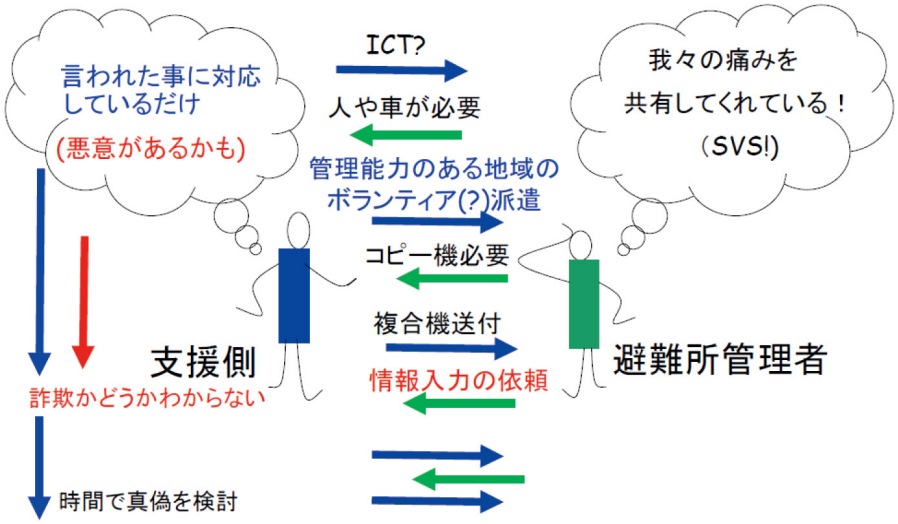


**図2-6-4　　災害時のトラスト構築での経験から**

図2–6–4のように、災害時に支援側と避難所側の間で、何が必要かに関するやり取りが行われる。避難所側から何が必要かを支援側に伝え、支援側がそれに応答することをどんどんやっていくと、ある種トラストが構築されていく。SVSモデルというものを先ほど説明したが、避難所側は支援側が自分たちの痛みを共有してくれていると思い込む。それで信用されるわけなので、私は支援側の立場にいたが、もし支援側に悪意を持った人がいたら、詐欺のようなことができてしまうのではないかと不安になった。2011年の災害対応から経験的にそんなことも学んだ。

前半の説明でMica Endsley[19]が提唱した状況把握（Situation Awareness）の3段階モデルに触れたが、第1段階は情報収集（Perception of the Elements in the Environment）で、何が起こっているか知ること、第2段階は理解（Comprehension of the Situation）で、どうして起きているのか理解すること、第3段階は予測（Projection of Future Status）で、このままにするとどうなるか予測することである。だから、災害時は情報を収集・共有するだけでなく、それが一体どういう意味合いを持つかまで共有しないと、本当の共有にならないということだ。

この状況把握をAI・ロボットのトラストと併せて考えてみたい。実際に起きた話として、UC Berkeley校の近くを歩いていたら、デリバリーロボットがいきなり襲ってきたということがあったという。そのとき感じた脅威は、野犬のようだったが、野犬とは異なりノンバーバルな合図がなく、また、こちらが脅威に感じていることを相手に知らせる術（プロトコル）がないといったことに要因があった。この経験から、ロボットがいきなり襲ってきたらどうしよう、どうやって意思疎通するんだということを考えて、Ross AndersonのところでAI・ロボットの状況把握の研究が始められたそうだ。動物では何かノンバーバルな合図がある。ロボットやドローンでも、状況把握の面からノンバーバルな合図やコミュニケーションが必要ではないか、アクシデントを避けるだけでなく、相互にトラストを構築できるのか、といったことを考えていく必要がある。



**【主な質疑応答】**

Q：図2–6–1と「トラスト」の関係は？　右上のTargetのところだけが「トラスト」か？

A：縦軸は「安心と不安」、最近は「安心と不信」と言ってもいいと思っているが感情軸である。安心はトラストの感情的な部分だと考えているので、この図全体がトラストのことを言っていると考えてよいと思う。

Q：安全と安心について考えている。安全は相手がいなくとも基準を超えているかどうかだが、安心の方は相手があるものではないか？　その相手が、何が起こるか予測できて、納得できるときに安心が得られるのではないか？

A：説明されるだけでなく、理解できて納得できることというのは、面白い視点だと思う。

Q：安心と安心感の違い、信頼と信頼感の違い、何かしらありそうにも思うのだが、どうか？

A：私が「安心」と言っているのは「安心感」と言った方がしっくりくるかもしれない。しかし、そのとき「安心」はどう定義したらよいのか、もう少し考えてみたい。

Q：安全・安心、セキュリティーの分野で、トラスト改善の目的や事例を教えてほしい。

A：私は実際にトラストの災害時応用に取り組んだ。災害の復旧時はパーフェクトでなくてもよいからスピードが重要になる。そのとき、トラストがあると意思決定に時間がかからない。相手に不信感を持っていると、相手が言ったことを素直に受け入れられず、迅速な意思決定ができなくなってしまう。また、トラストがあると相手の意見が受け入れやすくなるというのは、災害にかかわらずいろいろな意思疎通がうまくいくことになるので、障害者支援とか、さまざまなサービスのクオリティー向上につながると思う。

Q：人と人の間のトラストと比べて、人と機械やシステムとの間のトラストはどう考えればよいか？

A：目に見えるのは機械やシステムだが、その裏にはそれを開発・提供する会社の人がいて、そのポリシーが表れる。それを介して、結局は人対人の関係になるのではないか。

Q：セキュリティーで扱われるのは本人認証だが、本人だと認証されても、その本人の発言は偽だということがある。情報の中身の信頼性の扱いをどう考えるか？

A：誰も全てのことに対して専門家だということはないので、その発言の真偽は、本人認証では決まらず、Beliefでしかないだろう。

Q：情報科学・コンピューターサイエンスの研究には、人文・社会科学系の研究者があまり混じってきておらず、別々に取り組まれている印象を持っているのだが、村山先生はどう感じておられるか？

A：日本では結構縦割りで分かれているかもしれないが、海外のコンファレンスではそんなことはなく、例えばACMのヒューマンインターフェイス関連のCHIやグループウェア関連のCSCWなどには心理学者が多数参加しており、災害関連では情報システムの技術者から危機管理の立場の人や人文・社会系の人まで集まる。そういう風潮が広がってきていると思う。

## 2.7 中島 震[14]「ソフトウェア品質保証におけるトラスト」

**概要**

機械学習について、さまざまなところでTrustやTrustworthyという言葉が出てくる。従来のソフトウェアの分野でいわれていた似た概念、Trustworthy ComputingやDependabilityと何が違うのか、トラスト（Trust）とは何かということを少し整理した。それぞれの分野で同じような言葉が違ったニュアンスで使われていたり、時代とともに少しずつニュアンスが変わってきたりしている。これらについて歴史的な流れを踏まえて考える。

トラストする人とトラストされる対象があって、その間の関係性としてトラストが定義される。特に情報関係、ソフトウェアシステム関係では、トラストには2つの大きな役割がある。一つは複雑さを軽減する、つまりトラストすることによって、対象が何なのかということの複雑さを軽減するという役割。もう一つは、新しい技術の社会受容性と関係し、トラストすることでリスクをあえて取ろうとすることである。

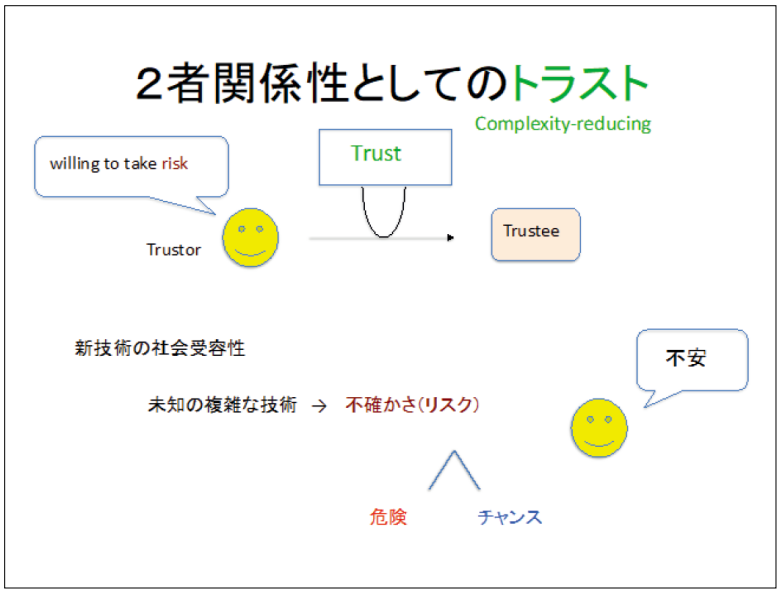


**図2-7-1　　2者関係性としてのトラスト**

一般に、未知の技術、複雑な技術は不確かさを伴う。不確かさをリスクということができる。この不確かさは害がある（Harm）ばかりではなく、複雑な未知の技術が便益（Benefit）をもたらすこともある。つまり、危険とともにチャンスという2つの観点があり、それがどのくらいのばらつきなのかということが不確かさ、リスクと考えられる。害を被ることへの恐れと便益への期待の両方に大きく広がっていると、人間は不安になり、新しい技術をなかなか積極的に使えない。そこでトラストという概念が登場する。

14　国立情報学研究所（名誉教授）、総合研究大学院大学（名誉教授）、放送大学 客員教授、国立研究開発法人産業技術総合研究所 デジタルアーキテクチャ研究センター 招聘研究員
https://researchmap.jp/nkjm/

### ITリスクとトラスト

ITの世界では、リスクの意味が先とは少し違う。そこで、ここではITリスクと呼ぶ。ITリスクは、危険側のことを考える。危険事象の発生確率と、それが起きたときの影響の深刻さ、これらの掛け算によってどのくらいの危険があるか、安全を脅かすか、を考えることが、ITリスクである。ところが、ITリスクはゼロにはできない。そうすると、ソフトウェアシステムの品質が良いというのは、ITリスクが小さいことが確かなことである、システムを使う前に不確かさの度合いが小さいと分かっていることを品質が良いと考えることになる。

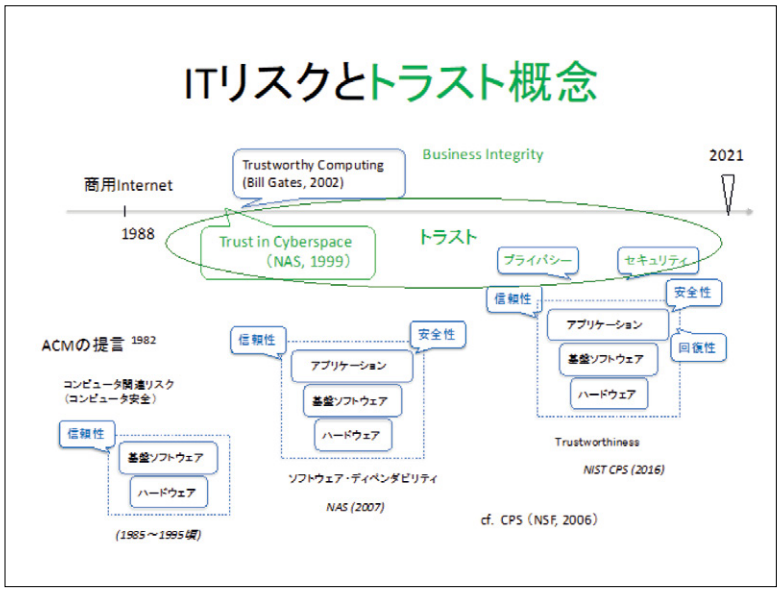


**図 2-7-2　　ITリスクとトラスト概念**

コンピューターシステムの世界でこういったITリスクのことを言い始めたのは、1982年アメリカのACM（Association for Computing Machinery）だった。コンピューターシステムが広まっているが、それが社会にもたらすITリスクが大きくなってきている、だからそのリスクを低減する、あるいは許容レベルにITリスクを抑え込めるような新しい技術の研究をしなければいけない、と論じた。この流れで、SRIのPeter Neumann（ピーター・ノイマン）がコンピューター安全について考えるという活動を始めた。リスクフォーラムを立ち上げ、実際に起こった不具合の収集、分析を行い、理由を調べていく。その狙いは、同じような不具合が世界のさまざまなところで二重三重に起きないように、不具合の状況を共有しようとすることだった。さらに、このコンピューター安全を達成する研究テーマが進められるようになった。

この頃の信頼性（Reliability）の対象は、コンピューター関連リスクといわれていることから分かるように、ハードウェアとしてのコンピューターシステムと、その上で作動する基盤ソフトウェア、オペレーティングシステムのレイヤーを対象に考えていた。その後、ITリスクに関連して、アプリケーションの比重が高まってくる。その理由の一つは、1988年頃にインターネットの商用化が始まり、それによってコンピューターシステムの使い方が変わって来たということが挙げられる。そして、ディペンダブルシステムのソフトウェア（ソフトウェアディペンダビリティー）に関して2007年にNational Academy of Scienceからレポートが出され、アプリケーションまで含めたシステムを対象にして、リスクフォーラムと同じようなこと、すなわち不具合を集めるとか、あるいは新しい技術開発をすることを推進することが提言された。ここで、信頼性（Reliability）だけではなく、安全性（Safety）も重要視することになった。簡単に言うと、ディペンダビリティー（Dependability）は、信頼性と安全性の2つの品質特性から成ると考えられる。

ソフトウェア工学の立場では、信頼性は仕様が与えられたとき、その仕様を満たすようにシステムが確かに

作られていることを言う。一方、安全性は信頼性が壊れたとき、つまり仕様通りに動作しなくなった状況であっても外界に危害を及ぼさないことである。従って、信頼性は正常系の機能振る舞いをきちんと作り込むことであり、安全性は正常系から逸脱した例外的な状況であってもITリスクを生じないことと言える。信頼性と安全性は、お互いに関連はあるが、ディペンダビリティーの2つの側面を表している。

**サイバースペースにおけるトラスト**

インターネットの登場によってコンピューターがつながることから状況が変わってきた。そこで、1999年にNASがTrust in Cyberspace というレポートを公表し、トラスティッドコンピューティング（Trusted Computing）に関わる研究開発を推進することを提言した。この頃からトラストというキーワードが、主にアメリカの研究開発支援の施策としてNSF（National Science Foundation）で用いられるようになり、その後、継続してさまざまな研究開発プロジェクトが実施された。そして、信頼性、安全性に加えて、例えばプライバシー（Privacy）とかサイバーセキュリティー（Cybersecurity）という側面が明確に認識された。



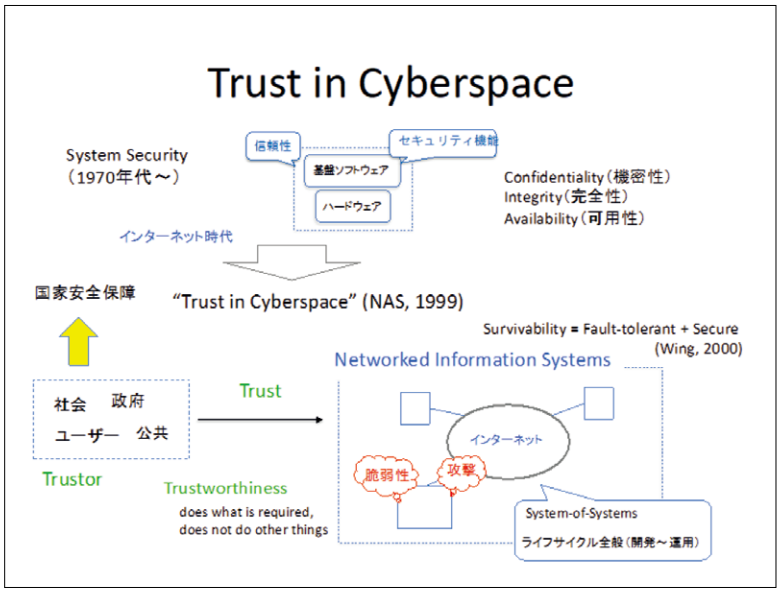


**図 2-7-3　Trust in Cyberspace**

2006年に、NSFがCPS（Cyber Physical Systems）[15]という新しい考え方を出した。これは特定の技術をいうのではなくて、CPSというソフトウェア中心システムが今後社会に浸透していく、そういう時代に合わせて新しいアプリケーション、応用、使い方や、そのための基礎技術を両輪で研究開発を進めようという位置付けの研究支援施策のキーワードである。それから約10年が経過し、CPSの考え方が浸透してきた頃、NIST（National Institute of Standards and Technology）がCPSフレームワークというレポート[16]を公表し、その中でTrustworthinessというキーワードを導入した。このときには、Trustworthinessは信頼性（Reliability）、安全性（Safety）、プライバシー（Privacy）、セキュリティー（Security）、回復性（Resilience）という5つの品質特性からなると整理した。この流れをみると、TrustとかTrustworthyという用語は、どう

15 https://www.nsf.gov/pubs/2021/nsf21551/nsf21551.htm

16 https://www.nist.gov/publications/framework-cyber-physical-systems-volume-1-overview,
https://www.nist.gov/publications/framework-cyber-physical-systems-volume-2-working-group-reports

もTrust in Cyberspaceの頃からだんだん入ってきたものであって、主としてセキュリティーが中心になっているという見方ができる。

さて、システムセキュリティーにおいてはさまざまな研究が1970年代から進められていた。特にCIAという3つの言葉（Confidentiality 機密性、Integrity 完全性、Availability 可用性）がセキュリティー機能要件であるとし、これらを満たすようなシステムを基盤ソフトウェア、ハードウェアの側から、高い信頼性によって作るという観点で見ていた。これがインターネットの時代になって少し変化し、Trust in Cyberspaceという前述のレポートが出現する。この対象は、Networked Information Systemsである。インターネットにさまざまなコンピューターシステムがつながっているという状況である。そうすると、外部から攻撃を受けるかもしれないという懸念が出てくる。さらに、セキュリティー機能が正しく高い信頼性で作られていないこと、つまり脆弱性が残るという懸念が生じる。従来のシステムセキュリティーに比して、新たな課題を考えなくてはいけなくなった。このレポートでのTrustworthinessは品質特性として定義されているわけではなくて、『期待や要求されていることをすること、さらに期待や要求されていること以外のことはしないこと』と説明されている。語感としては日常の言葉の使い方に近く、ユーザーがシステムをトラストするということを、ディペンダビリティーとの対比で、その新しい観点を強調した。ソフトウェアやシステム開発の観点でのポイントは、Trustworthinessを開発から運用までのライフサイクル全般について考えること、「トラスト」の作り込みである。



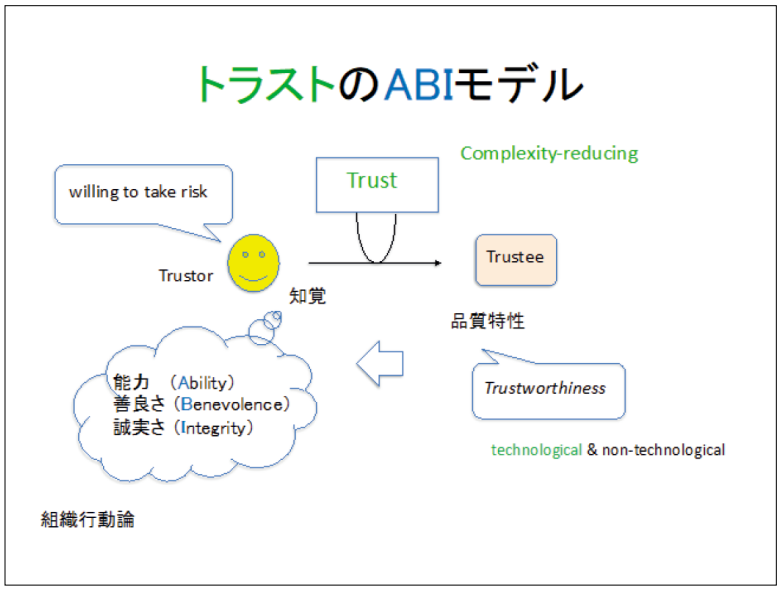


**図2-7-4　　トラストのABIモデル**

**トラストのABIモデル**

次に、トラスト関係が生まれる理由を説明する考え方を紹介する。トラストに関して、Trustworthy AIシステムに関する議論から広まったABIモデル（Ability　能力、Benevolence　善良さ、Integrity 誠実さ）がある。ここで、Trustorがユーザーであり、TrusteeがAIシステムであるとする。ユーザーのシステムに対するトラスト関係が生じることを、システムがAとBとIを示すとユーザーが知覚することと考える。しかし、AとBとIは人間が知覚する抽象的な概念である。ユーザーがこれらを知覚するために、システムの品質特性がどうあるべきか、システムとしてどのような品質特性が満たされればユーザーがABIを知覚しトラスト関係が生じるかが議論になる。つまり、機械学習AIシステムとしてTrustworthinessが直交する品質特性にどのように分解されるのかに関心が移る。ここでいう品質特性は、機械的な検査ができるということだけではなく、開発プロセスの中で担保していく側面もあり、Trustworthinessの構成を広く考えている。

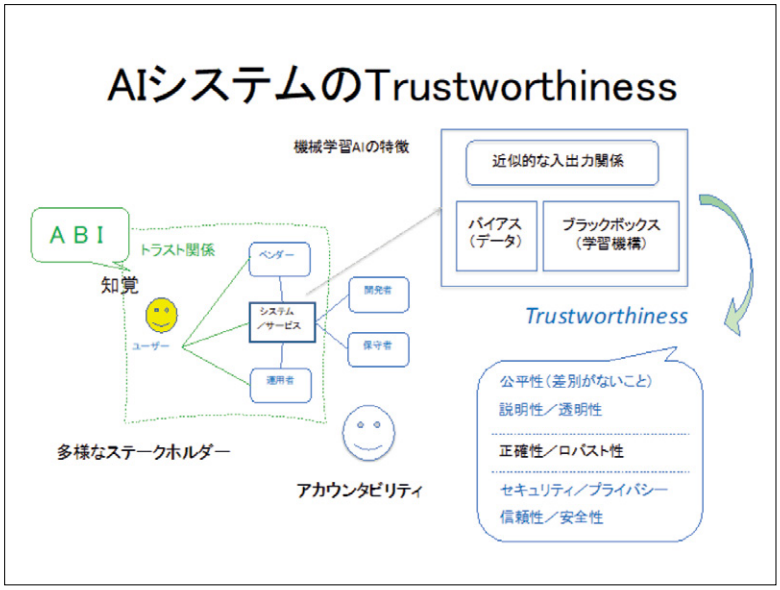


**図 2-7-5　　AI システムの Trustworthiness**

### AI システムの Trustworthiness

機械学習、特に深層学習は、基本的には近似的な入出力関係を大規模な学習データから帰納的に獲得するものである。このとき、データにはバイアス、何らかの偏りがあるかもしれない。また、入出力関係を構築するプロセスである学習機構はブラックボックスになっていて、外側から中身の細かいところまで知ることができない。一方、従来のソフトウェアでは、ソフトウェア技術者が仕様から設計し、プログラムにしていく。その過程で、きちんと中身を押さえているので、ホワイトボックスとして知ることができる。中間的な成果物や、成果物からプログラムにしていく過程で、さまざまな品質特性を確認することができる。ところが、機械学習では、そういうことができないので、いくつかの新しい性質が出てくる。図 2-7-5 では、いくつか代表的なものを挙げているが、これだけが Trustworthy AI の品質特性だということではない。多くの論文などでここに挙げたような品質特性を議論しているということである。なお、ユーザーが ABI を知覚するが、その対象は、システムやサービスそのものだけではなくて、ベンダーとのトラスト関係もあるだろうし、サービス運用者とのトラスト関係もある。このように、対象はシステムだけではなく、さまざまなステークホルダーである。

機械学習システムは運用しながら学習もする。すると、ユーザーから見たシステムの像が変わっていく。機能はよくなるかもしれないが、それにしても変わっていくので、ある種の予測性が低下する。機能が改善する、性能がよくなるという意味で便益が大きい側に振れるかもしれないが、ユーザーの予測性が低下する。つまり、良い方に振れるとしても不確かさが大きくなる。 IT リスクではなく、不確かさの増大という意味でリスクが大きくなり、ユーザーの不安が増大する。 NIST が発行した Trust and Artificial Intelligence レポート[17]は User Trust と呼んでいる。使う側の立場から見て運用中の変化に起因するリスクの低減を論じる見方としてトラスト概念を持ち込んでいる。このレポートでは、Trustworthiness がいくつかの品質特性からなると考え、これの重みづけによって、ユーザーが知覚する Trustworthiness を測定するという考え方を提案している。これは学習によって機能が変わっていくという機械学習の特徴に焦点をあてた議論である。

次に、Trustworthy AI の特徴を分析しようというのが欧州の AI 倫理ガイド[18]である。トップダウンな考え方で、基本的人権から始めて、3 つの原理、すなわち Lawful、Ethical、Robust を導入し、それを満たすた

17　https://www.nist.gov/publications/trust-and-artificial-intelligence
18　https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

めの要件（Requirements）をまとめようとしている。

基本的な発想は倫理規範（Ethical）に従うこと、システムが頑健（Robust）であること、というトップダウンな要請から始まり、4つのAI倫理原則を抽出し、これをさらにブレークダウンすることで、Trustworthy AIに対する7つの要件を洗い出している。7つの要件は今までソフトウェアシステムに対する品質特性として考えていたものよりも、抽象レベルの高いものになっている。トラストする側は、例えばユーザーであり、トラストされる側は、機械学習システムそのものだけではなくて、開発ライフサイクルに関わるさまざまなステークホルダーになる。従って、技術的な話だけではなく、もう少し高いレベルの概念が入っている。

この欧州のレポートの他にもAI倫理の側からTrustworthy AIへの要求を引き出そうというようなアプローチが進められているが、Brent Mittelstadtは、機械学習の世界ではそういうアプローチが難しいことを論じている。一般的に倫理原則を考えるときには、医療倫理原則に模範を取ることが多い。医療倫理原則はヒポクラテスの時代から長い歴史があり、実践の経験も豊富にある。医療倫理原則を社会の中にインプリメントしていく方法が明確になっており、それがトラストにつながっている。一方、AIは歴史が短く、実践の経験も少ない。AI倫理原則をうまく生かすような、原則と実務との間のギャップが大きい。

また、別の難しさもある。ソフトウェアや機械学習の技術というものは基盤的であり、実際にアプリケーションあるいはサービスとして便益を享受するのは、特定のアプリケーション、例えば医療AIシステムである。すると、倫理原則には2つの観点があることになる。仮にAI倫理原則がうまく整理できたとすると、アプリケーションの基盤としてのAIが倫理原則を満たしていることが最初の観点。次に、アプリケーションの世界での倫理原則、この場合は医療倫理原則であるが、それが満たされていることという観点がある。場合によっては2つの観点が整合しないかもしれない。 AI倫理はアプリケーションに依存しないメタレベルの議論になるので、個別分野の倫理との関係も見なければならない。 AI倫理原則のみから始めて、Trustworthy AIの技術特性を考える、あるいはそれを満たすような実務を考えることだけでは無理がある。

**Trustworthy AI エコシステム**

Trustworthy AI エコシステムでいうエコシステムは技術コミュニティーを指す。そして、機械学習AIに関わる研究者やエンジニアあるいはベンダーからなる技術コミュニティーが何をしなければいけないのかを論じている。例えば、敵対データ例（Adversarial Examples）の不具合を、技術コミュニティー全体が協力して低減していこうということが述べられている。そのために、発生障害の事例を収集し、どのような不具合が起きているのかを整理し、さらに、基礎的な研究を行うと論じられている。また、ここではVerifiable Claimsというキーワードが挙がっている。 Verifiableは検査できるという意味であり、検査可能な表明があれば、誰かが開発した機械学習AIシステムを第三者がチェックでき、監査に使えることが論じられている。複雑なソフトウェアのシステムに対して認定証（Certificate）を与えるというのは、ディペンダブルシステムについても論じられていた。北米のコミュニティーにおいて、認定証を与えることを可能にする技術を開発することが繰り返し議論されている。

Trustworthy AIがいくつかの品質特性からなると考えたとき、従来のソフトウェアのDependabilityやCPSのTrustworthinessで論じられてきた品質特性に加えて、AI倫理原則、つまりトップダウン的な分析あるいは思想から出てきた品質特性が一体になってTrustworthy AIの特性が論じられている。

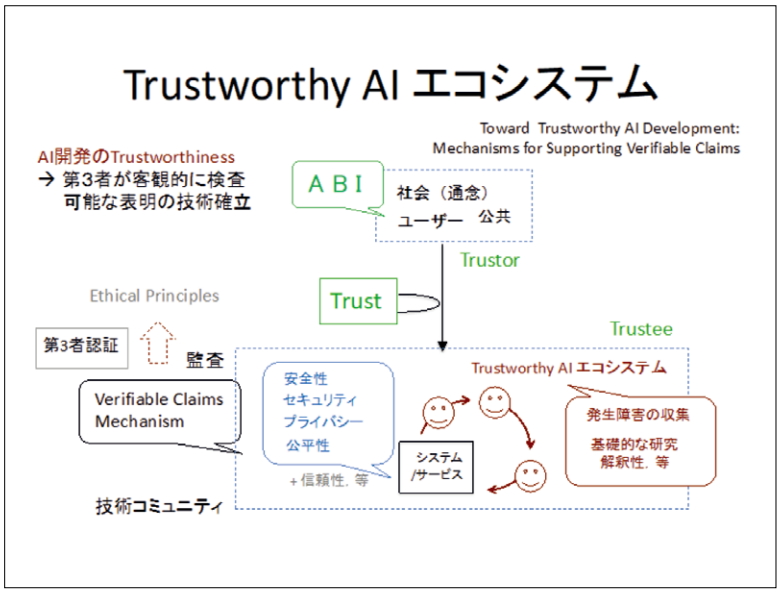


**図 2-7-6　　Trustworthy AI エコシステム**



**国際標準**

ソフトウェアの品質に対するモデルにSQuaREがある。データ品質と製品品質の2つの観点で開発時の品質を評価し、ユーザーが使うときの利用時の品質に影響する、という考え方で、品質モデルと品質特性が定義されている。SQuaREでは、トラスト（信用性）という品質特性が挙げられているが、意図通りに動作することの確信という意味であって、狭い範囲でトラストを捉えている。最近になって、Trustworthy AIの特徴的な品質特性を整理追加することでSQuaREをAI拡張することが議論されている[19]。これまでのところ、AIシステム全体の品質に大きく影響する学習データの品質モデルが明確には論じられていないようである。

また、従来ソフトウェアを対象として、Trustworthinessというキーワードを使っている国際標準がいくつかあるが、広義の信頼性ということであって、NISTのCPSフレームワークで言っていたTrustworthinessと同じように、信頼性（Reliability）、安全性、セキュリティー、プライバシー、回復性のことを指している。現在、ISO/IEC JTC1 WG13[20]でTrustworthinessを明確に定義しようとしている。

AIに絡んでTrustworthiness in artificial intelligence[21]という国際標準の議論が進行中である。ここでは、Trustworthinessというのは検査可能な方法によってステークホルダーの期待に沿う能力であると言っている。具体的な品質特性を定義するのではなく、いくつかを例示するにとどめている。今後、技術の発展とともに、Trustworthy AIについて、新しい品質特性が出て来ることを想定している。

**まとめ**

トラストという言葉はソフトウェア工学やソフトウェア品質保証の観点から見ても多様な解釈があり、何を議論したいか、目的によって定義を適宜選択している。 Trustworthy AIにおけるトラスト定義に着目すると、3つのアプローチがある。心理学的なアプローチ、組織行動論的なアプローチ、AI倫理から見たトップダウンのアプローチである。

そして、コンピューターソフトウェアとしてのAIの品質特性の全てを同時に考えるのではなく、おのおのの

19 https://www.iso.org/standard/80655.html
20 https://jtc1info.org/technology/working-groups/trustworthiness/
21 https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:24028:ed-1:v1:en

品質特性をカテゴライズして考える方が分かりやすい。従来の高度なソフトウェアシステムに対して議論されていたこと、機械学習の特徴的なメカニズムから新しく入ってきたこと、さらにAI倫理からの要請で考えるべきこと、こういうものに分けることである。また、従来から議論されてきた品質特性でも機械学習の特徴を考慮して再考すべきだろう。例えば、敵対データ例と安全性の関わりなどである。

**【主な質疑応答】**

Q：第三者認証の話が出たが、セキュリティーの世界でも必ずしもうまく行っていない。

A：第三者認証の例としては機能安全がある。ソフトウェアの場合は偶発的な故障が起きないので、ハードウェア中心の機能安全の枠組みに、どう合わせるかに苦心し、そこにプロセスで担保するという考えが導入された。機械学習AIにおいても、このようなプロセス的な担保になるのではないだろうか。それがどんな具体的な性質なのかについて、コンセンサスはまだ得られていない。Verifiable Claimsが例示されてはいるが、トップダウンのEthical AIのような概念からくる性質と今後結びついていくかもしれない。

Q：使い方というところについては利用時品質が関係するのではないか？ 製品の仕様にはなかなか反映されにくいが、だからこそその製品を使っても大丈夫なのというところが利用時品質に関係するのではないか？

A：利用時の品質を詳細に見たわけではないが、例えば製品品質についてはどのような拡張が必要か、利用時の品質、特にリスク回避性をどのように考えていけばよいのか、何を新しく考えるかという視点の中に、Ethical AIで議論され、整理されるものが入ってくるというだろうとNataleたちは述べている。例えば、公平性について考えてみると、データ公平性、機械学習機構と最適化の目的関数の選び方、結果の活用の仕方などいくつかの要因が関わる。これら全てをSQuaREと同じような品質定義の詳細レベルで論じるのは難しい。しかしながら、公平性をひとくくりにして、利用時の品質に入れるのは、抽象的過ぎる。例えば、データ品質のところにデータ公平性、製品品質のところに公平性に関わる訓練済み学習モデル構築品質の担保、というように、分けて考えるのではないだろうか。

Q：トラストとTrustworthinessの違いについて聞きたい。トラストというのは全てを検証できてはいないけれども、そこも含めてこの人を信じるとか、このシステムを信じるということだと思う。一方、標準化や工学的な保証というのは、検証可能な範囲の話であり、それをTrustworthinessとして定義しているという印象を持った。こういう理解でよいのか？

A：個人的な意見であるが、多くの議論が、質問者の見方に賛成していると思う。トラストを知覚するのは、人間の側の話であり、複雑さを軽減するために、トラストに頼ってしまい、あとは見なくしてしまう。しかし、見ないけれども、リスクをあえて取ろうという気にさせるというのがトラストである。
一方、何の根拠もなくトラストを持つわけではないので、システム側もある種の歩み寄りが必要になる。検証可能な品質特性としてのTrustworthinessを決めておき、そこまでは何とか技術サイドがやる。しかしながら、そこにはギャップが生じる。そのギャップについては、今の議論では何も話すことができていない。まずVerifiable Claimsのセットを作り出すことが第一歩であろう。

Q：Trustworthy AIでは、検査できるものとノンテクニカルなものがあるということであり、ノンテクニカルなものとして開発プロセスが挙げられていた。例えばISO9000の監査においては、品質の責任者が決まっているとか、品質目標を設けているといったチェックリストを作っていた。こういうことがノンテクニカルな評価項目だと思えばよいか？

A：従来のソフトウェア工学との対比でいうと、確かにおっしゃったようなプロセスモデルになる。そこで掲げていることは、多分参考にすべきである。
そして機械学習システムで困るところは、いわゆるProof of Conceptのところである。ISO9000でやっているような開発プロセスの成熟度モデルがアジャイルな世界でどのように論じられているのかという

と、まだ課題がありそうである。

ABIというのはあくまでも人間が頭の中で知覚するものである。従って、Trustworthinessの品質特性とか、Verifiable Claimsとかがいくら定まったところで、そこには大きなギャップがある。そこで第三者認証の考え方が出てくる。有力な認証機関でお墨つきをもらえば、ユーザーは安心して使うことができる。認証機関をトラストすることによって、複雑さを軽減している。つまり複雑さを軽減するやり方というのは技術だけではないし、認証のような社会システムも含めて、価値共創のエコシステムとして考えていこうというのが、Trustworthy AIエコシステムであると思われる。

Q：学習する商品がフィールドに出てから学習することによって、工場出荷時から仕様が変わってしまい、振る舞いが予測できなくてユーザーの不安が増すというのはよくあると思う。そのような状況では、どのようにしてTrustworthinessを決めるのか？

A：機械学習システムとしてはより高度になっていたとしても、人間との知覚が合わない。それによっては不安を感じたり、そのシステムを使うことに対するトラストが減じてしまったりする。 NISTのUser TrustはTrustworthinessを仮に決めた後の計算方法を提案している。

Q：技術者コミュニティーという話があったが、技術者自身にもバイアスがある。 AIを使ったシステムではそのバイアスが露見した。コミュニティーは、技術者だけではなく、広くあるべきかと思う。

A：ここでは、分かりやすさのために技術コミュニティーと書いた。ここでいうTrustworthy AIエコシステムの中には、エンジニアサイド、あるいはアカデミックな人間だけではなくて、ちょっと違った見方ができるアクターもいる。 AIシステムに関しては、どういうことができるのか、何が検査できるのかという技術的な要素というのが、非常に未知で大きいので、技術コミュニティーの役割としてのオーケストレーションが有用である。むしろ技術側の人間がリードするべきであるという考えもある。

Q：AIに限らず、大きなシステムになると、サプライチェーンが非常に長くなる。すると、供給者をトラストするというのが問題になり、最終的にユーザーが使う製品までたどり着くというときのトラストの流れはどうなるのか？

A：ベンダーから納品されたものに対してユーザーが検収をするときにどんな検査をすればよいかは、今までのシステム構築の経験から分かっている。しかし、機械学習システムの場合には、そういう検収そのものが技術的には難しいので、サプライチェーンをぶつ切りにするのではなくて、全体が一つの価値共創（注：AI開発での包摂性）という形で見ていかなければならない。一方で、Verifiable Claimsがうまく整理できれば、サプライチェーンの一つの受渡しのところでチェックできるかもしれない。

Q：自動運転の話が出ていたが、仮にトラストできるものを作ったとしても、それはあくまでもテスト環境におけるトラストにすぎない。現実の場が予測できるような形のトラストを製品レベルで求めることは可能か？

A：確かに機械学習システムの場合にはそういうことが起きるが、それは従来からオープンシステムといわれているものに共通すると思われる。従って、今までに培われたさまざまな経験から少しずつそれをアダプトしていくことになりそうである。そのときにはもちろん、機械学習ならではの特性が新しい技術課題を生むのかもしれない。

C：自動運転にしても、他のソフトウェアにしても、お互いにインダクションするようなとき、非常に複雑であり、かつ相手のスペックがよく分からない状況で動いてしまう。自動運転車同士がどういうように情報交換するかということについては、プロトコルを作ることは可能だが、人間が運転する車が相手の場合、技術的な予測が困難な問題になる。そのときにどうするかということがエコシステムとしての目的の一つだというふうに考えざるを得ない。

## 2.8 松本 泰[22]「ゼロトラストから考えるトラストアーキテクチャー ～トラストのメカニズムのパラダイムシフト～」

### トラストのベースとなる考え方、用語の整理

トラストについては、Niklas Luhmann（ニクラス・ルーマン）著作の「信頼―社会的な複雑性の縮減メカニズム」[1]が引用されることが多い。著作の中でルーマンは、トラストは社会生活の基本的な事実であるとし、我々が他者や社会に対して一定の信頼を置いていることから社会生活が可能になっている、と説明している。トラストというのは、水や空気と同じような存在であって、トラストがなければそもそも社会生活が成り立たない、そのようなものだと捉えることができる。

トラストを議論する際に前提として整理すべきは、誰が誰にトラストをするか、という関係である。ここに齟齬があると議論が噛み合わない。図2–8–1に示すように、「Relying Party」がトラストする主体、「Trusted Party」がトラストされる対象者、そして「Trustworthiness」はRelying PartyがTrusted Partyに期待する性質、と定義することができる。例えば、我々が医師と対する際、そこには何らかの社会的な複雑性の縮減メカニズムがあって、Relying Partyである我々が、Trusted Partyである医師をトラストするのだろう。その際の社会的複雑性縮減のメカニズムは、医師資格という国家資格や、医師免許証という医師資格の証明、医療機関の認可制度、その他医療の公平性を支える国民皆保険制度などの多くの法制度がベースになっていることが多いと考えられる。

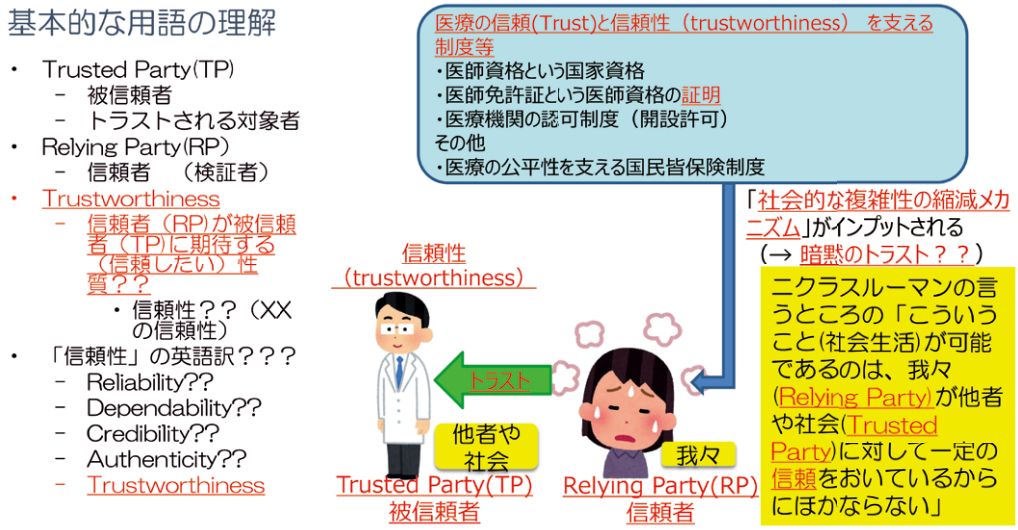


**図2–8–1　　トラスト　用語の整理**

### デジタル社会におけるトラスト

複雑性やブラックボックス性が増すデジタル社会において、Relying Partyである我々が、Trusted Partyをトラストするというのはどういうことだろうか。これまで社会的な複雑性縮減のメカニズムとして社会に深く組み込まれてきたものに、例えば、紙台帳、ハンコなどの押印、手書き署名、物理的な証明書、紙幣などが

22　セコム（株）IS研究所ディビジョンマネージャー
https://www.secom.co.jp/isl/

ある。しかし、これらは複雑性やブラックボックス性に対応できず、また、現在のデジタル技術によって壊されてきており、何らかの新しい仕組みを構築する必要に迫られていると感じている。

従来、Relying Partyである人間に注目したトラスト研究は、社会心理学などの分野で多く行われてきた。一方、デジタル社会において重要なのは、Relying Partyが人間だけではなく、マシンもあるという点である。スマート化・自律化したシステム、サイバーフィジカルシステム、デジタルツインなどの多くにおいて、マシン対マシンの信頼関係が必要になっているが、こうしたデジタル社会においては、Trusted Party をVerifiableとすることでトラストをスケールアウト可能にする方向に向かっている。

今後のSociety 5.0社会に必要となるデジタルトラストの要求・技術は、図2–8–2のように整理することができる[23]。

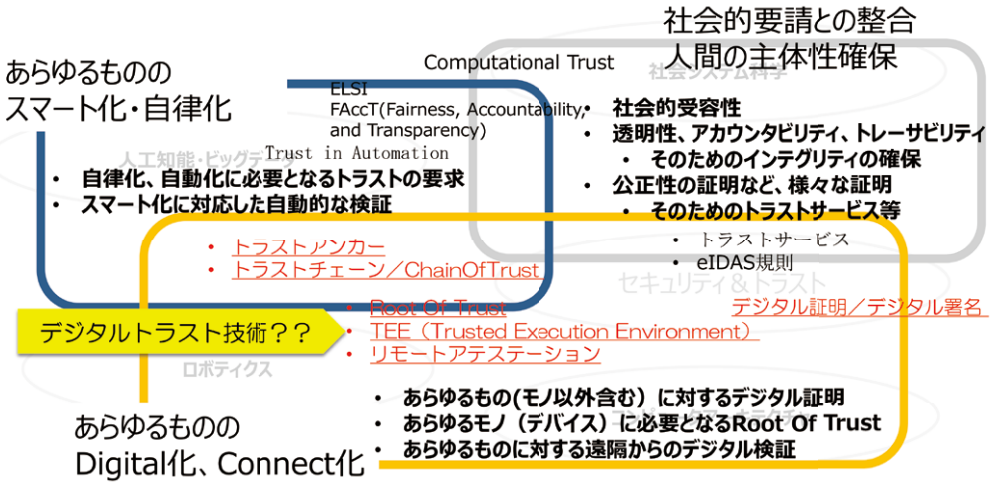


**図2–8–2　　Society 5.0社会に必要となるデジタルトラスト**

**ゼロトラスト**

「ゼロトラスト」は、セキュリティー業界ではバズワード化し、また、マーケーティング用語化している。いろいろな解釈が後付けで登場しているが、従来のセキュリティーが、境界線防御・物理ゾーニングで守られたネットワーク（トラステッドネットワーク）前提のアプローチだったのに対して、別のアプローチのセキュリティーになる。同じ建物内などの境界線内で守られている相手であればトラストするという境界線防御の考えに対し、そのようなトラステッドネットワークは存在しないと考えるのがゼロトラストの世界観になる。

このようなゼロトラスト環境の中で、何らかの形で相手をトラストするために、「Never Trust, Always Verify」といった考え方がある。ここで、NIST（National Institute of Standards and Technology：アメリカ国立標準技術研究所）による定義[2]として、信頼者であるRelying Partyを「リソース」、被信頼者であるTrusted Partyを「サブジェクト」として説明する。図2–8–3に示すように、リソースは、ゼロトラスト環境下のサブジェクトに対し、常に検証（Always Verify）を実行する。具体的には、サブジェクトであるユーザー、使用するデバイス、アプリケーション全てに対し、リソースへのアクセス時にVerify（Always Verify）を実行する。このVerifyが成立すれば、境界防御の考えなしに、リソースはサブジェクトを直接トラストする

23　図2–8–2における「あらゆるもののDigital、Connected化」「あらゆるもののスマート化・自律化」「社会的要請との整合、人間の主体性確保」という枠組みは、JST CRDS「研究開発の俯瞰報告書：システム・情報科学技術分野（2021年）」に示された分野全体観を踏まえた。

ことができるし、Verifyできなければ Never Trustとするといった考え方になる。サブジェクトの Trustworthinessは、サブジェクトであるユーザー、デバイス、アプリケーションのセキュリティー上の信頼性（Trustworthiness）などによって決定する。

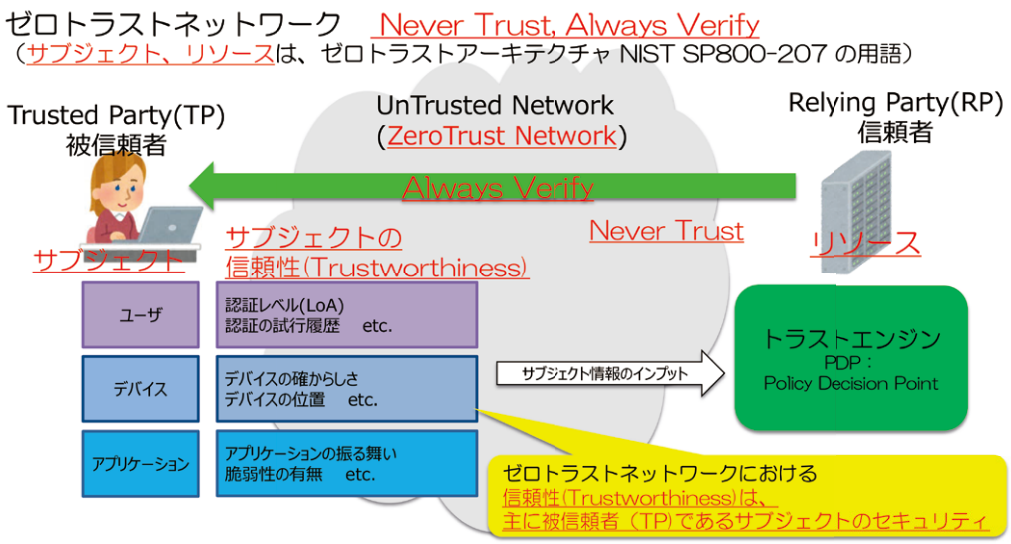


**図 2–8–3　　ゼロトラスト環境における Always Verify の仕組み**

### PKI（Public Key Infrastructure：公開鍵暗号技術基盤）

「Verify」を実行する基本的な仕組みとして、PKI（Public Key Infrastructure：公開鍵暗号技術基盤）がよく利用される。PKIでは、信頼者が「Relying Party」、被信頼者である Trusted Partyが「Subscriber（署名者）」と呼ばれる。

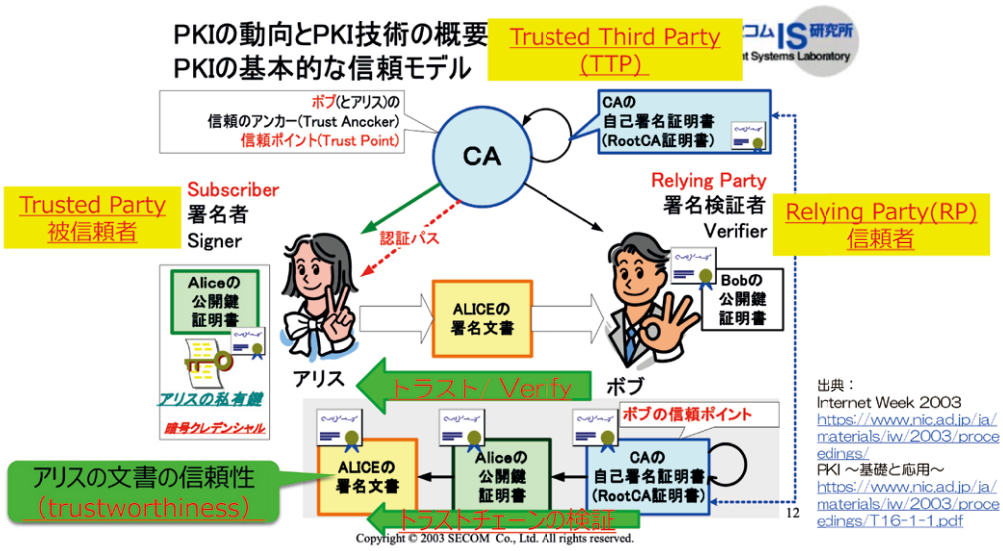


**図 2–8–4　　PKIによる簡易な信頼モデルの例**

図 2–8–4 はPKIによる基本的な信頼モデルである。 Subscriberであるアリスのデジタル署名が付された文

書があり、それをどのようにRelying Partyであるボブが検証するかについて示している。具体的には、信頼できる第三者機関であるTrusted Third Party（TTP）がアリスに公開鍵証明書を発行する。 PKIでは基本的にCA（Certification Authority：認証局）がTTPに相当する。 Relying Partyであるボブは、CAの発行したルートCA証明書の中に格納された公開鍵をトラストし、それを手がかりに、アリスの文書に至るまでのトラストチェーンの検証、すなわち署名の連鎖の検証を行うという仕組みである。ゼロトラスト環境におけるVerifyは、基本的にこのような流れで実施される。

このPKIによる検証の仕組み自体には、ネットワークをトラストする必要はない。従って、Relying PartyとTrusted Partyの間には、物理的な境界線防御も必要ないことになる。ただし、公開鍵証明書を発行するCA自体は、強固な境界防御が施され、HSM（Hardware Security Module）と呼ばれる、署名鍵を守るための高いセキュリティーを実現するハードウェアが使用されることが多い。このことにより、トラストできない環境における証明書などの署名データの信頼性（Trustworthiness）を確保している。



**リモートアテステーション**

「Always Verify」を可能とする仕組みとして今注目されているのが、リモートアテステーションである。リモートアテステーションとは、Relying Partyから見て遠隔にいるTrusted Partyが、Relying Partyの期待した通りの信頼性（Trustworthiness）を維持しているか、Relying Partyが期待した通りの公正を保っているのかをVerifyする仕組みであるが、ここではデジタル署名が多用される。

リモートアテステーションは既にいろいろな分野で実装されているが、標準化があまり進んでいないという課題がある。具体的なリモートアテステーションの概念や、アーキテクチャー、用語の整理などがIETF（The Internet Engineering Task Force）において議論されているところであり、この議論が進めば、リモートアテステーションの必要性や、あるべき姿がより具体的に見えてくるだろう。

リモートアテステーションの全体像を、医師ないし医療システムへのトラストの例で紹介する（図2–8–5）。リモートアテステーションでは、アテステーションの対象である医師を「Target–Environment」と呼ぶ。「Attester」は、その医師の信頼性に関する情報を「Evidence」として「Verifier」へ送る。 Verifierは「Evidence」を検証して、アテステーション結果としてRelying Partyに返し、それによってRelying Partyは医師をトラストするかどうかを決定するという流れである。

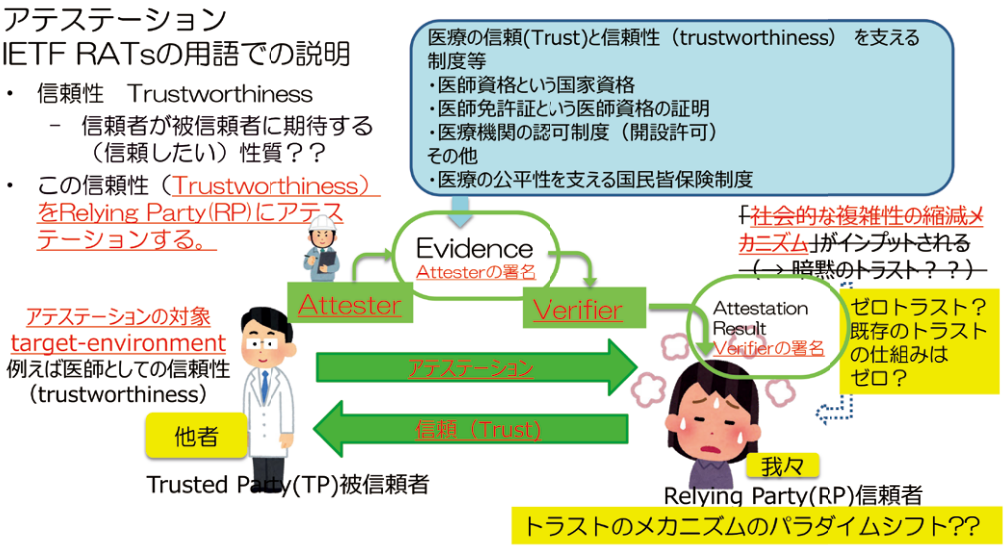


**図2–8–5　　リモートアテステーションの流れ**

IETFのRATS WGドラフト[3]では、7つのユースケースを示し、リモートアテステーションのアーキテクチャーを議論している。ユースケースとしては、「機械学習モデルの保護」「機密データ保護」「FIDO（Fast IDentity Online）バイオメトリクス認証」などが挙げられている。「機械学習モデルの保護」では、AIエッジデバイスに格納されている機械学習モデルが、提供者が意図した通り設定され、（知的財産的に）保護されているかどうかをアテステーションする仕組みである。「機密データ保護」では、コンフィデンシャルコンピューティング技術などの利用が想定され、「機密データ保護」が機能しているかどうかをアテステーションする。 FIDOは、パスワードレス認証であることで注目されているが、その中でバイオメトリクス認証を使う場合、デバイスに実装されるバイオメトリクス認証器（Authenticator）が意図したものかや、バイオメトリクス認証によるローカル認証をどのようにリモートに伝えることができるかが従来から課題であったが、これらをアテステーションの仕組みで行おうとしている。

**信頼の基点、変貌するトラストアーキテクチャー**

ゼロトラスト環境において、上述のようなリモートアテステーションのユースケースを考えると、具体的には使用するデバイス（サブジェクト）とサービス（リソース）間でのアテステーションを行う必要がある。Always Verifyは、ユーザーのデバイスにAttesterが組み込まれ、Attesterから送られたEvidenceをVerifierが検証する仕組みで行われる。

これを実現するため、例えばAppleのiPhoneでは、「セキュアエンクレーブ」と呼ばれるハードウェアが実装されている。これは、メインのアプリケーションプロセッサーとは別の、独立したプロセッサー（セキュアエンクレーブプロセッサー）であり、強固なセキュリティーを実現する信頼の基点（RoT：Root of Trust）になっている。この分離されたハードウェアにおいて、iOSとは別のトラステッドOSが起動し、通常のアプリケーションからは論理的に隔離されたトラスト領域（TEE：Trusted Execution Environment）が動いている。 TEEには内部にAttesterが組み込まれており、ユーザー認証やデバイス認証、アプリケーションの監視を行い、Attesterがリモートにアテステーション結果を伝えるという仕組みである（図2–8–6）。

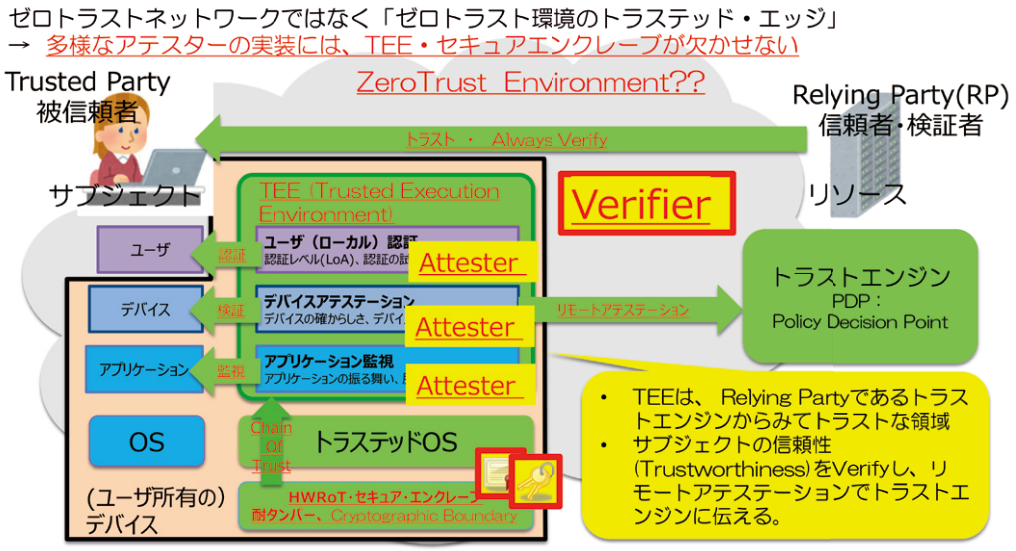


**図2–8–6　　ゼロトラストにおけるRoTとアテステーション**

境界線防御の効かないゼロトラスト環境においては、このようなハードウェア（チップ）内の境界線防御により信頼の基点を構築することが重要になってきている（図2–8–7）。 Appleのiphoneなどでは、デバイス

の識別子（UID：User Identifier）、メインCPU、セキュアエンクレーブプロセッサーも含めSoC（System On Chip）として不可分な「Physical Boundary」を形成している。そして一番重要となるデジタル署名や暗号化を司るセキュアエンクレーブプロセッサー内では、暗号鍵を内包し暗号演算を境界線内で行う「Cryptographic Boundary」で、耐タンパー性などの物理的にも強固な境界線防御が施されている。こうしたことにより、ゼロトラスト環境における物理的な攻撃にも対応できることになる。

また近年では、信頼の基点（Root of Trust）と暗号技術が組み込まれた新しいコンピューターアーキテクチャーにより、チップ外にさまざまなバリエーションでトラストな環境を拡張する動きが見られている（Trusted Boundary）。代表的なものとして、Intel社のCPUに実装されているIntel SGX（Software Guard Extension）というセキュリティー機構がある。SGXは、メモリー上にエンクレーブ（飛び地）と呼ばれるセキュアな領域を生成することで、使用中のデータ（Data in Use）を暗号化し、保護したまま処理を実行する。これは、クラウド上におけるトラストを実現するものとして近年注目を集めているコンフィデンシャルコンピューティング技術の一つである。これらの技術によって、クラウド利用者は、クラウド事業者が信頼のおけるものでなくとも、Intelなどのシリコンベンダーのみを信頼するだけで高いセキュリティーの恩恵を受けるといった世界観になりつつある。

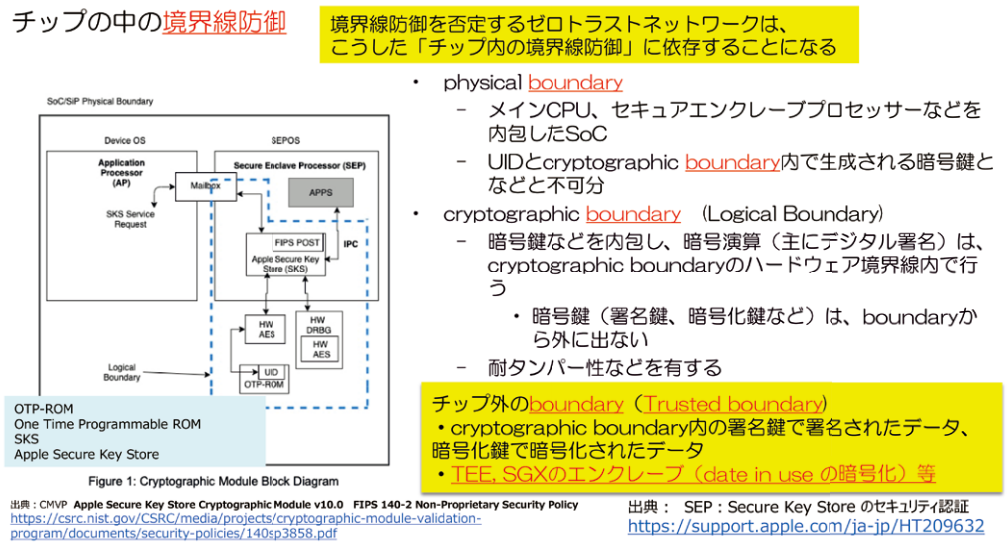


**図2-8-7　　チップ内外に構築されるさまざまな境界防御**

**まとめ**

Society 5.0時代においては膨大な数のデバイスが登場すると考えられ、そのためには、スケーラビリティーのあるトラストの仕組みが求められるが、これは、過去にはなかったトラストへの要求となる。このような要求に対応したマシン対マシンのトラストを形成するための技術を今回紹介した。これらは、セキュリティー技術でもあるが、サイバー攻撃への対抗というよりは、ビジネスを原動力として発展してきたような向きがある。

膨大な数のスマートフォンに組み込まれるTEEや、コンフィデンシャルコンピューティングのようなTrusted Boundaryを自在に組めるような仕組みが、Society 5.0時代のトラストを形成していくのではないだろうか。

**【主な質疑応答】**

Q：セキュリティーにおける認証では、お互いに認証し合うものであるが、トラストもお互いにトラストするような手続きがあると考えてよいか。

A：実際には、トラストは相互に行うものである。今回は、話を単純化するためにRelying Partyから

Trusted Partyに対する一方向の見方で説明を行ったが、ゼロトラストということに関して言えば、トラストできない環境（ゼロトラスト環境）に置かれたTrusted Partyとなるデバイスを、いかにRelying Partyからトラストできるようになるかが重要になるため、このような説明を行った。

Q：リモートアテステーションはいろいろなユースケースがあるようだが、どれも同じような仕組みで検証されるのか。

A：アテステーションが実現できるコアなデバイス（セキュアエンクレーブやTEEの搭載されたデバイス）が実現できれば、どのようなユースケース／デバイスであっても、実現できる。ただし、膨大な数のデバイスをアテステーションで管理することはまだ難しく、それを可能にしようとしているのが、現在の標準化の動きでもある。

Q：信頼の基点やトラストチェーンにおいて、現在の課題はどこにあるか。

A：標準化があまり進んでいない。標準化は、多くのステークホルダーが登場するSociety 5.0時代には非常に重要だが、例えばTEEやエンクレーブは、CPUのアーキテクチャーに依存する部分が多いため、標準化が単純ではない。また、このようなトラストの構築に熱心なのは、プラットフォーマーとシリコンベンダーであり、日本がそのような海外企業だけに依存してよいかという問題はある。

Q：アテステーションを行うには、PKIにおけるCAのように、外部の信頼できる機関や情報が必要か。

A：一般的には必要であり、あった方が仕組みはシンプルになる。アテステーションにはさまざまなトラストモデルがあるが、IETFはパスポートモデルとバックエンドチェックモデルという2つのモデルを示している。パスポートモデルはビザのようなもので、本人ないしAttesterが、Verifierとなる大使館で取得するビザを取得し、本人が、Relying Partyとなる海外の入国管理にアテステーション結果であるビザを提示するというような流れで行う。

C：ビザを取得する際、何か別のものと照合してビザを発行するわけで、その情報を実はシリコンベンダーが握ってしまうことになる。何が正しいかという情報そのものをシリコンベンダーが握り、彼らがOKと見なせるものしか使えなくなっていく（自由さが奪われる）ことは危ういと感じる。

C：そのような方向に進んでいるのは間違いない。日本では、信頼の基点をベースにTEEを作るといった、基盤となるハードウェア、ソフトウェアの研究開発を担えるような人材がいなくなっていると感じている。

Q：今回は基本的にデバイス認証（Authentication）の話が中心だったが、認証だけでなく中身をどうトラストするかという視点もある。それについてはどう考えているか。

A：まずは、リモートに存在するものであれば、認証・識別できないものはトラストに値しないことが前提となる。その上で例えば、従来計測データのトラストを確保するものとして、計測機器の紙ベースの校正証明書が存在する。これを、計測機器のキャリブレーション（校正）が施されたものかどうかが、欧州のeIDASにおけるeSealなどの利用により信頼のおける組織が発行したデジタル校正証明書によって保証され、これによって計測結果の中身の正しさを保証されるようになればよい。こうしたことによりデジタル証明がなされたCPS（Cyber–Physical System）が実現できる。デバイスへのデジタル証明・署名が、能力、権限、資格のある署名者により行われ、このデジタル証明自体が、トラストチェーンにより検証できることが重要になる。

Q：トラストの俯瞰をする際に、重要な軸や考え方はあるか。

A：Trustworthinessの話なのか、トラストの話なのかは区別して考える必要がある。また、IoTの文脈ではReliabilityやDependabilityという言葉も出てくるが、Trustworthinessは、Relying Partyから期待されている性質であって、その中にはReliabilityもあるし、Dependabilityも入ると考えられる。また、先の法定計量の計測器であれば、法定計量で要求される校正のトレーサビリティーもTrustworthinessに含まれる。

Q：我々はAppleやIntel、Nvidiaなどのシリコンベンダーを信頼することが前提になるのか。もしそうであれば、シリコンベンダーはどのようなTrustworthinessを我々に提供してくれるのか。また我々がシ

リコンベンダーに対してできることはあるか。

A：シリコンベンダーを信頼しないと生活できない世界になってしまっている。欧州はそのような支配を嫌っていて、規制をかけるなど、自分たちで対応する方向に進もうとしている。

## 2.9 佐古 和恵[24]「暗号プロトコルとトラスト」

**暗号研究者の観点から見た一般的なトラストについて**

「AliceがTrustor、BobがTrusteeで、AliceがBobをトラストするかしないか」というトラストのモデル（図2–9–1）において、私が考えたいのは、AliceがBobの提供するITサービスを使うかどうかという意思決定をするシチュエーションである。 Bobがトラストできれば使う、トラストできなければ使わないことを想定している。従って、トラストするかしないかより、使うという意思決定をするかどうかがポイントである。なお、「Bobがどのくらいの Aliceのトラストに値するかがTrustworthinessである」と私は理解している。

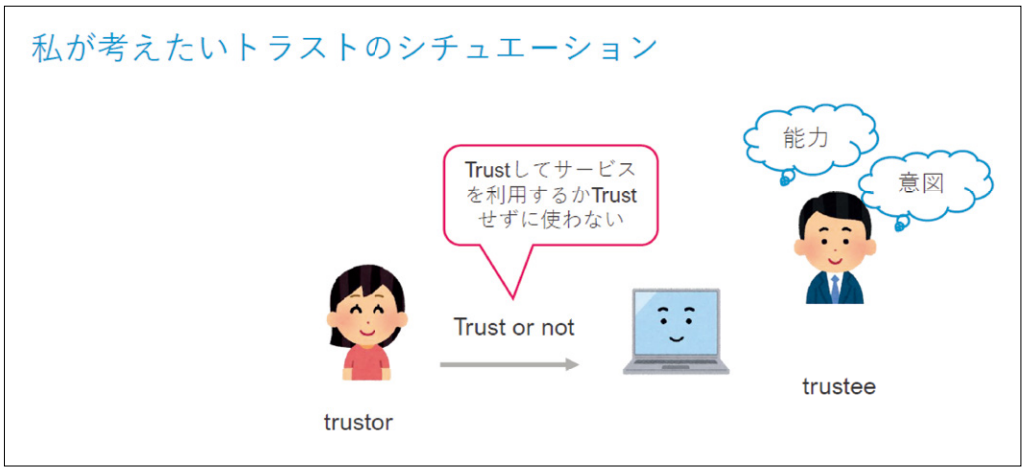


**図2–9–1　　トラストのシチュエーション**

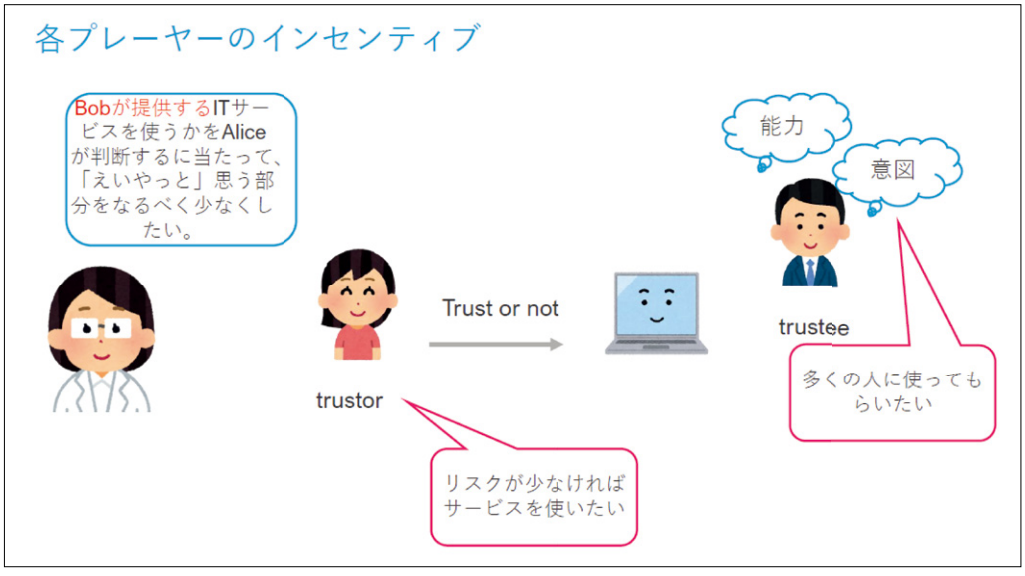


**図2–9–2　　各プレーヤーのインセンティブ**

24　早稲田大学基幹理工学部教授
https://sako-lab.jp/faculty.html

Bobが提供しているITサービスの内部の仕組みやアルゴリズムに対して、Aliceから見ると、全部Bobが責任を持っていると考えるのが妥当であろう。「AIを信用する、信用しない」というような、AIが人格を持っているような（コンピューターの擬人化的な）表現を使うのではなく、AIはAIのプログラムを作った人の意思が表れているものであると私は思う。従って、Aliceは、Bobが提供するITサービスを、Bobを信頼して使うか使わないかという判断を（Bobの「能力」や「意図」を見て）行うと考える。 Aliceは、まずITサービスを使いたいという気持ちがあってITサービスに対峙し、「リスクが少なければ、サービスを使いたい」と思っているのに対して、Bobには、多くの人に使ってもらって利益を得たいというインセンティブがある（図2–9–2）。 Bobが提供するITサービスを使うかをAliceが判断するに当たって、「いろいろ不安があるけど、でも、えいやっ、これ使っちゃえ」と思って「意を決して使う判断をする」ときのギャップがトラストであり、ここのギャップをなるべく少なくしたいと私は考えている。



**トラストを言語化する試み**

「情報セキュリティーに深く関連する言葉としてトラストがしばしば使われるが、この言葉は多義的であり、使われ方によって解釈が異なるため、一見つじつまが合っているようでいて、実際には話が合っていないことがある」ことを鑑み、コンピュータセキュリティシンポジウム2019（CSS 2019）で木村・島岡・菅野・佐古の連名で「トラストの言語化を試みる ～トラストとセキュリティ技術の関係～」を発表した。本論文執筆中も4者4様のトラストに対する捉え方であったことは興味深かった。

ITサービスシステムの仕組みは複雑である。基盤の上にミドルウェアがあり、その上でアプリケーションが動く階層的なITサービスシステムで、それぞれに別に開発している場合、連携されたものが安心・安全だと言い切れるだろうか。結局のところ、Bobが自分で設計したのはシステムの一部分であり、その下の部品（サブ機能）が悪さをして、Aliceに対して安全ではないITサービスを提供してしまう可能性はある。そのような状況を考えて、AliceはBobだけを見て、どう信頼を考えればよいのか、を上記論文で議論した。一つは、Bobは、その下のサブ機能を開発した他人を信頼し、AliceはそのようなBob（信頼した人のサブ機能を活用している）を信頼してサービスを提供していると考えられると思い、これを「内包モデル」と呼称した。別の形態では、Bobにお墨つきを与える偉い人がいて、Aliceはそのお墨つきを与える人を信頼するから、間接的にBobのサービスを信頼している場合が考えられ、これを「推移モデル」と呼称した。このようにさまざまな信頼の形について議論した結果を、CSS 2019で発表した。

**ビットコイン**

「誰も、誰にも邪魔されることなく事前に定められたルールが執行される」支払いシステムがビットコインである（図2–9–3）。サトシ・ナカモトが作ったルール（ポリシー）に基づいて、暗号技術を使うことで、ルールから逸脱できないように作られているところに信頼が置かれている。お金を発行する政府やマネートランスファーする銀行を信頼しなくても、自分のお金の価値が変わらず、また、確実に相手に送金できるので、ビットコインはトラストレスな仕組みであるといわれてきた。しかし、実際はいろいろな前提をトラストした上でビットコインの仕組みが成り立っていることから、「トラストレス」という表現は語弊がある（図2–9–4）。図中に記述しているビットコインがうまく回る前提が崩れてしまえば、安全ではなくなる。例えば、量子コンピューターが出現して暗号アルゴリズムが解かれてしまえば、未来永劫安全であるとは言えない。

また、ビットコインのルールはサトシ・ナカモトが決めたものだが、それが公平だと思ってみんな使っている。しかし、想定外の使い方をされたら公平ではないかもしれない。なので、このルールも公平だと「トラスト」されている状況である。また、正しくソフト実装されていることを「えいやっ」と信用しないといけないところがある。オープンソースになっているので、いろいろな人が見ていて多分大丈夫だろうと思っているが、本当にバグなくソフトが実装されているかは分からない。イーサリアムの方はバグが見つかり、もう一度ルールを変更したことも記憶に新しいかと思うが、正しくソフトが実装をされていることを信頼して使わないといけない。

また、ネットが正常稼働しているかというのもビットコインの安全性の前提としてある。インターネットの仕組みがハックされたことによって、自分が支払おうと思っていたビットコインが相手に届かなかったことがイロイロ起こった。また、サトシ・ナカモトは、本人が自分の鍵管理をしっかり実施する前提でこのシステムを作ったが、他人に任せた結果、そこから流出したという事件も枚挙にいとまない。このように信頼を揺るがすことがいくつか起こっているが、ビットコインは何となくみんなに受け入れられている状況である。

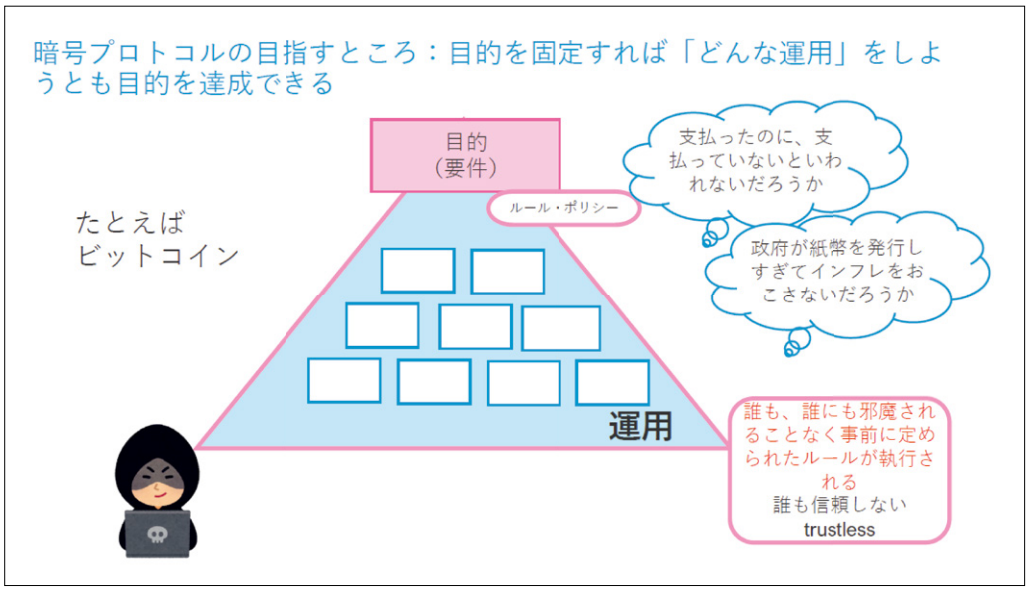


**図 2–9–3　　暗号プロトコルの目指すところ**

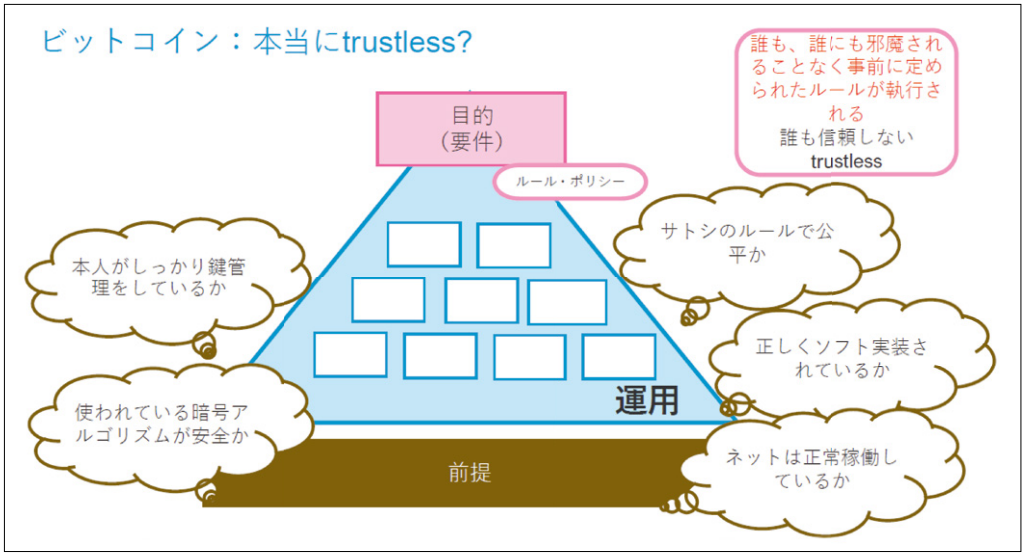


**図 2–9–4　　ビットコイン：本当に Trustless？**

**ベリファイ**

「えいやっ」と盲目的に信頼する部分を減らすには、「ここは大丈夫」と確認できる検証可能な部分が増えると良いと思う。 Bob がうまく暗号プロトコルを使って、事実を確認せずに信頼している部分（青い斜線の部分）がピンク色に染まる（検証可能な部分が増える）ような仕組みを IT サービスの中に入れることによってベリファイ（検証）できると良い。2021 年 3 月 12 日付で発表された Trusted Web 推進協議会の「Trusted

Web ホワイトペーパー ver1.0」の12ページの図にあるように、ベリファイできる部分を多くして、ITサービスの運用がおかしかったらAliceに分かるように暗号プロトコル技術などを使って広めていくことが、これからの社会に必要なITシステムの要件だと思う。

ブロックチェーンなどは、かなりの部分がベリファイアブル（検証可能）であるが、先ほど述べたように事実を確認せずに信頼している部分（青い部分）も残っていて、前提など「えいやっ」と信じないといけない（図2–9–5）。一方で、その部分をゼロにするためには、目的が限定されてしまうことや、システムを運用するのにコストがかかってしまうところがあるので、Trusted Web推進協議会では、検証可能な部分（ピンクの部分）をなるべく大きくしつつ、全体のコストを抑えた仕組みというのを模索していくべきと議論した。

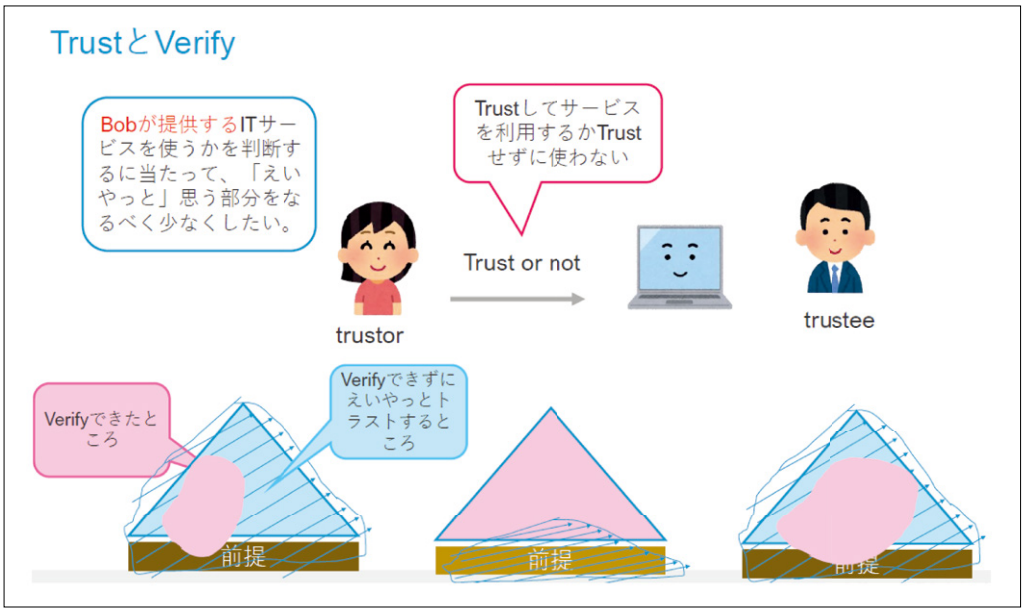


**図2–9–5　　TrustとVerify**

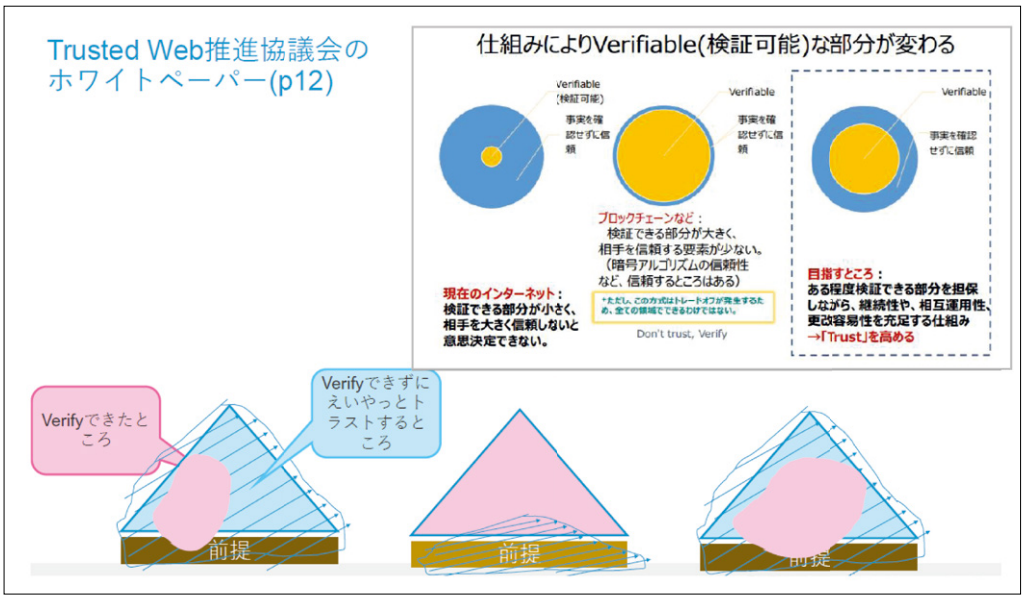


**図2–9–6　　Trusted Web ホワイトペーパー ver1.0（p.12）**

図2–9–6に示すように、「Trusted Web ホワイトペーパー ver1.0」では、トラストを「相手が期待したとおりに振る舞う度合い」と仮置きして議論をし、①ユーザーがデータへのアクセスをコントロールできる（Identifier）管理機能、②相手やデータに関する信頼を第三者によるレビューも含めて検証できる（Trustable Communication）機能、③双方の意思を反映した動的な合意形成（Dynamic Consent）機能、④その合意形成プロセスやその後の履行状況を検証できる（Trace）機能、の4つの機能をこれから考えていくことを提言するとともに、マルチステークホルダーによって、これらの機能をどうガバナンスしていくかということを考えるべきであるという提言もした。今年度は、バージョン2.0を作って、実際にどうやって、この機能を作り、マルチステークホルダーによるガバナンスができるのかというところに関するホワイトペーパーを作ろうとしているというのが、今の現在状況である。

**暗号プロトコル技術**

暗号プロトコル技術は、目的を固定すれば「どんな運用」をしようとも目的を達成できることを目指している。その目的の範囲であれば、ずさんな運用をしていても、それがおかしいということがばれるような（ある程度安全な）仕組みを入れることができるのが、暗号プロトコルである。例えば、「同時」が実現できないネット上で公平に勝者を決めることも、暗号アルゴリズムを使えば実現可能になる。それが公平だということをどのように暗号プロトコル研究者は説明しているのか？　それを以下に説明する。

例えば、AliceとBobが「じゃんけん」をするときに、まず神様（Trusted Third Party）がいる世界というのを考え、Aliceは自分の手を神様に伝えて、Bobも同じように自分の手を神様に伝えることを考える。神様は情報を漏らさないという全面的な信頼があり、なおかつ神様はAliceとBobの手から正しいじゃんけん結果を返すので、神様がいる世界では「じゃんけん」はできる。暗号プロトコルで目指すのは、そのような神様がいなかったときにも、神様がいたときと同じアウトプットが出てくるということを証明することによって、安全に「じゃんけん」ができることにある（図2–9–7）。

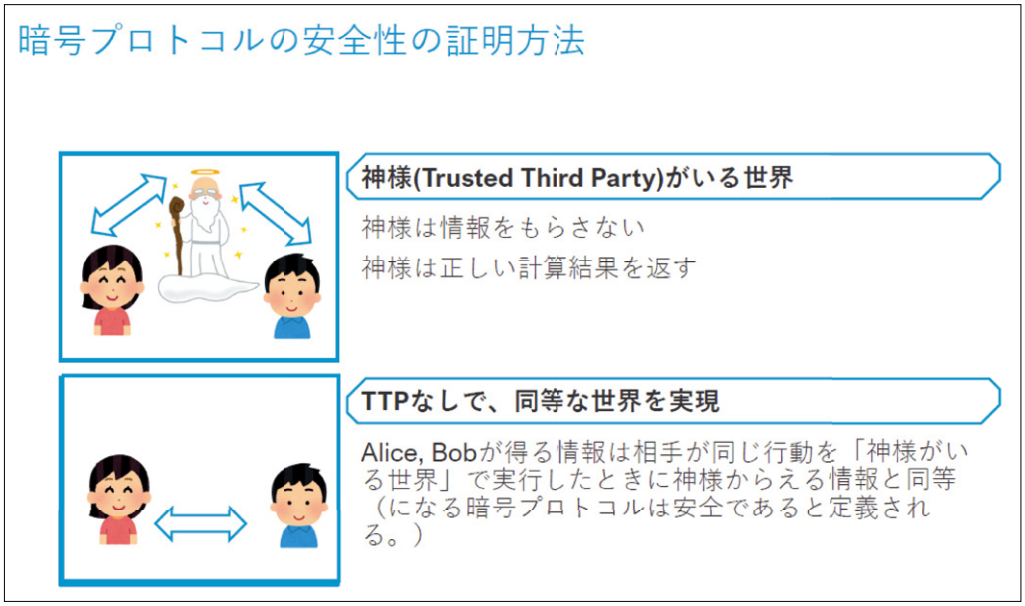


**図2–9–7　　暗号プロトコルの安全性の証明方法**

暗号プロトコル技術は、「じゃんけんプロトコル」や「電子投票プロトコル」などに代表されるように、ある目的を限定すれば、その目的がそのルールに従って実行されることをITの世界でエンフォースする技術である。いわゆる暗号化技術というのは、権利が付与されている人しかデータを読み出せないというようなルールをエ

ンフォースしており、デジタル署名技術では、その権利が付与されている人しかデータを書き込めないというルールになっている。また、暗号プロトコル技術は、正しく運営されていることを確認（Verify）できる機能を提供しており、正しい開票結果（電子入札プロトコル）や正しい通貨流通量増加（ビットコイン）を誰もが確認できて安心して使えるような仕組みを支援する技術である。しかしながら、暗号プロトコルの多くは、うまくいくための「前提」を置いていることが多く、ビットコインのように実装の問題も残っている。このように暗号プロトコル技術にも限界があることは事実なので、それを正しく一般の人に伝えていかなければならない。

**【主な質疑応答】**

Q：「じゃんけんプロトコル」には、最後に同時に復号鍵を送り合うという同時性を検証するための神様（Trusted Third Party）は必要ないか？

A：そこに神様は必要ない。復号鍵が同時じゃなくても後出しができないような暗号アルゴリズムというのを定義することによって、神様がいなくて同時性が担保されていなくても、最後の人が後出しできない仕組みになっている。

Q：佐古さんが考えているトラストは、ツール（さまざまなソフトウェア）をトラストできるかどうかという意味でのトラストだと思うが、ツールを使う人をトラストできるかは全く別問題である。ビットコインのシステムがトラストできるかどうかという話と、ビットコインが正常に運営されているかどうかという話は、マネーロンダリングに使っている、使っていないというような観点からすると、全く別の種類のトラストである。つまり「社会的な意味でのトラスト」と「ツールとしてのトラスト」というのは違う。ただ、一つ問題になるのは、AI的なシステムでは、必ずしも設計者が善意の意図で設計したことがトラストにつながらないのではないかという部分であり、論点だと思う。なぜなら、単独のAIシステムが非常に複雑であるということ以外に、実はAIシステムが社会の中に埋め込まれて使われると複数のAIシステム同士が相互作用し始めることが当然起こるからである。そのときに何が起きるかまで設計者が全部予測し切れないという問題があるので、やはりトラストが設計者の善意だけから外れてしまった崩れ方をしてくると思う。それに対する手段をどのように講ずるかという意味でのトラストは非常に難しい問題であるが、そこは技術系としてはどうなのか？

A：ツールとしてのビットコインでもマネーロンダリングに使われるというのはまさしくその通りである。従って、ビットコインは誰にも邪魔されずに支払いができるツールであって、そのように動いているということは皆に信用してもらえる。しかし、それを使って、どういう悪いことができるかというところまでのトラストは考えていない。

Q：多分、法律家の方がトラストの議論を始めると、そこの議論になって、どのレベルで妥当な解を社会的に見つけるかという議論になりそうな気がするので、そこは射程外ということか？

A：はい。

Q：技術的にも、実は複数のシステムが相互作用したときのトラストを作り切れるかという問題はどうか？

A：私は、設計者でさえ予見できないAIシステムを市場に出すべきではないという気持ちだ。もしもそれを市場に出すのであれば、そのシステムが複数のシステムと相互作用して起こった事案をログとして残して、その事案を調査する責任を取る覚悟があるのなら市場に出してもよいと思っている。

Q：それが設計者の理想的なポジションだと思うが、実は、なかなか現実はそうではない。例えば、AIトレーダーが引き起こした（株価などの相場が瞬間的に急落する）フラッシュクラッシュは、まさにそういうことに背くようなタイプの例である。現実的にものすごい経済的損失を引き起こしたということもあり、それどうするかというのは微妙な（企業秘密がある）問題であり難しいと思うが、どうか？

A：フラッシュクラッシュに関しては、そのようなAIトレーダーがいたことを前提として、本来どのようなルールで株取引をするべきなのかというところから考えなければならない。そのようなルールを検討するのが次の研究課題。

Q：結局、社会システム的なものを持ち込んでいかなければならない部分はどうしても出てくる。それともう一つ、将来的な問題としては、自動運転は果たして、設計者の理想的な立場だけで作り込めるか？

A：自動運転の方は、個人的には、安全性に懸念の残る自動運転車を世に出してほしくないところだが、人間は失敗を積み重ねて賢くなるということを考えると、サンドボックス的に、取りあえず被害が少なさそうなところから徐々に始めていくしかない…。

C：それはみんな考えており、Operational Design Domain（ODD）を決めて、自動運転がレベル3になれるエリアを限定して運用しようとしている。それにもかかわらず、そのような限定エリアと外側とのつなぎに関する問題が生じると非常に複雑な問題に発展するので、なかなか大変だと思う。

Q：イギリスのコンピューター科学者Roger Needhamらは「認証プロトコルのBeliefに基づき、True Beliefsというのは数を全部調べなければならない」と述べているが、BeliefとTrustはどういう関係になるのか、私は今もって明確に言うことはできないでいる（2.6節参照）。スライドの9（TrustorとTrusteeの関係）で、Trusteeのところに「能力」や「意図」が書かれており、トラストするかしないかは、Trusteeの「能力」や「意図」に対するTrustorが思っている期待値であると佐古さんは仰った。Trusteeの「能力」や「意図」を、自分の分かる範囲でのつかみどころのあるベリファイをすることによって、Beliefが思っていた通りだったとすると、True Beliefsだと分かってトラストするという感じを受けたが、いかがか？　また、Trustorが「えいやっと思う」部分というのは、そのBeliefが低いにもかかわらずトラストする決定をしてしまうことに対してベリファイすることによって、思っていた通りになると、もっとトラストするようになると考えればよいのではないか？

A：仰る通り。トラストするかしないかより、Aliceが「えいやっと思って使うか、使わないか」（どのように意思決定するか）というところを想定していた。確かに、Aliceの意思決定した結果がよかったら、次にBobのサービスを使うときには、かなり「えいやっと思う」部分が減っていると思う。

Q：ベリファイしているのは何か？ 健全な意図に基づいたサービスの使用に対してベリファイするというのは一つイメージとしてあると思うが、その意図とベリファイの関係を、もう少しお話しいただきたい。

A：何をもってベリファイできたとするかというところにも、またトラストがあると思っている。 Bobの方が信用してもらいたいから自分が提供するサービス中にベリファイできるような仕組みを盛り込んでおいて、Aliceはそれを確認していき、それが確認できたら、確かにBobが言った通りの機能が入っているに違いないとAliceは思えるようになるということを考えていた。

C：今日の話は、社会のトラストではなく、デジタルトラストの話に聞こえる。デジタルトラストには、ベリファイがやはりポイントになるような気がする。今までの世界でもベリファイはあるが、それは人間の目視によるベリファイで、そこがツールの話だと言ってしまえばそこまでだが、実はそこが大きく変貌していくと考える。そこでは、まさに（中川先生が表現したような）社会システムとしての制度設計も必要になる。ただ、社会のトラストとデジタルトラストとの違いが（今回で8回目の議論になっても、まだ）もやもやとしている…。

A：まさしく私が考えたいのは、ITツールやITサービスを使うときに、一般ユーザーがどう思うかというところであり、まず「デジタル」に限定して考えている。また、デジタルのITのところのトラストも、もともとの社会のトラストの上にあるものだと考えていて、そもそもこの社会が信用できない中ではデジタルトラストもない。ただ、実際の物理の世界は影響範囲がある程度限定されているのが、デジタルの世界ではアンプリフィケーション（増幅）がかかって、被害が甚大になり得る怖さがある。

C：そこは、私の関心事と同じだが、デジタル社会におけるトラストとは何か！？と、いまだにもやもやとしている。それは、きっとこれまでの社会のトラストの話でもあり、デジタルという属性がついたときのトラストの話がどうなっていくか！？といったところが、今回の俯瞰セミナーシリーズにおいて皆で共有すべき点であると理解している。

A：私も、まだ、もやもやしている。

C：意図やベリファイというところに関して、人間の側の結構高いスキル（あるいは相手に関する理解力）を要求している点が気になる。その点に関するリテラシーがない人間がAIを擬人化して考えるのはよくないという話は理解できる一方、ユーザーに対してヒューマンインターフェイス的な面から考えると、ある種、人のように見せることによって、システムのモデルをユーザーの側に思い起こさせやすい（想像しやすいようにしてしまう）テクニックとして使われている面もある。一般的なユーザーがシステムに対峙するときに親しみを持たせることで分かりやすくするという話と、意図を推測してベリファイもできるようなリテラシーをユーザーに求めるという話のギャップが、悩ましく感じた。

C：ユーザーとのやり取りをしていて、人にどう見せるか、それを見ることとか触れることによって、使う前にどういう反応が起こるかが想像できるようなことというのが結構トラストにつながるところの一面であると思うので、直前のコメントに結構同感できる。もちろんそれだけではないが、結局インターフェイスは、その人が使うだけのものではなくて、先ほど自動運転の話でも出たように、いろいろなところへの影響があるので、やはり（作る側としても想定し切れないが）できる限り作る側はもちろんのこと、使う側にとっても反応が想像できるようなものにしていくということは結構重要であると思う。

A：ユーザーインターフェイスでだまされてしまうのではないかという恐怖が私には常にある。本当は違うのにユーザーインターフェイスでだまされて信頼感を持ってしまうというのは、常に警戒したい。あと、ベリファイの話だが、つい先日、倫理資本主義を謳うドイツ人哲学者Markus Gabrielが、ルールに従っていることだけが確認できたら（意図がどうあっても）社会がうまく機能するようなルールを人間が英知を持って作るべきだ、とある対談で語っていた。そのようなことができたらすごいと思っている。相手の意図というのは多分ベリファイできないものなので、ルールの側でこれさえ従っていれば、どんな意図でもOKというものが作れたらいいと考えている。

C：AI分野におけるメカニズムデザインは、それと似た考え方のように思う。

Q：少し大きめのツールシステムを考えるにしても、サプライチェーンが長くなると思う。つまり基本ツールを作り、それを組み合わせ、さらにそれをユーザーインターフェイスまで持ち込むことで、全体的にシステムが複雑になればなるほど、サプライチェーンが長くなる。そうすると、各段階で、サプライチェーンの各インターフェイスの段階ごとにベリファイの条件を付けて一つ一つやっていないと、絶対に漏れが生じると思うが、それが果たしていろいろなケースでうまく機能するだろうか？ セキュリティー分野では非常に重要なポイントなので、おそらく機能するように設計されていると思うが、一般的なAIも含めたシステムについては、サプライチェーンの各段階をトラストできるようにベリファイしろ！とEUは言ってくることが心配であり、そこをどうするか！？というのがすごく心配になってくる。

A：本当にコストが高くなってしまう。本当に、そのコストを払ってでも、そのようなトラストが必要なのか、どこか良いバランス点はないのか、と思う。

Q：バランス点のところで、ある種の妥協が入っていく。やはり、そこにはトラストレスのトラストができない部分が若干入ってきているということが積み重なり、どこかでアウトになるような微妙な問題が起こる気がする。システムが大きくなればなるほど、そういうことがしばしば起きることは僕の少ない経験でもあるので、なかなか大変である。だからこそ、何かうまい解決策なり、指針なりが出せると嬉しいが、アイデアは僕自身もあまりない…。

A：先日のMyData Japanの会議でも議論に上がったのだが、そのようなサプライチェーンを管理する高いコストは強い巨大企業しか払えないので、強い巨大企業が一人勝ち（Winner takes all）になってしまう。これは、悩ましい問題だ。

C：まさに、Google一人勝ちの世界を我々は作ろうとしているような感じの話にもなってくるので非常に難しい問題だと思いつつ、今回もやはりそういう問題がありそうだということに気がついたので、一言コメントした。

Q：コストというのは、お金ではかる場合もあるし、処理プロトコルの重さではかることもあるが、トラスト

のコストという問題をお聞きしたい。

A：ベリファイできるような機能を埋め込み確認するためのコストを一番に考えていた。（処理プロトコルの）重さについては考えていなかったが、遅ければ早くする仕組みをいれることにより、やはり開発費に一番跳ね返ってくると思う。確認ができる仕組みを入れることによって、ユーザーインターフェイスが分かりにくくなってきて、ITサービス提供者側の方でやらなければならないことが雪だるま式に増えていく。しかしながら、もしかしたらそれは過渡的なもので、人間の英知が入ればコストが抑えられて、トラストのための機能が組み込めるようになると期待したい。

Q：セキュリティーのコストはシステム全体の約5%までが限度であると（今にして思えば古きよきのどかな時代に）聞いたことがあるが、そのコストは、増えてきているのか？

A：はい。ITによって、その費用限界がどんどん小さくなっていくときに、差別化する部分はセキュリティーやトラストだと考えるので、他のところがどんどんコストが抑えられていく反面、そこの要素が増えていってもおかしくはない。

Q：それに関連して、「製造業的観点からはトラストはコストと考えられているかもしれないが、サービス業的観点からはトラストは運用コストを減らす要素でもあり、必ずしもコストではないかと思う」というコメントがあって、確かにそういう面はあると思う。また、「意図とベリファイの議論なども含め、客観的再現性のある評価プロセスであるベリファイと、主観的な評価プロセスであるバリデートを使い分けると、より具体的な議論ができるかと思う」というコメントや、「トラストの曖昧性にどう向き合っていくかという意味では、図2–9–5の右下の図で、ベリファイアブルな部分とバリデートしかできない部分を明確にすることが重要で、ここに取り組むのがまさにセキュリティー分野であると思う」というコメントがある。これらの点に関しては、どうか？

A：仰る通り。主観的な評価プロセスへのトラストというか、その評価プロセスでいいのかどうなのかというところも、また議論が必要になってくる。多分それはルールメイキングの一環として考えていくことになる。また、ベリファイアブルな部分と、バリデートしかできない部分を明確にすることも心に留めていきたい。

Q：図2–9–7「暗号プロトコルの安全性の証明方法」のAliceとBobは、図2–9–2「各プレーヤーのインセンティブ」などで出てきたAliceとBobとは、違うタイプである。図2–9–7のAliceとBobは、ある程度セキュリティーが分かっている人であり、全体のシステムに対するビリーフが（一般ユーザーよりもはるかに）強い。このようなAliceとBobの描かれ方の違いが、技術系のトラストと、社会科学・心理学などでいわれるトラストとの食い違う部分ではないか！？と思った。

A：仰る通り。

## 2.10 山田 誠二[25]「ヒューマンエージェントインタラクションと信頼工学」

**ヒューマンエージェントインタラクションについて**

HCI（Human–Computer Interaction）は長い歴史を持っているが、我々は新たに「人間とインタラクションを持つ対象をエージェントに限定」した日本発の学問分野を立ち上げ、HAI（Human–Agent Interaction）に関する研究を精力的に続けている。 HAIに関する国際会議も今年で9回目の開催となり、世界的にも認知された研究分野に展開しつつある。

従来のエージェントの定義は「自律的に行動する認知主体」とされる場合が多いが、HAIでのエージェントの定義は「人間が人間のように感じる人工物の総称」である。要するに、擬人化というプロパティーを持っている人工物を広くエージェントと呼ぶ。簡単に言うと、日頃使っているボールペンに非常に人間らしさを感じるなら、そのボールペンはエージェントという概念に入る。 HAIで扱うインタラクションは、（1）人間−擬人化エージェントインタラクション、（2）人間−ロボットインタラクション、（3）エージェントを介した人間−人間インタラクションの3つがある。なお、インタラクションとは、2つのエンティティー間でやり取りされる全ての情報のことを指す。それは物理的な情報であってもいいし、ビジュアルな情報でもいいし、さまざまなモダリティーを持っている情報全てを指す。例えば、外見は光を通して情報が伝わるので、一つの重要なインタラクションである。

HAI研究の目的は人間とエージェントの間のインタラクションデザイン技術を開発することである。最終的には、人間とエージェントがうまく付き合う（共生する、協働する）ために、エージェントが人の良きパートナーになるようなデザインの方法論を開発することを目指している。インタラクションデザインを大きく分けると、①エージェント自身のデザイン、②人間とエージェント間でやり取りされる情報のデザイン、③人間とエージェントの関係のデザイン、の3つがある。上記①としては、アピアランス（外見、身体）、それから表出される情報の表現（人間に伝えるべき情報、情報の表現）のデザインがあり、説明可能AI（XAI: Explainable AI）やユーザーインターフェイス（UI: User Interface）に関連した部分である。また、上記②としては、エージェント自身がどういうような機能を持つべきか、人間が理解しやすい（適応しやすい）エージェントの学習アルゴリズムをどう開発するかというところに主眼を置いている。そして、上記③としては、人間とエージェントがどのような関係を持つのがいいのかという研究（協調タスク、ゲーム・エンターテーメント、癒し・親和性、ゲーミフィケーション、ヒューマンコンピュテーション）が関係してくる。

HAIは非常に学際的な研究分野であり、統一理論的なデザインの方法論はまだない。よく使われるのは、人工知能、特に機械学習であり、少ない訓練例からの学習、能動学習、トランスダクティブ学習、インタラクティブ機械学習である。そして、発達心理学、社会心理学が非常に関係しており、心の理論（Theory of Mind）、情報伝染（Emotional Contagion）がよく使われる。あとは、心理学、哲学であれば、Media Equation、Daniel Dennettの3つのスタンス（Intentional/Design/Physical Stance）がよく援用されている。 HAIはあくまでも工学的なアプローチなので、社会科学で使われるような説明原理を追求するのではなく、生成原理を目指すことになる。これはデザインの方法論を開発したいというところにも表れている。

**信頼工学**

信頼工学は元の英語はTrust Engineeringであり、主にヒューマンファクターやエルゴノミクスなどに関する人間工学分野で数十年前から研究されている。山岸俊男が世界的にTrustの信頼の研究で有名であり、社

25 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所教授
http://www.ymd.nii.ac.jp/lab/seiji/

会科学におけるTrustの定義は、山岸の定義と非常に近いものになっている。簡単に言えば、「相手に対する期待値」あるいは「相手の能力の推定値」を「信頼」と呼んでいる。図2–10–1は、2019年のthe Human Factors and Ergonomics Society Annual Meetingにおいて、Trust Engineeringの位置付けが議論された際に使われた図である。Trust Engineeringは非常に多岐に渡っており、細かい分野に分かれているが、一般的に考えられているTrustとは少し趣が異なる定義になっている。

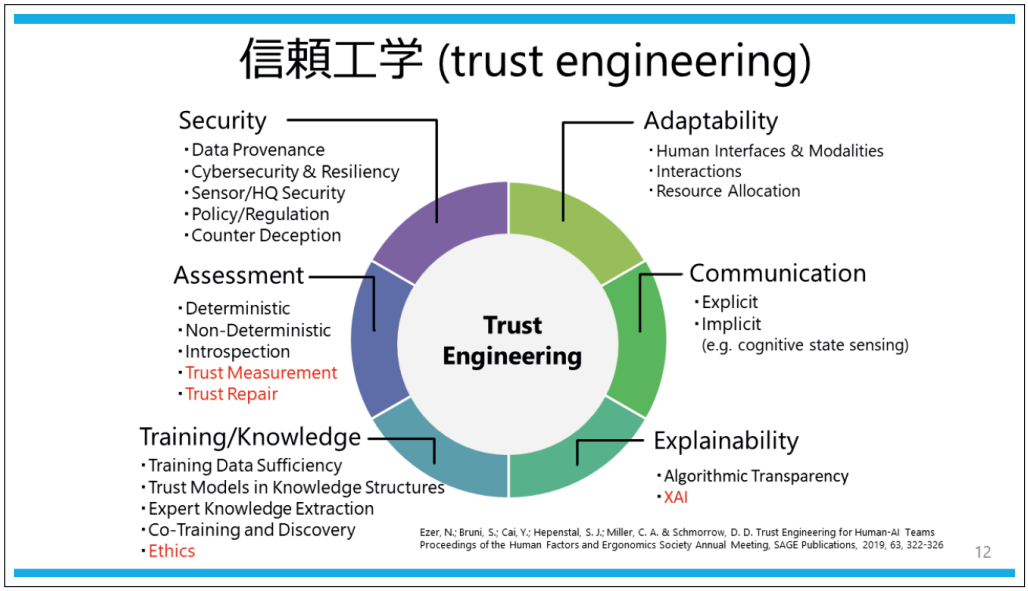


**図2–10–1　　信頼工学**

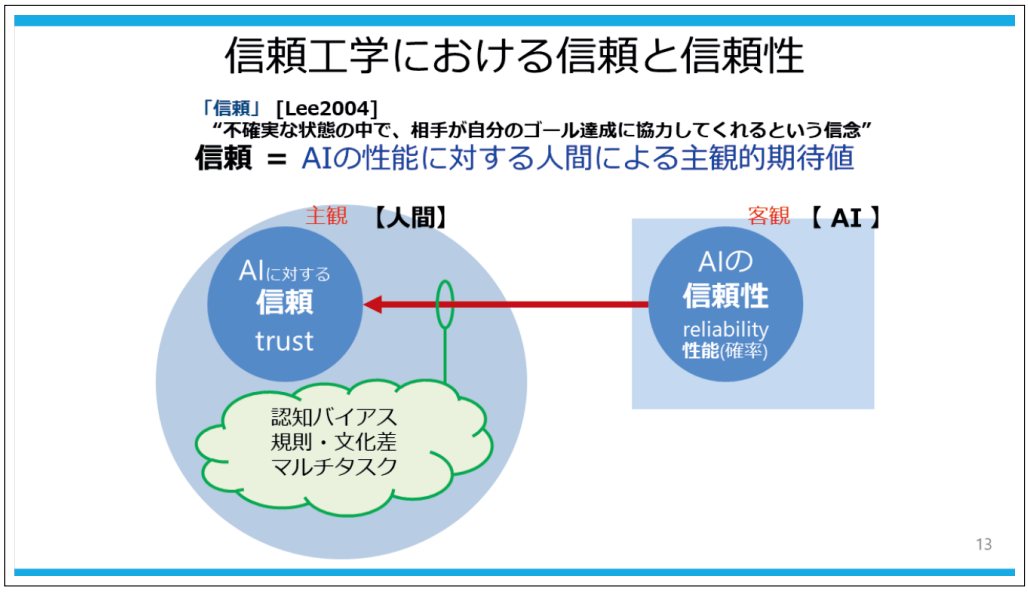


**図2–10–2　　信頼工学における信頼と信頼性**

信頼工学における信頼の定義を説明する。図2–10–2（右）はシステムがどれぐらいの性能を示すかという意味での信頼性（Reliability）であり、信頼工学が扱う信頼性のベースとなるものである。このようなReliabilityはAIにもあり、性能（あるいは、あるタスクをAIに実行させた場合の成功確率）で定量化できる。つまり、AIの性能に対する人間による主観的期待値を信頼（Trust）と定義し、この確率と、タスクがうまく

いったときのユーティリティーの掛け算（確率×ユーティリティー）で計算できるという立場を取る。ただし、この性能は人間の主観で解釈されるので、図2–10–2（左）に示すように、認知バイアス、レギュレーション・文化差、マルチタスクの影響などのいろいろな要因が絡み、少し主観が入った（ある種のバイアスがかかった）推定値が出てくる。従って、AIの信頼性（Reliability）は客観的なものだが、AIに対する信頼（Trust）は主観的なものであると言える。

**Trustへ影響する要因**

Trustへ影響する要因としては、多面的な要因があるといわれている。人（Culture, Personality, Knowledge, Self–Confidence）・システム（Reliability, System Failure, Predictability）・環境（Task, Risk, Workload, Rule）の3つのカテゴリーに分類される要因、および性能（Reliability, Predictability）・プロセス（Algorithm, Operation）・目的（Goal, Designer's Intension）の3次元に関する要因がある。例えば、ロボットやAIを見たときに非常に拒否反応を示すようなネガティブな先入観を持っている一般人は数多くいて、パーソナリティーの一つとなって、非常にバイアスがかかってくる。簡単に言うと、AIが、客観的に見て高い性能（つまり高い信頼性）を持っているにもかかわらず、主観的な評価をした信頼という意味では過小に評価されてしまうような状況には、パーソナリティーが影響している。

信頼性がバイアスを超えて正しく人間に伝われば、信頼と信頼性は等しくなる。そのような望ましい状態を最適信頼と我々は呼ぶ。図2–10–3に示すように、信頼性にバイアスがかかって過大評価されている状態が過信（オーバートラスト）といわれる状態、過小評価される場合が不信（アンダートラスト）といわれる状態である。過信・不信の状況においては、これは人間とAIを含んだシステム全体のパフォーマンスが悪くなることが原理的に分かっているので、信頼性を正しく人間に伝えるために、バイアスをできる範囲で排除するのが重要である。それを行うためには、システムのReliabilityをいかに人間に的確に伝えるかという、いわゆる「透明性」の問題になり、重要な役割を果たすUIを工夫する研究が精力的に行われている。

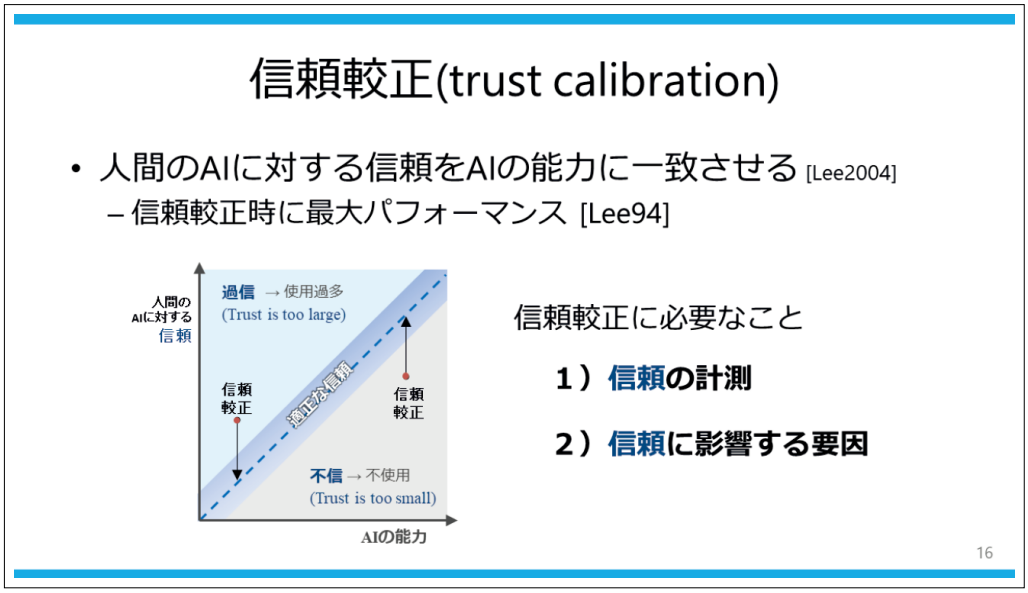


**図2–10–3　　信頼較正（Trust Calibration）**

**信頼較正（Trust Calibration）**

過信・不信という望ましくない状態から最適な信頼（あるいは適正な信頼）に戻してやることが大事であり、これを信頼較正（Trust Calibration）と呼ぶ（図2–10–3）。ただし、実際にどうやって信頼較正を実現す

るかという方法論は、それほどまだ出されていないので、我々は適応的信頼較正（Adaptive Trust Calibration）というフレームワークの提案を行い、ここ2、3年研究している。図2–10–3の縦軸がAIに対する信頼、横軸がAIの能力（AIのReliability）の方の信頼性であり、45度の線がAIの本当の能力と人間の主観的な期待値が一致するところ（正しい適正な最適信頼）を表す。

過信・不信という望ましくない状態では全体のパフォーマンスが低下することが分かっているので、最適信頼のところに状態を持っていきたい。しかしながら、あくまでも信頼というのは主観的なものなので、結局人間が納得して自分で適正な信頼に自分で修正してもらう必要がある。つまり、信頼較正がうまくいくかどうかは、最終的には人間にかかっている。そのためには、信頼が今どういうような状況にあるかを計測して、かつ、それに対して信頼をある種の刺激としてのパータベーション（Perturbation）を与えて、望ましい方向に人間を誘導する作業が必要になる。

信頼の計測に関する先行研究では、直接計測は困難なので、自己申告アンケート・生体信号・人間の行為・モデルベース推定などの研究があるが、安定性に欠ける。そこで、我々は、人間、AI、どっちがやるかという選択行動を繰り返すというフレームワークで、その選択行動の履歴から過信・不信を判定する方法を提案している。また、あるモデルから演繹的にトップダウン的に計算するようなモデルベースアプローチも行っている。

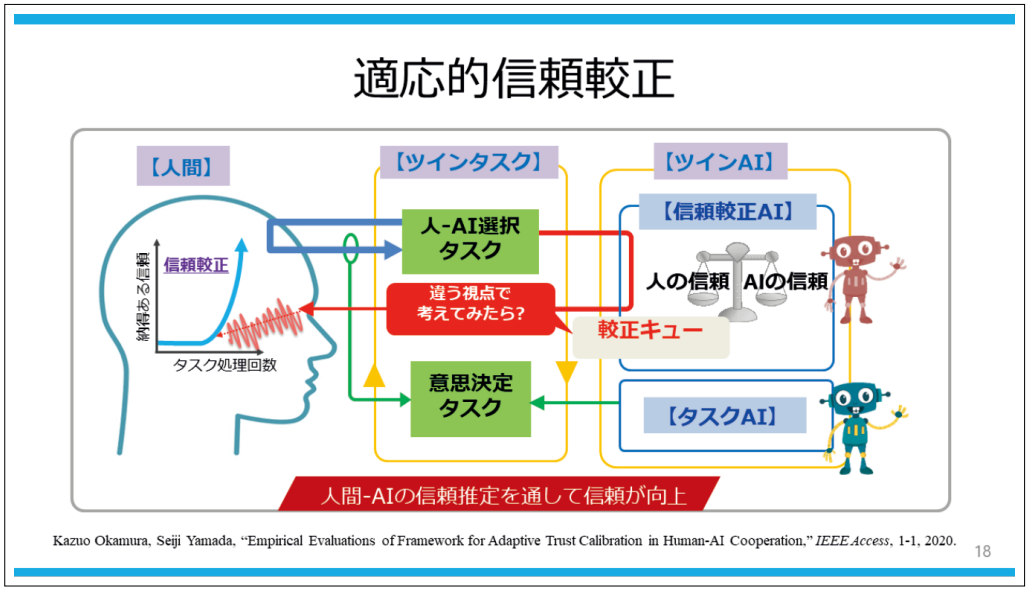


**図2–10–4　　適応的信頼較正（Adaptive Trust Calibration）**

**適応的信頼較正（Adaptive Trust Calibration）**

図2–10–4に、我々が提案している適応的信頼較正のフレームワークを簡略化したものを示す。まずタスクとしては、ある人間とAIがどちらでもできるタスクを繰り返してやるというフレームワークにしてある。実際にタスク自体を実行するタスクAIというのを用意しておいて、信頼較正AIと実際にタスクを遂行するタスクAIの2つのツインAIの構成になっている。応用先としては、人間とAIの協調画像診断や協調健診をターゲットにしている。人間は「エックス線の画像の読影を、AIに任せるより自分でやった方が正しいと思ったら自分でやるし、AIの方が正しそうだと思ったらAIに任せる」という選択行動を繰り返してループで何回か行う。そして、その選択行動を信頼較正AI（過信か不信かを判定するAI）がモニターしていて、人間のAIに対する信頼をモデルベースで計算して、過信・不信をチェックしている。もし過信や不信が検出された場合には、信頼較正キュー（TCC: Trust Calibration Cue）と呼ぶ刺激を与えて、人間自身に信頼較正してもらう。

**HAIと信頼工学、信頼較正の関係について**

信頼性を人間にいかに伝達するかというのが、信頼工学におけるメイントピックの一つである。モダリティーの工夫に、HAIのさまざまな知見が使えるのではないかと考えている。

実際、我々の“Adaptive Trust Calibration for Human-AI Collaboration”と題するPLOS ONEに2020年に掲載された論文では、3Dドローンシミュレーターを使用したオンライン実験において4つのモダリティーの異なる信頼較正キュー（①Visual-sign TCC、②Audible TCC、③Verbal TCC、④Anthropomorphic TCC）を評価した結果、単純な手がかりを適応的に提示することで過信頼時の信頼の調整を大幅に促進できることを示した。図2-10-5に示した4つの信頼較正キューを簡単に説明すると、①逆三角形の赤い警告サインを提示する「ビジュアルサインTCC」、②音による人工的な微妙な表現（400 Hzから250 Hzへ下げる）で、エージェントの信頼度を伝えることができる「聞こえるTCC」、③マンガのような顔のパーツを使ったドローンのアニメーションによりエージェントの状態を表す「言葉によるTCC」、④YESボタンが選択されたときにツールチップとして警告のテキストメッセージを表示する「擬人化TCC」である。これら4つの比較実験を行った結果では、4つともそれぞれ効果はあるが、統計検定的に一番効果があったのは（我々が期待した「擬人化TCC」ではなく）「言葉によるTCC」であった。今後、擬人化エージェントとHAIが絡んでくるような「擬人化TCC」をもう少し突き詰めて性能を向上させ、HAI的にロボットや擬人化エージェントで信頼較正キューを実装したいと考えている。

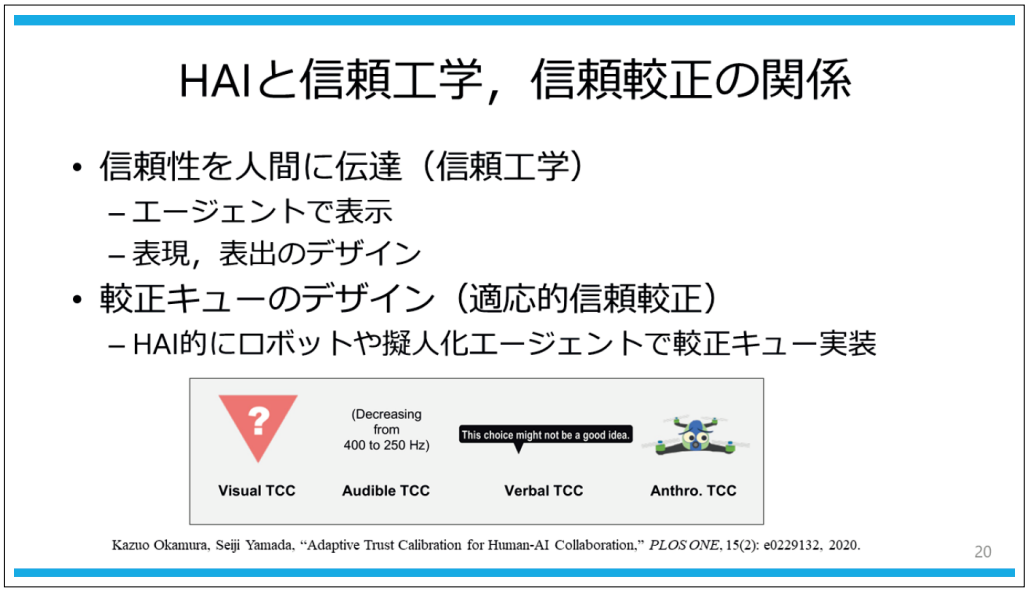


**図2-10-5　　HAIと信頼工学、信頼較正の関係**

**【主な質疑応答】**

Q：AIは必ずしも人間に対して善意で動いているとは言えない。特に、セキュリティー周りでAIを使い始めると、必ず敵対的な人もAIを使ってくるという問題も出てくる。従って、善意であるか悪意であるかを見分ける技術が必要にならないか？

A：仰る通りだが、個人的には、AIに対して善意・悪意という概念はあまり適用したくない。 AIはあくまでも人間がデザインした通りに動いているだけであって、もし悪意が存在するとすれば、それはデザインした人間の中に存在していると思う。

Q：例えば、悪いボットを作った敵対的な人の中に存在している？

A：はい。そういう意味では、我々の信頼較正や、過信・不信を判定するようなAIが悪意を持って作られ

ていると、過信・不信の判定がわざと操作されて、逆転したりおかしくなったりしてしまう…。

Q：例えば、ナッジ（Nudge）なども関係してくると思うが…。

A：なるほど。実はナッジについては、ここの信頼較正キューをナッジの考え方で作ろうという考えもある。我々が考えている今のフレームワークでは、最終判断を人間が自分の意思で行うことは入っているが、信頼較正AI自体が悪意を持って間違った判定をするように作られていることへの対策として、それを判定するようなメタなAIは入っていない。

Q：そのようなメタなAIが必要ではないか？

A：それを言い出すと、さらにそのメタなAIが必要になってくる。

Q：倫理後退（Ethical Fading）せずに、どこかでうまくストップできるのではないか？

A：人間には、行動経済学でいわれているような自制の参照点（損得という心理的な感覚の評価を主観的に設定した基準で判断している点）があり、限られた認知能力と限られた時間的制約の中で損失を回避するような行動を取るといわれている。従って、一つの考えとしては、もう一個メタなAIを作るのが解決方法だと思うが、今は考えていない。

Q：不信と過信を分けるというのは、すごくきれいな分け方でよく分かる。ただ、過信の方が実はもう少し構造が複雑であり、2つのタイプがあると思う。一つはAIに対する思い入れのような（例えば、ペットロボットAIBOに感情移入してしまう）過信のタイプ。もう一つは、（AIにお任せする）思考停止のタイプ。ともに、難しい問題であると思うが、今は思考停止に関してコメントする。人間が思考停止するというのは非常に自然な流れであって、いつも怪しいと思いながら行動するのは非常に疲れるので思考停止したがる。要するに、人間は脳を無駄に使わずに、他の方にエネルギーを振り向けようとするのが自然の流れである。これは社会の効率化などにもトータルに見て役立っているが、どの辺でユーザーにサジェスチョンするかというのは、大きなフレームワークで捉えると難しい問題になってくる。

A：思考停止という考えは、人間のAIに対する信頼に及ぼすファクターとして、我々が（従来の信頼工学の研究から）メタ解析的に考え出したものだ。例えば、車を運転するときに音楽を聴くようなセカンダリータスクをやることが頻繁にあるが、もしプライマリータスク（車の運転）がAIに任せられたときには、人間はセカンダリータスクに集中できる。従って、他のマルチタスクの状況において、運転以外のタスクというセカンダリー（あるいはサードの）タスクが、どれぐらいのユーティリティーを持っているかを定量化することで、思考停止（任せっきり）をモデル化できると考えている。

Q：なるほど、よく分かる。ところで、ReliabilityとTrustの関係に加えて、EUなどでいわれているRobustは、一体どういう関係にあるのだろうか。 Reliabilityはテストコンテキストで動くものであるのに対して、おそらくRobustはマルチコンテキストで動くReliabilityのことだろう。ただ、Trustというのはこの両者に対して一体どういうような位置関係にあるのか。私はまだよく分かっていない。この点に関して、どう考えるか？

A：Robustの方もマルチタスクが関係していると思う。

Q：AIの性能の真値とは何か？それが真かどうかというのを私たちはどう信頼したらいいのか？

A：これは何か自然現象で起こったことを観測しているわけではなく、ある計算アルゴリズムで出てきた値である。従って、これが真値だと言われればそれを信じるしかないが、それを疑うということもあり得る。

Q：しかし、その計算アルゴリズムで出てきた値に沿うように較正されてしまうのでは？

A：仰る通り。 AIの性能の真値に関する研究は、いまだに続いている。 Ground Truthに対してどれぐらい正しく答えを出せるかという性能を決めている裏には人間が持っているGround Truthがある。

Q：しかし、人間が持っているGround Truthを、逆にAI性能真値に較正するところにすごく違和感がある…。

A：信頼較正で言っているのは、Ground Truthを較正しているのではない。

Q：信頼較正をするかしないかの最終判断は人間が決めるのか？

A ：はい。

Q：実は較正しなくてもいいとか、何%較正した方がいいという、シチュエーションアウェアネス（Situation Awareness）という状況把握する能力が人間に備わっているのではないかと考えており、研究を始めたところだ。ところで、AIに対する信頼を考えるにあたって、信頼モデルの中でメンタルモデル（人間が無自覚のうちに持っている、思い込みや価値観）はどこかにあるのか？

A ：信頼モデル全体がメンタルモデルであり、人間がAIに対する信頼をどう構築するかというモデルであると期待している。

Q：人間は無意識のうちに機械のようなAIではなく生身の人間でないと信頼できないように、そのような説明しにくい感情的なものが背後に強くあって、そこを何とかしない限り、なかなかAIに対する信頼を向上させる（あるいは不信を解消する）ことはできないような気がする。今回提案されている手法は、そこまで踏み込んでいる感じなのか？

A ：できればそれをやりたい。しかし、先入観がものすごく強い人が結構な数いる。そのような人々は「AIの出した答えは絶対信用できない。 AIは単なる計算をやっているだけで人の気持ちを酌んでいないから、こんな計算結果なんか意味がない」と思い込んでいる。もちろん、それはある意味正しい部分もあるが、凝り固まった強い先入観をほぐして取り除くところまで信頼較正キューの効果を向上させられるかどうかは、今後の研究次第だと思う。現段階では、そこまでバイアスが強い人が、較正を自分で行うところまでいくのは無理だと思う。

Q：「Trustは主観的なもので、Reliabilityは客観的なものである」とした図2–10–2に関して、社会学的な観点から質問したい。医療を例にすると、患者によっては、自分が受ける治療方法がTrustworthyであると思えることに関して、信頼性の程度に違いがある。従って、Reliabilityには、主観的な要因もある程度入るのではないか？

A ：ここで言っているのは、そういうある程度の信頼性があればTrustworthyであるというような判断をしている信頼ではない。もともとシステムにReliabilityはあり、それを主観的に期待値で解釈したものがTrustであるという信頼工学の定義である。だから、対象がTrustworthy否かという判断は入っていない。要するに、このReliabilityの絶対値の高低は信頼較正には関係ない。高いなら高いで低いなら低いでどっちでもよくて、その値を正しく主観的に評価できているかの問題である。つまり、本当に低ければ自分でやればいいし、本当に高ければほとんどAIに任せればいい。そのようなことができれば、全体のパフォーマンスがベストになると考えている。

Q：AIから遠くになるほど多分信頼を構築することが難しいと思う。そして、患者・医者・市民・開発者のポジショナリティーにより、AIに関するリスクも変わってくる。

A ：そのようなポジショナリティーは、HAIの方ではソーシャリティーとかソーシャルリレーションシップと呼んでいるものである。しかし、ソーシャルリレーションシップとは人間社会、人間対人間の間のリレーションシップであるのに対して、AIは機械なので、AIと人間の間にはソーシャルリレーションシップが基本的にはないと考えている。

Q：AI技術への期待がリレーションシップに入るのではないか…。

A ：なるほど。信頼工学のこれまでの研究には導入されていない概念だ。

Q：今の議論を聞いて、ソーシャリティーやポジショナリティーの話はAI側にもあるのではないかと少し思った。要は、どのようなAIが情報提供しているのかによって、人間の理解、感覚、受ける理解が変わる。例えば、GoogleのAIスピーカーがレコメンドしたものであれば信頼するとか…。

A ：なるほど、デザイナーの方の信頼性はあると思う。仰るように、GoogleとかMicrosoftが出しているAIだから信用しようというバイアスは、かなりかかると思うので、検討してみたい。

Q：人間同士でもあることだが、雑談も含めたAIと人間の対話によって、相手（AI）の性能は全然変わっていないのに人間は親しみを持って信用するようになると思う。今回の信頼較正では、どう考えればよ

いか？

A：HAIでも、ファミリアリティーと言って、親和性を高めるために少し対話させたり、たまごっち系ゲームをやらせたりすることをよくやる。そうすることによりバイアスがプラスの方向に行くということは十分あり得るので、そのようなものを信頼較正キューにするのは十分あり得える。

Q：なるほど、キャラクターをデザインすることと会話で親しみを持たせることは確かに近い。

A：ただ、今は余分なタスクを極力控えたいので、（会話ではなく）1秒ぐらいで認識できるような刺激にとどめている。

Q：今回のモデルが適用できて効果がありそうなケースと、いろいろな複雑なものが絡んでいて簡単にはいかないケースがあると感じた。それらをタスクやシチュエーションなどで適切に分類するような研究テーマがありそうな気がした…。

A：今回のフレームワーク（図2–10–4）が使えるタスクは何なのか、タスクのクラスは何なのかという、タスク分析は行っている。 UML（Unified Modeling Language）を使って、繰り返しAIか人間かどちらかを選択するというタスク分析には適用できている。ただし、精度やファクターの洗い出しなどに関する研究がほとんど始まっていないので、これから頑張りたい。

Q：4つ信頼較正キュー（TCC）を試したところ「言葉によるTCC」が一番効果的であったという話だが、話している当事者にとっては話している相手を擬人化するようなことで物事が進んでいるのではないか。その擬人化された相手は、AIを作った人かもしれないし、TCCのデザインそのものかもしれない。会話の場合も、実は会話のテキストだけを見ているわけではなくて、自分の中で話す人を想像しているのが影響していると思うが、いかがか？

A：100%仰る通りだ。出された信頼較正キューに対して、ユーザーは、どこから発せられたものか（バックで人間が操っていてその人から出たのか、あるいは全部プログラムで動いているAIが出したものなのか）の区別も分からない。オートマチックに動くAIが出すキューで、ディセプティブな（人をだますような）ことは簡単にできてしまうので、信頼較正キューのデザインが一番本質的な部分の一つである。

Q：信頼較正のアイデアはとても興味深い。しかしながら、実は日常普通の人がやっていることだという気もする。例えば、結構使っていて、その結果を見て、ああ、こいつはこれぐらいできるはずというのを、過信と不信の間を行ったり来たりしながら、いいところに落ち着いていくようなことは、人間関係でもあるのではないか？

A：はい。信頼較正キューのようなものを時々出していることはある。

Q：ふだんは信頼較正キューなしで結果を見て人間側が判断することがあるかもしれないが、たまには（自分はもっとできるんだという）信頼較正キューを出してくる人もいる。

A：確かにある。結果だけでここに自分で持っていってもらえる人だと、このシステムは要らない。実験をやると、なかなかそうならない人がたくさんいる。それは人間対人間ではないというのが効いているのかもしれない…。

Q：やはり、ある程度機械の気持ちが分かる（中身の分かる）人は、落ち着きやすいという気がする。

A：仰る通り。だからこそ、中身を何とかうまく人間に伝えようという研究になる。 AIの中身が分かっていない人にも、現状のAIの能力が分かるようにしておくと、自分で勝手に較正してくれるだろうという立場だが、すぐに見なくなってしまう。

Q：ヒューマンインターフェイスでシステムモデルとメンタルモデルを一致させることになるのか？

A：それは、一般に難しいといわれている。メンタルモデルはその人間の中にあり直接的に外から操作できないので、自分で変えるしかない。

Q：AIと一緒に共同作業をしているときのパフォーマンスを考えるに当たっては、信頼較正キューをどのように変更して目的変数のパフォーマンスを最大化するかという問題に帰着しているように見えるが…。

A：はい。ざっくり言うとそうだ。

Q：結局バイアス（主観的なところ）を除去する観点で信頼較正キューのデザインを考える上で、途中に信頼というものをかませないといけないというところがポイントだと思うが、信頼というものが中間変数になっているという理解でよろしいか？

A：はい。今のフレームワークでは中間変数になっていると考えるのがいい。

Q：すると、やっぱりこのフレームワークのメリットは、主観的な部分と客観的な部分を整理しやすいフレームワークにしている点であり、そこが新しいポイントであると思えばよろしいか？

A：はい、今まで信頼較正を定式化した例はほとんどない。状態をどう判定（つまり信頼の計測）して、図2–10–3に示すように、どうやって適正な信頼領域へ持っていくかという具体的な方法論はほとんどない。我々のオリジナリティーとしては、まずこのAI–人間選択タスクを作って、この上に乗っかった選択行動を見ることによって人間が不信か過信かを判定することができることは、今まではなかった。また、不信か過信かを判定したときに信頼較正キューという刺激を与えて、それによって較正を実際にしてもらう信頼較正キューのいくつかデザインをしたというところも新しい。

Q：我々はコミュニケーションロボットをやっているが、その目的は、判断能力が低下している（信頼を得ることが難しい）高齢者の見守りに近い。信頼を得るというより、会話を推進して活性化することで毎日の生活クオリティーが高まることを狙っている。

A：我々もHAIでのコミュニケーションロボットの研究を行っており、高齢者の見守りは一つのメインターゲットだ。ただ、その目的は、バイアスを軽減しようとする方向とは違って、Reliabilityがある程度低くてもAIに対する主観的な信頼を高く持っていく方向である。従って、目的によっては、バイアスを強くしようという方向が必要な目的のタスクもあると思う。親和性を高めることによって、主観的な信頼を高める方法が、HAIでよく使われている。親和性を高めるというのは、信頼性が高まっているわけではなくバイアスが高まっているのだ。つまり、人間が感じるのは、客観的な信頼性ではなく、主観的な信頼の方なので、そこが高まればいいという考え方だ。また、別のアプローチとして、信頼できるようなエージェントの外見をデザインする（例えば、眼鏡をかけたり、ひげを生やしたり、ちょっとした小物を持たせる）ことによって信頼性が一気に高まる方法もあり、現在探求中である。そこはエージェントの外見のデザインに入るので、HAIのメインターゲットの一つとなる。

Q：親和性を高める手法は、使い方によってはよろしくない場合もある？

A：はい。

Q：性能は全然上がっていないのに見かけで信頼させるのは、ある意味でAIへの抵抗感（あるいはシステムに対する抵抗感）を持つ人やゼロリスクをすごく考えている人に対して少し敷居を下げてあげるという点で、使い方はいいと思う。しかし、ある種の詐欺のような使い方になってしまってはいけない場合には注意が要ると思うが、いかがか？

A：はい。厳密に言うと、使っている信頼較正キューにも、ある種のディセプティブな（人をだますような）要素が含まれるので、参加者実験をやるときはすごく気を使う。

# 2.11 中川 裕志[26]「AIのトラスト」

### パーソナルAIエージェントとは

最初にパーソナルAIエージェント（PAI Agentと略記）を取り上げる。

従来、データ主体の個人は、世の中のさまざまなサービス事業者に直接向き合っていた。サービス事業者はさまざまで、なかには悪い業者もいるし、情報環境は一般人には複雑すぎる。そこで、各個人がPAI Agentを持ち、PAI Agentを介してサービス事業者に対することで、この複雑な状況に個人が対応できるようにしようという考えが出てきた。

PAI Agentは、データ主体の個人の購買履歴や医者にかかった履歴など、さまざまな履歴を把握し、さらに、どんなサービス事業者に対して個人情報をどこまで開示したかとか、どのようなタスクや条件のときにはそれを止めたかとか、個人の嗜好や意思決定のルールブックのようなものも蓄積する。このようなPAI Agentは長く使っていくと、データ主体の個人の分身・デジタルツインとも言えるようなものになってくる。

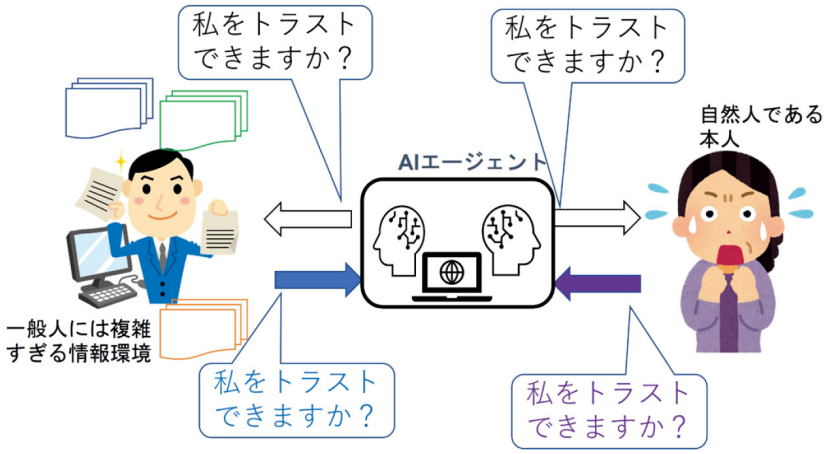


**図2-11-1　　パーソナルAIエージェント**

### パーソナルAIエージェントと事業者の間のトラスト

ここで、PAI Agentに関わるトラストとして、図2-11-1のような4種類がある。

1つ目は、事業者がPAI Agentをトラストできるか（図2-11-1の左上）。これは認証の問題なので、技術的には、IdP（Identity Provider）と呼ばれる認証サーバーが外にあるならば、このIdPを通して事業者がPAI Agentを認証できるかという話である。デジタルアイデンティティー[1]として検討され、既に使われている。ここでは、まず、エンティティーとしての認証にはFIDO 2.0という標準がある。これが認証されたら、複数のIDが存在する場合にID連携認証が行われるが、これにはOpenID Connectという標準がある。そして、1回のアクセスごとのアクセス認可にはOAuth 2.0が使われている。

今後注目したいものとしてSelf-Sovereign Identity（自己主権型アイデンティティー：SSI）が挙げられる。これは、管理主体が介在せず、個人が自身のアイデンティティーをコントロールできるようにしようというもの

26　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研究グループ 社会におけるAI利活用と法制度チーム チームリーダー（東京大学名誉教授）
https://sites.google.com/site/nakagawa3/home
https://researchmap.jp/nakagawa3

である。この場合は多分、PAI Agentが自分自身をIdPとして個人認証を行うようなことになる。できるだけ少ない情報で個人認証をさせるということも目標としている。現状はIdPをGoogleなどの事業者が押さえているため、アカウントBAN（アカウントを剥奪・停止されること）をされると、もう誰も認証してくれなくなり、身動きが取れなくなるが、SSIはアカウントBAN対策になる。 SSIは、分散型IDと鍵ペアを受け付けるシステムが普及すれば可能になることが分かっている。つまり、プラットフォームが分散IDと鍵ペアを他者に通過させてくれればよいだけなのだが、なかなかそうはならない。

2つ目は1つ目の逆、PAI Agentが事業者をトラストできるか（図2–11–1の左下）。これは古くからある問題で、サービスの利用契約に基づき、法的かつ形式的なトラストである。個人データの利用法について、目的の定義が広過ぎるとトラストしにくいし、第三者移転の有無も注意すべき点である。このようなチェックは結構大変で、PAI Agentがこのチェック能力を持つことは理想だが、実現は容易ではない。現状、組織としてのデジタル認証をインターネット経由で行えたり、PマークやCEマークなどの認証マークが設けられたりしている。



**パーソナルAIエージェントと個人（データ主体）の間のトラスト**

3つ目は、PAI Agentがデータ主体である個人をトラストできるか（図2–11–1の右下）。つまり、データ主体の個人がなりすましかもしれないので、本人確認をする必要があるということである。 PAI Agentがスマホ やPC内のソフトならば、持ち主がそれらを使うときの認証（パスワード、生体認証、2段階認証、FIDO 2.0）が用いられる。スマホなどが乗っ取られることもあるわけで、PAI Agentが乗っ取られたことを認識して自分自身をKillできれば、損害を回避できるかもしれないが、簡単ではなさそうである。

また、PAI Agentがプラットフォームの個別ユーザーインターフェイスとして実現され得る。Googleアカウントや、FacebookなどのSNSアカウントのような形でPIA Agentが実装されるというのは、むしろ可能性が高そうである。このとき、悪意のある人が別ユーザーになりすますというのは、プラットフォームとしても嫌なことなので、プラットフォーム側がAI技術を駆使して監視するだろう。例えば、ユーザーがいつもと違う行動をしているかを監視するというのが考えられるが、正しい利用者でも、いつもと違う行動をするかもしれないし、考え方を変えることもあり得る。ログインするときに正しい利用者しか知らない鍵を用いる2段階認証は効果的だが、敵もさらに強力な手段を使ってくるかもしれないので、いろいろ考えていかねばならない。

4つ目は、データ主体である個人がPAI Agentをトラストできるか（図2–11–1の右上）。この問題は結構難しい。PAI Agentの開発・販売会社をトラストできるか、その利用契約はどうなっているかに注意する必要がある。 PAI Agentがプラットフォームの個別ユーザーインターフェイスとして実装された場合には、その利用契約がさらに重要になる。

また、PAI Agentは、データ主体であるユーザーの思惑と違う行動をすることがあり得る。ユーザーは、PAI Agentが自分の意思を学習してくれて、自分がふだん思っている通りにやっておいてほしいわけだが、そうお任せしてしまうと思惑と違うことをされてしまうかもしれない。そこで、PAI Agentの行動をユーザーが遅滞なく認識できるかが問題になる。 PAI Agentを監視するPAI Agentが必要かもしれない。行動停止をさせることはできるだろうが、例えば第三者との契約のような、既に行った行動を破棄できるかというと、これは法的問題として解決策を考えておく必要がある。 PAI Agentの免責事項もあらかじめ決めておくことも必要である。このような面も含めて、ユーザーがPAI Agentをトラストできてこそ、Agentと協働したり、任せたりといったことができることになる。

改めて問題点を考えると、PAI Agentの問題は、いわゆるCybernetic Avatarと共通するところが多い。Cybernetic Avatarはリアルロボットであることが多く、問題はさらに複雑になる。それから、場合によっては、PAI Agentの他言語対応能力もクリティカルである。ユーザーの言語圏だけでなく、英語・中国語など異なる言語圏とのやり取りも必要になる。

### EUにおけるAIトラスト

ここで話題を変える。AIとトラストを最初に扱ったのはEUである。EUにおけるトラストの系譜は、まず2018年に欧州委員会（European Commission）のHigh-Level Expert Group on AI（HLEG）が「Ethics Guidelines for Trustworthy AI」[2]を出し、その後2020年に「AI白書」（AI White Paper）[3]が出された。これをもとに2021年には「AI規制法案」（Artificial Intelligence Act）[4]が出された。

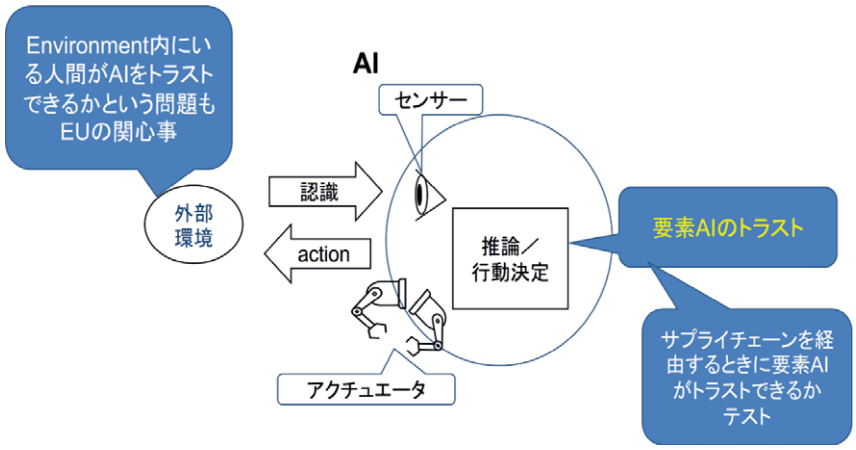


**図2-11-2　　EUのAIとトラスト**

まず「Ethics Guidelines for Trustworthy AI」では、「AIとは、特定の目標を達成するために、環境を分析し、ある程度の自律性を持って行動を起こすことにより、知的な行動を行うシステム」と定義している。図2-11-2のように、外界に対するセンサーやアクチュエーターを持ち、いろいろと推論や意思決定をするAIがあるというわけで、やはり要素AIをトラストできるかという問題が出てくる。さらに、サプライチェーンが非常に長く複雑なシステムになることから、サプライチェーンごとに要素AIをトラストできるかをテストする必要がある。また、環境（Environment）の中にいる人間がAIをトラストできるかという問題もある。こういったことがEUの関心事であり、さきほどのPAI Agentのトラストとつながる面もある。

### Trustworthy AI

Trustworthy AIの発想では、Training Dataに恣意的なバイアスが入ってないこと、倫理性、人間中心といったことを求める。倫理性は、基本権として、個人の尊厳、自由、平等と連帯、市民の権利と公正があり、公益を最大化しながら個人の権利と自由を保護するものであり、当然、法令遵守も含む。人間中心というのは、常に人間が上位の決定者であると言っている。EUだけか分からないが、キリスト教圏の国は、このように人間を絶対上位にもってくる。

EUは「Trustworthy AI＝倫理性＋人間中心＋技術的なトラスト」と捉えている。技術的なトラストというのは、工学的なツールとして信頼性が高いという意味である。ロバストと言ってもよく、こういうインプットをすればこういうアウトプットをいつも返してくれて、壊れたりしない、という意味でのトラストがTrustworthy AIのトラストである。

Trustworthy AIの条件は、①アカウンタビリティー、②データガバナンス（バイアスの除去）、③誰でも使えること、④AIの自律性をガバナンス、つまり、人間が監視・早期介入できること、⑤差別しないこと、⑥人間の自律性を尊重し促進すること、⑦プライバシーの尊重、⑧ロバスト、⑨安全性、⑩透明性。ロバストは技術的には重要で、信頼性と再生成性、正確さ、攻撃への耐久性と立ち直り、止めるときの計画をあらかじめ立てておくことを求める。これは普通のソフトウェアに対する考え方と一緒で、EUはやはり、AIは特別

なものではなく、普通のソフトウェアだと思いたがっている。要するに、EUはAIを信じていなくて、人間が制御できるように限定したAIにしたいという意図が非常に強い。

**透明性や説明可能性とトラストの関係**

ここで、透明性や説明可能性とトラストの関係を整理しておきたい。図2-11-3は、大屋先生から話をうかがって、私なりにまとめたものである。

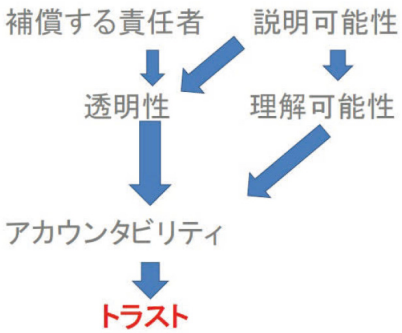


**図2-11-3　　透明性、説明可能性、トラスト**

まず、説明可能性とよくいわれるが、これだけが独り歩きするのはまずくて、理解できるものでなくてはならない。一方、透明性はそれだけではなくて、補償する責任者、つまり、何かまずいことが起こったときに補償は誰がしてくれるのかが求められる。アカウンタビリティーというのは、この2つのルートが合わさって得られるが、これは日本語で言う説明責任とは違う。説明すればいいよというだけのものではない。

理解可能性については、AIの内部の動きを理解できる形で説明するのはほとんど無理で、近似的なモデルを作って説明することが行われている。例えば、非常に複雑な分類問題も、部分的に見ると結構直線で分けられて、その分かりやすい直線的な分け方をつなげていくことで、分かりやすい説明を近似的に作れる。説明可能AI（XAI）というのは、このような分かりやすさを目指す技術になっている。

また、トラストの実態として、例えば、大きな会社だし、みんな使っているから信用しようかとか、あるいは、自分と似たようなケースは、自分と同じ結果が出ているようだから、まあトラストしておこうとかいうことが多い。

**EU的トラストを実現する技術**

次にEU的トラストを実現する技術には、Technical MethodsとNon-Technical Methodsがある。Technical Methodsとしては、まずアーキテクチャーがあり、デザインレベルからきちんとやっているか、いわゆるX-by-designがある。例えば、Privacy-by-designや、Security-by-designや、Ethics-by-designといった考え方のことである。また、実用に供する前はテストと検証をしっかりやる。実用に供した後はトレーサビリティーと監査可能性が大切で、先ほどのXAIもある。

Non-Technicalな方も重要で、法的規則、標準化、アカウンタビリティーを確保するガバナンス、行動規範（Codes of Conduct）、倫理的な考え方を育むための意識化と教育、ステークホルダー間の対話、社会との対話が挙げられ、さらに、多様性の確保と「誰かを取り残さない」ような設計を行うチームを作れとも言っている。Non-Technicalだけれど具体的で、これは技術者としても心得ておくべき問題だと思う。

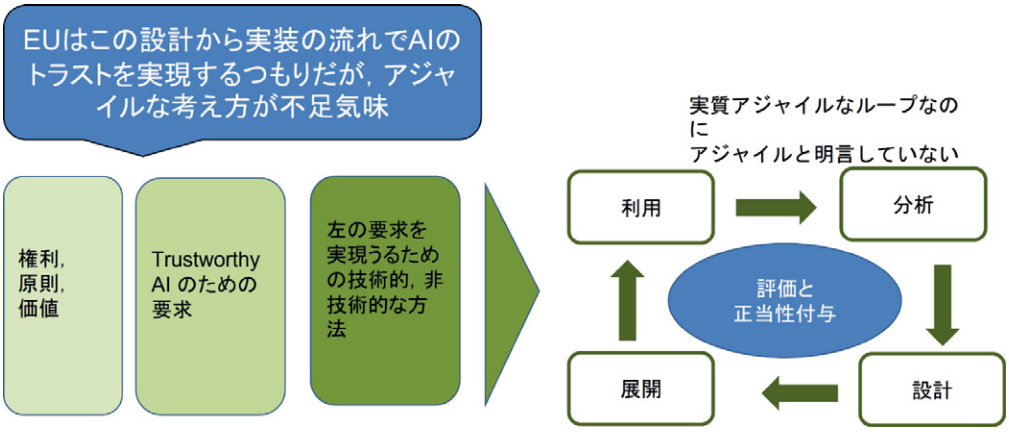


**図 2–11–4　　設計から実装への流れ**



これまで話したような設計から実装への流れを図 2–11–4 に示す。人権・原則・価値といったところから、Trustworthy AI の要件として、先ほど示したような 10 個の条件が出され、今話した Technical および Non–Technical な Methods を用いてシステムを作る。すると、実際にデザインして、開発・配備して、使用し、分析するという、評価と正当性付与のループが回り始める。これはアジャイルのループだと思うのだが、残念ながら EU ではそういう言い方はしていない。アジャイル的な考え方を強調してもよいのではないか。

**信頼できる AI のためのチェックリスト**

欧州委員会 HLEG は Trustworthy Assessment List[5]というものを作っている。約 60 個のチェックリストで、具体的でなるほどと思える内容になっている。日本でも QA4AI コンソーシアムが非常に充実したチェックリスト[6]を出している。

EU のチェックリストの具体的な内容としては、基本的権利に悪影響が及ぶかとか、異なる原則と権利の間の潜在的なトレードオフを見ているかとか、AI システムがエンドユーザーによる決定と相互作用するかとか、エンドユーザーがチャットボットなどと会話をするとき会話の相手が人間ではないと知っているかとか、いろいろ非常にうまく書かれている。 AI プロセスへの過度の信頼や依存を防止するために作業プロセスの安全性対策を講じたか、というのも非常に重要だと思う。人間の監督について、管理下にある人間は誰かとか、人間の介入のための瞬間やツールは何かとか、非常に技術的に解釈できるレベルまで落とし込んで書いてくれている。フォールバック計画と一般的な安全性についての項目もとても重要で、駄目になったときにどうするかという問題、閾値を定義してフォールバックをトリガーするようなガバナンス水準を作るとか、いろいろ考えておくべきであるとされる。

**EU AI 白書**

次に、EU は 2020 年に AI 白書[3]というものを出してきた。AI サービスは事前に徹底的なリスク予測を行うべきという考えである一方、リスク発生の予測は、AI 技術の複雑さや発展の早さから、技術的に抑えることは困難だということも意識している。とにかくリスクに注目し、AI システム製造のサプライチェーンの各段階で、倫理指針や法制度に基づくリスク管理や公平性・非差別を求めるとしている。このサプライチェーンをチェックするということは、EU だけではなくてアメリカも、中国産の基本ツールを念頭に置いてやっていることで、それをしていない日本は少し甘いのかもしれない。

それから、基本権、すなわち人権、人間の尊厳、多様性、非差別、プライバシーと個人データの保護などに対するリスペクトを推進すると言っている。これは非常にまともな価値観ではあるが、AI は既存の EU 法お

よびEU加盟国の国内法がまず適用されると言って、だんだんAIを敵視し始める。EUは、AIアプリケーションに関する潜在的なリスクに基づく証拠を明確化するために、既存の法的ないし技術的手段を最大限に活用するとしている。 AIを悪者扱いするものの、そうは言っても、AIの進歩を認めざるを得ないので、効果的な適用と施行を確実にするために、責任に関する特定の分野の既存の法律を調整または明確にする必要があるとしている。要するに、法律の方を変えても、あくまでも法律がコントロールするのだという非常に強い意志を感じる。さらに、サプライチェーン内の異なる経済事業者間の責任の割り当てに関する不確実性をなくせという方向であり、AI白書にはEUの非常に強い監督意識が強く出ている。

また、AIシステムが全てのライフサイクルフェーズでエラーや不整合に適切に対処できることを保証するとか、AIシステムの出力は人間によって事前にレビューおよび検証されていない限り有効にはならないとか、動作中のAIシステムの監視およびリアルタイムで介入して非アクティブ化するというのを人間がやるとか、書かれている。しかし、人間によって事前にレビューするとか、リアルタイムに人間が介入するとか、実際にできることなのか、何か現実感がないような気がしてならない。例えばAIトレーダーはマイクロセカンドオーダーで動いているわけで、これに人間が介入してクラッシュを防げるかと考えれば、非現実的と思えてしまう。このような考え方は、この後に話すAI規制法案に引き継がれている。

EUの保護主義的な傾向が非常に感じられるものになっている。 AIシステムを調整する必要がある場合は、例えばEUで再トレーニングすることで修正する必要があるとか、標準化されたEU全体のベンチマークに準拠していなければならないとか、言っている。AIシステムの仕様を倫理指針・法制度で管理しようとする指針を打ち出していて、開発者にとって非常な重荷になると予想される。これを懸念している企業は、日本でもとても多い。

**EU AI規制法案**

次は今年2月に出されたAI規制法案[4]である。その前文に書かれた目的を読むと、アメリカ・中国などEU域外で進むAI技術に対する不信感がまずある。EU加盟国はAIが安全であり、基本的権利の義務に従って開発および使用されることを保証するために、国内規則をフル動員する。ただし、EU内の国ごとの規則が異なると、国内市場が細分化されて、AIシステムを開発または使用する事業者が個別の国の法律に振り回される。そこで、EU全体で一貫して高レベルの保護規則を確保したい。オペレーターに統一された義務を課し、公益と国内市場全体の人の権利を最優先する保護規則が必要だという。非常に高邁で美しい文句だが、EUファースト的なものに思える。

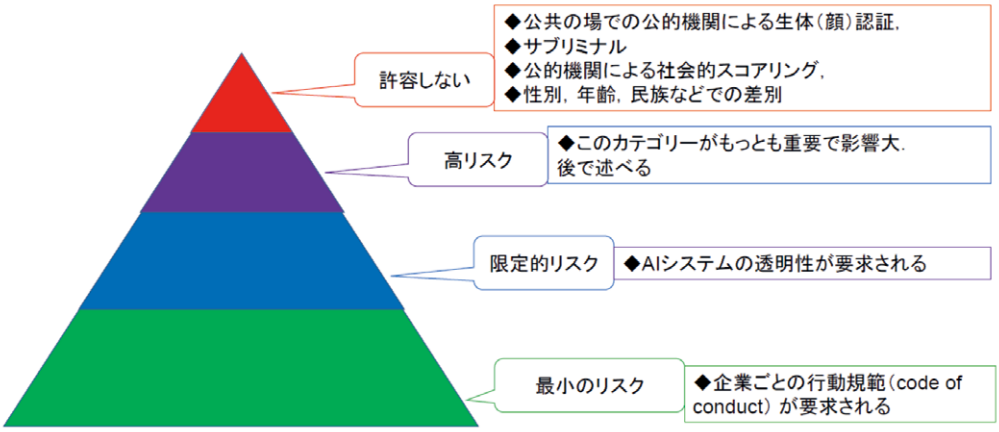


**図2-11-5　　EU AI規制法案の階層**

では、AI規制法案の内容に入る。図2–11–5のピラミッドのような4階層がある。

最下位層は「最小のリスク」、これは非常にリスクの少ないものなので、これは企業ごとに行動規範を作るぐらいでよい。次の階層は「限定的なリスク」、リスクが限定的な場合はAIシステムの透明性が要求される。ここまではよいとして、上の2階層が問題になる。3階層目は「高リスク」、これが実質的に最も影響が大きいもので、この高リスクAIのカテゴリーの詳細を表2–11–1に示した。そして、最上位層は「許容しない」とされ、例としては、公共の場での公的機関による生体認証、サブリミナル、公的機関による社会的スコアリング、性別・年齢・民族などでの差別が挙げられている。ある国を非常に意識したものになっていて、その国でやっていることを強く否定する立場が見て取れる。

**表2–11–1　　高リスクAIのカテゴリーの内容**

| | | |
|---|---|---|
| 1 | 生体認証とそれによる分類（リアルタイムと事後）差別があってはならない | ・リアルタイムおよび「ポスト」リモート生体認証に使用することを目的としたAI |
| 2 | 重要な生活インフラストラクチャーの管理と運用 | ・道路の管理と運用における安全設備として使用することを目的としたAI<br>・交通と水、ガス、暖房、電気の供給のためのAI |
| 3 | 教育と職業訓練へのアクセスの可否決定 | ・教育および職業訓練機関へのアクセス可否の決定または自然な割り当ての目的で使用することを目的としたAIシステム<br>・職業訓練機関および一般の教育機関への入学のためのテストの者を評価するためAI |
| 4 | 採用における利用、人事評価、労働者管理、雇用と解雇 | ・採用または選択、特に求人広告、アプリケーションのスクリーニングまたはフィルタリング<br>・候補者の評価面接またはテストのためのAI<br>・昇進と解雇の決定を行うために使用されるAI<br>・契約関係、タスクの割り当て、パフォーマンスの監視と評価のために使われるAI |
| 5 | 不可欠な民間サービスおよび公共サービスへのアクセス適格性評価順位付け | ・公的機関によって、または公的機関に代わって使用されることを目的としたAI<br>・公的支援の給付とサービスに対する適格性を評価するAI<br>・そのような利益およびサービスを付与、削減、取り消し、または再請求するAI<br>・クレジットスコアを計算するAI<br>・消防士や医療援助などの緊急性ある処理に対する優先順位を計算するために使用されるAI |
| 6 | 法執行機関が個人の状況に立ち入る | ・個人のリスクを高めるために法執行機関が使用することを目的としたAI<br>・再犯または刑事犯罪の潜在的な犠牲者のリスクを評価するAI<br>・法執行機関がポリグラフなど感情状態を検出するために使用するAI<br>・法執行機関がディープフェイクを検出するために使用するAI：第52条（3）で透明性に関与するとして言及されている |
| 7 | 移住、庇護および国境管理での利用 | ・公的機関がポリグラフや感情状態を検出するためとして使用することを目的としたAI<br>・公的機関がセキュリティーリスク、不法移民のリスク、または健康リスクを含み、加盟国の領土に入ろうとする、または入国した人を検査、検証するAI<br>・渡航文書の信憑性と裏付けとなる文書セキュリティー機能をチェックすることにより、本物ではないドキュメントを検出するAI<br>・公的機関による審査を支援することを目的としたAI<br>・庇護、ビザ、居住許可の申請および関連する苦情ステータスを申請する人の適格性を審査するAI |
| 8 | 司法当局が事実を調査および解釈するのを支援するAI | |

高リスクAIをEU市場に出す前にやらなければならないことが、製造業者、配布業者、サービス事業開発者（プロバイダー）、事業者代表、公的ユーザーなどのグループに対して細かく指示されている。当然、EU域外企業もEUで商売したい場合は適用される。全利用期間におけるリスク管理、データガバナンスの問題、技術文書をきちんと作れということ、利用状況レコードを保存する機能、ユーザーへの内容説明や監視を2名以上で確認すること、技術的ロバストさ、セキュリティーの確保などが求められている。CEマーキングを取れということもいわれている。 CEマーキングは、アメリカで特定の電子機器を販売するときに使用されるFCC（Federal Communication Commission）適合宣言のようなものである。製品が健康、安全、および環境保護に関するEU基準を満たしているというメーカーの宣言で、違反したときの罰則として3千万ユーロまたは全世界の年間総売上げの6%が上限といったことも書かれている。まずいことが起きたときは必要な訂正行動を即時行うとか、高リスクAIをEUのデータベースに登録するとか、市販後の動作のモニタリングと稼働状況レコードを保存するとか、インシデント報告義務なども書かれている。



**【主な質疑応答】**

Q：PAI Agentが事業者をトラストできるかのところで、事業者の認証は、単に事業者本人だという本物認証ではなく、その事業者は悪いことをしないとか、そういったことも含むか？

A：もちろん含む。そういった面も含めて、認証機関がきちんと認可した事業者であることが確認される必要がある。それが銀行のような規制が非常に強い分野は分かりやすいが、AIのように柔軟な分野でうまく実現できるかは課題である。また、事業者やそのサービスの評価においては、サービス利用契約を正確に理解することが重要だが、それは利用契約の表面的な理解で済むものではないはずで、PAI Agentにとって大きな技術的課題だと思う。そうすると、PAI Agentが単独で判定するのではなく、同じような条件で他がどうしているのかとか、口コミ的な情報を参考にするとか、PAI Agentネットワークのようなものも必要なのかもしれない。そういった仕組みを作りやすいという面でも、PAI Agentはプラットフォームでの実現が進みそうに思っている。

Q：PAI Agentのアプリケーションとして早く立ち上がりそうなものは何か？

A：FacebookのLegacy ContactやGoogleのInactive Account Managerなど、死後のアカウント管理が既に提供されている。また、個人データを集めて、その人の仮想AIエージェントや仮想アバターを作るビジネスもたくさん出てきている。母子手帳やお薬手帳の電子化も、個人情報の管理が非常に重要なのでPAI Agent的な可能性が高い。小さな分野で一つ一つ進んでいくと思う。

Q：AIのコミュニティーの中でトラストはいつ頃から考えられるようになったのか？

A：今回の紹介したEUのAIトラストは公的文書として出たのが2018年頃からなので、研究としてはもっと前から取り組まれているはず。 AI倫理としてまともな文書化されたのは、2017年のアシロマAI 23原則が知られている。その後、IEEEのEAD（Ethically Aligned Design）がよく知られている。その間、非常に短い間に、言っていることがすごいスピードで変化したという感じがある。

Q：図2–11–3でアカウンタビリティーからトラストへのジャンプには、技術以外の要素、例えば保険のようなものが必要だと思うが、具体的な動きはどうか？

A：自動運転をはじめ、保険の出番は今後結構出てくると思うが、現状まだ、自動運転レベル3以上のケースはどう責任を問うかよりも、そのような事故が起こらないようにすることに主眼を置いて進められている状況だと思う。

Q：EU AI規制法案の高リスクAIについては、第三者認証をどこかの組織が担うことになるのだと思うが、どのような見込みか？

A：高リスクAIをどう監督していくかは、新たに設置されるというボード（European Artificial Intelligence Board）が担っていくものと思う。その監督の下でCEマーキングをどういう組織体制で運用していくかについては、まだよく分からない。

Q：AI規制の一般論とは別に、医療機器、自動車、航空機のような、もともと規制の強い産業分野の場合、その規制当局が何らかの認可の仕組みを作らない限り、その分野にAIは入らないのではないか？　第三者認証というのも、産業分野ごとに作られるということはないか？　EUは自動車産業が強いので、そのような動きはないのか？

A：やっていけないことと、やっていいことのバウンダリーを明確にすると民間はやりやすい。EUの自動車産業は強いので、決められたものを受け入れるだけでなく、自分たちで決めようという姿勢も強いと思われ、注目していきたい。

Q：EUの規制法案に日本でも追従しようと考えるセクターはどうか？

A：日本国内では、基本的にソフトロー、つまりあくまでガイドラインという考えで、EUのような規制法案は作らないというコンセンサスがあるように感じている。ただし、EUでビジネスをする際には、EUの規制に従わざるを得ないという立場を取る。アメリカの場合は、GAFAなどがEUに対して矢面に立ってやり合っていて、課徴金を食らったり、法的訴訟もやったりしているが、残念ながら日本にはそういうパワフルで資金力のある企業はいない。アメリカは政権や長官が変わって、どう動くかにも注目しておいた方がいいだろう。

## 2.12 工藤 郁子[27]「公共政策とトラスト」

**社会的関係資本としてのトラスト**

「トラスト」を考える上で示唆的な社会実験を紹介しよう。とあるカフェのFree Wi-Fiの利用規約の中に、「サービスの利用に際し、F-secure（セキュリティー企業）に第一子または最愛のペットを譲渡することに同意する」という条項を設けたところ、ほんの数十分で6名が同意した[1]。

このように、建前として利用者はプライバシーポリシーを読んで同意していることになっているが、実態としては目先の利益に釣られてプライバシーポリシーをほとんど読まずに同意しているというケースはしばしば見受けられる。多くの人が利用規約を読まずに済ませているにもかかわらず、円滑に社会が回っている。ここに法制度と「トラスト」の関係を考えるヒントが隠れている。



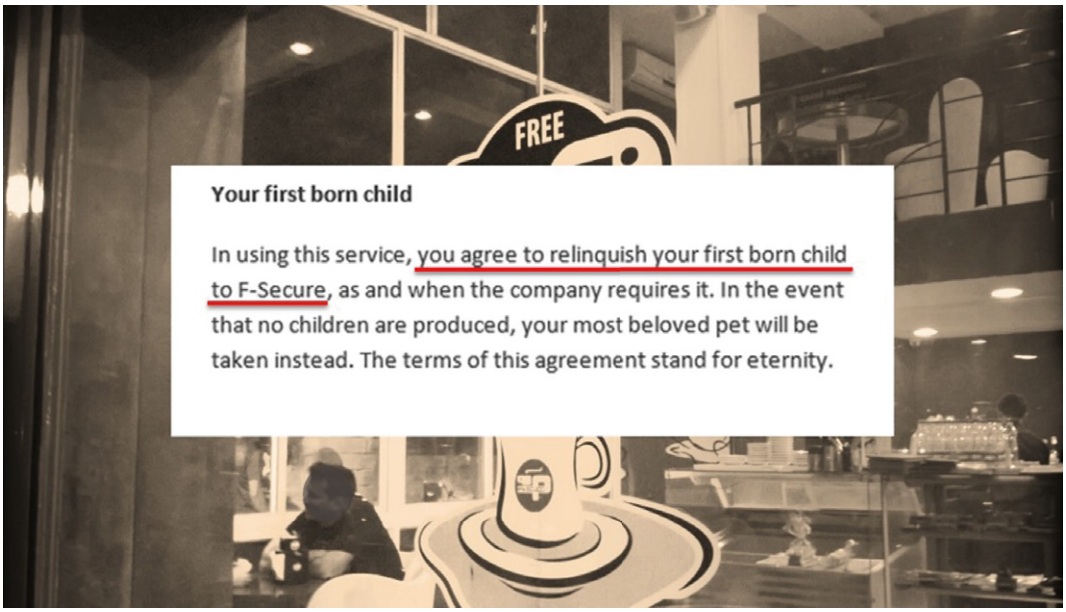


**図2-12-1　　ヘロデ条項（Herod Clause）実験**

多くの人は、著名企業が「悪」ではない、または、「悪」だとしても行政指導や社会的批判などで何らかの衡平（Equity）が図られる、と暗黙のうちに期待している。利用規約を読み飛ばして、SNSや動画を楽しむことができるのは、そこにトラストが存在しているためだ。取引・協力のコストを低減し、ある種の社会関係資本（Social Capital）としてトラストが機能していると言える。

このケースで見られるように、法制度と実態は乖離しているように見えるが、実際は相互に関係している。行政指導や紛争解決手段としての裁判があることで、裏切られるリスクが軽減される。トラストを促進・保証する機能を法制度は有している。

また、トラストを基盤とした取引・協力関係が日常化すると、自分は期待していないかもしれないが、「他の人たちは信頼しているはず」というメタなトラストに基づき好循環が生じる。これによりさまざまな社会的コ

27 大阪大学社会技術共創研究センター招聘教員、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
https://researchmap.jp/fumikok

ストが低減されている。

しかし、現在はトラストに揺らぎが生じている。「信頼が強く意識されるのは、それが壊れ失われつつあるとき」だ[2]。以下では、公共政策（Public Policy）においてトラストに揺らぎが生じるいくつかの場面を紹介したい。

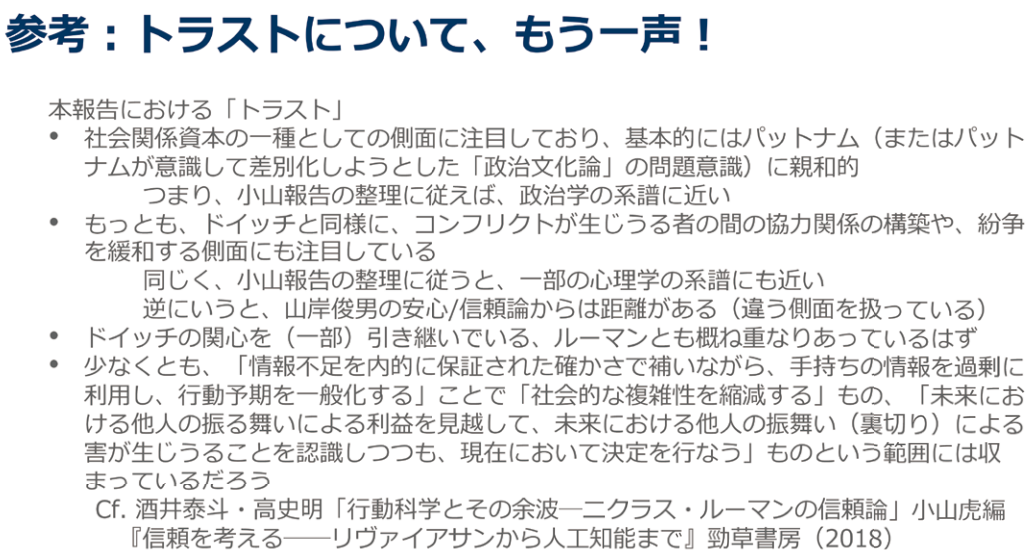


**図2-12-2　　トラスト研究についての補足**

**政策における専門知のトラスト**

政策における専門知のトラストについて紹介する。この課題は科学技術社会論（STS）や科学哲学などでも以前から議論されてきた。今回は特に、Theodore Porter（セオドア・ポーター）が指摘するような、数値（計量化、手続規格化、指標化）と専門知（Expert Judgment、Local Knowledge）との間で生じる、トラストをめぐる綱引きについて扱いたい[3]。

COVID-19パンデミックでは、多数の数値指標が飛び交った。新規患者報告件数、接触歴など不明者数、重症患者数、PCR検査の陽性率、病床の占有率などだ。しかし、何をどう測りどう解釈するかは、専門知に基づく判断の集積であり、ある程度「振れ幅」がある。既知の問題ですら妥当な指標を絞るのは難しいとされるが、COVID-19のように公共政策上のゴールが常に更新され得る場面においてはなおさら難しい。そのような状況で数値化を進めれば、ポーターが指摘しているように、専門家集団の裁量や自律性を失わせることにつながる。いったん受け入れられた指標に基づいて判断するということは、指標からこぼれ落ちる細かなニュアンスを含む複雑な情報や専門的知見に基づく総合的な判断を阻害してしまうからだ。加えて、測定対象だけに労力を割くことで本来の目標から逸脱してしまったり、長期的視座を見失いがちになってしまったりもする。これは、「測りすぎ」と呼ばれる課題の一つである[4]。

他方、数値化の要請は故のないものではない点も強調しておきたい。数値指標であれば門外漢でも把握可能であり、それに基づく決定は、没個性的で透明で明確に見える[28]。つまり、「政策的判断はできるだけオープンかつ公平であるべきだ」、「一握りのエリート集団の判断で重大な決定がなされるべきではない」、「誰もが

**28**　もっとも、不適切な数値指標を強引に導入すれば、かえって恣意性が高まり自由裁量が広がることもありうる。その際、より自由な裁量を手にするのは、自然科学の専門家というよりはむしろ政策の専門家である官僚かもしれない。

分かるような形で説明責任を果たすべきだ」といった価値観が重視される政治文化に応えようとして、数値化は生じる。こうした要請を直ちに棄却することはできないだろう。

アメリカの反知性主義（Anti-intellectualism）に関する議論は、この文脈において大変示唆的である。反知性主義は、知識を積極的に否定し、むしろ知識がない方が良いという考えを指す。この思潮を析出したRichard Hofstadter（リチャード・ホーフスタッター）によると、民主性・平等性の価値観に根差したものと指摘されている[5]。アメリカでは、例えば、象牙の塔にこもる経済学者よりも新進気鋭の起業家の発言を尊び、「机上の空論」よりも実践から得られる知見や技術を重視する傾向が強いとされる。この傾向は、権威や特権との折り合い方という歴史的・文化的背景に依拠している。つまり、欧州では、教会や貴族などの権威が知性・知識と不可分だったという歴史を持つが、封建制・身分制から距離をとったアメリカでは、その特権性に批判的な政治文化がある。エリート集団が強くなると、反知性主義の動きが盛り上がり、それがエリート集団すなわち知識人排斥につながって衆愚に堕ちていくことが何度も繰り返されてきたとされる[29]。

先述したように、これは、数値に寄せられるトラストと、エリート集団である専門家の判断に寄せられるトラストとの間で生じる綱引きであり、信頼相当性（Trustworthiness）を裏打ちするはずの「客観性」の形成を巡る議論だ[30]。



**政治における主権者のトラスト**

次に、政治における主権者のトラストについて紹介する。Thomas Hobbes（トマス・ホッブズ）以来、「眠れる主権者」という比喩は連綿と受け継がれてきた。統治者を選出する「民会（Convention）」のときに人々は目覚めるが、常日頃は眠り続けるというイメージである。

これは、法理論上、主権（権利）と統治（執行）の分離を示す[7]。本来、主権は絶対的で不可分であるはずで、それは君主から国民の手に渡ったはずであるが、実際のところは政府が掌握しているように見える。この落差を埋めるべく、主権は国民の手元に残されているが、執行が政府に委ねられていると説明される。裏返せば、人々は、古代都市国家で政治を支えた公民のようにはいられず、（政治などの公的生活だけでなく）私的生活を営んでいるので、実際の統治作用を政府に「信託」せざるを得ないということだ。日本国憲法前文の「国政は国民の厳粛な信託による」という文言の理論的背景にもなっている。

この「眠れる主権者」のイメージは、Jean-Jacques Rousseau（ジャン＝ジャック・ルソー）などのそれ以降の憲法学者や政治学者にも継承されている。例えば、現代では、憲法学者のBruce Ackerman（ブルース・アッカーマン）が、二元的民主政理論を提唱しており、米国における民主的プロセスを「通常政治」と「憲法政治」の二つに分ける[8]。そして、通常政治においては、人民が眠りについているので裁判所が司法審査を行うことで憲法を保障しなければならないが、憲法政治においては、人民は「我ら合衆国人民（We the People）」として行動し、熟議に基づく高次法形成を行うことで憲法の在り方を最終的に決定すると主張する。これは、人々は通常は「眠って」いるが、根本的改革が必要なきには「目覚めて」熟議するモデルと解すことができる。

このイメージをさらに敷衍すると、「睡眠不足問題」も想定される。従来、ルソーらは、政府による主権の簒奪を警戒し、定期的に「目覚める」ことを提唱していた。しかし、不確実性や流動性が高まり、根本的変革が連続しがちな現代において、覚醒を求め続ければ、「薄眼を開けて起き続け」ることになり、合理的な判断ができなくなる[9]。人間の認知資源は有限である点を考慮に入れなければならない。

**29** マッカーシズムが吹き荒れる1950年代前半に知識人排斥を体験した（知識人との自認がある）ホーフスタッターは、それでも民主性や平等性の尊重を肯定する姿勢を崩さなかったが、同時に、1950年代後半のスプートニクショックを起因とする知識人の再評価に戸惑いを示している。

**30** ロレイン・ダストンとピーター・ギャリソンが示す通り、「客観性」は多義的であって、その意味するところは、ある時代やあるコミュニティーにおいて重視される「規範」に左右される。文献[6]を参照のこと。

この側面において、「眠れる主権者」の比喩は、安心して眠る、すなわち、トラストできるような政治制度の整備が必要だという議論につながる。

トラストが機能不全だと、自分の力のみで複雑さとリスクに向き合わねばならない。アメリカの社会生物学者 Edward Osborne Wilson（エドワード・オズボーン・ウィルソン）は、「私たちは石器時代の感情や感覚で、中世の組織・体制・制度のまま、神のような技術を持って、21 世紀にうっかり入ってしまった」と指摘している。人間、制度、技術の間に埋めがたいギャップがある現状を受け入れた上で、トラストの再構築を考える必要がある。

**トラストの再構築に向けて**

近代の人間像である「合理的な存在」は、自然的な事実認識というよりは、規範的に構築されたモデルであり擬制である。我々は「あるべき姿」を実現するために、制度を構築し、技術を駆使してきた。つまり、機能分化と流動性上昇に対応するために、無理して拡張（Enhancement）してきた。しかしながら、今日の社会の複雑化と情報量の飛躍的増大に対し、個人も制度も追い付けていない。例えば、人々がSNSなどでさまざまな情報を発信できるようになったことは民主的で平等が実現されている点で良い社会であると考えられるが、事前にふるい分けされずに発信され流通する大量の情報には、偽情報や誹謗中傷が含まれ、情報環境全体、ひいては社会に対する、不信と冷笑を醸成する事態を招いている。

また、先に紹介した政治過程における通常政治と憲法政治のモード切替えというメタ決定は、機能不全に陥っているのではないか。

「合理的な存在」という理想・理念を実現するためには、技術をどのように使うかという視座が重要である。

**Data Free Flow with Trust（DFFT）**

トラストの再構築の一例として、DFFTを紹介する。 Data Free Flow with Trust（信頼性ある自由なデータ流通）は、日本が主導するデータ流通や電子商取引に関する国際的なルール形成に関する構想である。2019年1月にダボスで開催された年次総会にて安倍前首相がDFFTを提唱し[10]、同年6–7月に日本を議長国として開催されたG20にて「大阪トラック」として立ち上げられた[31]。その後も、G20、G7、WTO、OECD、世界経済フォーラムなどグローバルを舞台に議論・交渉・連携が進んでいる。


**図2–12–3　　DFFTのグローバル・ルールメイキング**

31　https://www.mofa.go.jp/mofaj/ecm/it/page25_001989.html

直近では、2021年4月に行われたG7デジタル・技術大臣会合では、DFFTに関する協力のためのロードマップが策定・合意された[11]。また、同年12月には、WTO電子商取引交渉に関する共同議長国閣僚声明が発出され、DFFTに関する交渉の進捗が確認されている[12]。

## 参考：DFFTトラスト白書

- 白書 “Rebuilding Trust and Governance: Towards DFFT”では、トラストの再構築にはガバナンスのアップデートが不可欠であることを指摘し、トラスト・ガバナンス・フレームワークを提案
- 経済産業省の報告書 「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」の問題意識とも連動

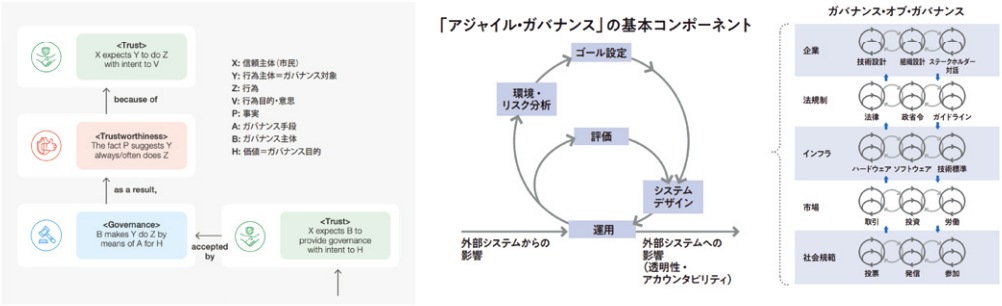


**図2-12-4　　DFFTトラスト白書**

## 参考：G7デジタル・技術大臣会合(UK)

- G7 デジタル・技術担当大臣は、「building back better」をテーマに、2021年4月28日に会合
  - 日本からは、武田総務大臣と佐藤経済産業大臣政務官が参加
- 大臣宣言における、**データ・フリー・フロー・ウィズ・トラストに関する協力のためのG7ロードマップ**該当部分：
  - 2019 年の G20 大阪首脳宣言及び G20 貿易・デジタル経済大臣会合閣僚声明、2020 年の G20 リヤド首脳宣言を踏まえ、我々は、志を同じくする、民主主義的で開かれた外向的な国として共有する価値に基づき、信頼性のある自由なデータ流通による利益を実現する取組を支持する
  - このことを実現するため、我々は、このアジェンダに関する具体的な進展をもたらし、企業や個人が技術を利用する際の信頼性を高め、経済的・社会的価値を高めるための方法を示した「データフリーフローウィズトラストに関する協力のための G7 ロードマップ」（附属書２）を承認する
  - 本ロードマップの一環として、我々は、合意した優先分野における相互に受入可能なデータ共有プラクティスの発展を加速化していく。また、データローカライゼーションによる経済・社会的影響を立証する。さらに、OECDによる「越境データ移転に対する規制アプローチの共通項マッピング」や、信頼性のある「民間セクター保有の個人情報に対するガバメントアクセス」に係る取組の進展を支持する

**図2-12-5　　G7デジタル・技術大臣会合（UK）**

**ガバナンスイノベーション**

DFFTを推進することは「データを活用する際にトラストを確保する」、「縦割りを打破してデータ活用を連携する」という話に限定されるわけではない。社会の前提が情報化で大きく変化しつつある中で、ガバナンスの在り方の問い直しを行なっていくことも同時に重要である。

例えば、トラストを確保したデータ流通のためにプライバシーの保護は重要であるが、そもそもなぜプライバシーを保護しているか、そもそも個人の尊厳や自律や自由とは何か、というところから考えなければ、政策を変える方向性は分からない。

冒頭のカフェの無料Wi-Fiの例では、利用目的を通知してユーザーに同意してもらうという「Notice and Consent」を個人情報保護の大きな柱としている。しかし、そもそもなぜ通知して同意することが保護になるのだろうか。IoTが普及して無数のデバイスから個人情報が取得されるとき、認知限界を超えそうな数の通知をし続けても、保護に資するのだろうか。そういったことを議論し、トラストを確保する方法を再構成することが重要である。

技術の変化によってガバナンスの在り方も変わる必要がある。例えば、これまでは工場と自動車は別の安全管理体制を採っていたが、共通項を抜き出して管理することが可能になるかもしれない。さらに自動運転や空飛ぶ車のような新しい技術が今後普及したとしても、ガバナンスに関する基本要素をOSのように捉えて、その上にアプリケーションレイヤーとしてさまざまな規制手法が走る状態にすると、技術や社会の変化に柔軟に対応できるだろう。


**図2-12-6　現行法によるガバナンスと第四次産業革命時代のガバナンス**

最後に、「量や複雑さが増大するペース」「情報過負荷」について、一言申し添えたい。私たちが現在直面している課題の構造自体は、それほど新しいものではないと思う。

実は、人類は情報化に度々悩まされてきた。そしてその都度、さまざまな手法や技術を生みだし、情報を整理・管理しようとしてきた。例えば、中世後期から初期近代の西ヨーロッパにおいては、記憶術、文献管理術、章タイトル、引用文献一覧表示、索引、レファレンス書、百科事典などの技術が開発された[13]。そのような歴史を振り返ることは、トラスト研究の参考になるだろう。

そして、情報過負荷への対応は、人間の認識構造の変化にも繋がった。数量化と視覚化が普及したことによって、学術研究で、目的論や本質・本性の探求に加えて、定量性を意識したアプローチが登場し、客観性や普遍性といった概念の意味が重層化したといわれる（例：インドアラビア数字の導入、機械時計、グレゴリオ暦、定量記譜法、ルネサンス遠近法、複式簿記など[14]）。我々が直面している情報過負荷が「新しい思考様式」をもたらす可能性があるかもしれない。

**【主な質疑応答】**

Q：市民の政治に対する信頼を上げるための方策と、無関心の市民に関心を持ってもらうための方策は似ているようで異なると感じているが、政策レベルでどのように整理しているか。

A：無関心の市民に関心を持ってもらうことと、信頼のある政治制度を組み上げることは、おそらく違う。たしかに、アッカーマンのモデルのように、必要なときに、眠っている市民に目覚めてもらう（＝関心を持ってもらう）ことが必要だという議論がある。他方、市民の関心は希少資源でもある。つまり、人々の関心や認知を一種の資源（リソース）として捉えれば、それを各陣営が奪い合う関係にある。政治はその一つにすぎない。従って、少し異なる議論をしている。政治という複雑なエコシステムの一部の側面をそれぞれ捉えたものとして理解すればよいのではないか。

Q：「トラストの再構築」と言うときの「トラスト」が何を指すのか説明できなくて困っている。図2–12–6のガバナンスとトラストを置き換えても大きく間違っていないと考えている。現行法では、紙の書類や人の目視前提でトラストが形成されてきたのに対し、これからの時代では、何らかのデータやデバイスが自動的・自律的にトラストを形成することを想定していると思われる。トラストの再構築のために誰が何をやればいいかということに関しては、何らかのルールや制度を作る以外には何かあるのか疑問である。今日のトラストはFace to Faceではなくて、マルチステークホルダーのエコシステムの中でどうやって形成していくかが課題となっている。各ステークホルダーに向けてトラストを形成する取り組みを促すのは実現可能性が低いと思われる。ご意見を頂戴したい。

A：ガバナンスとトラストの関係は入れ子構造になっていて難しい。機械やAIが、紙や人間に代替できそうだというとき、技術的に代替できるかだけではなく、代替することが規範的に望ましいのかも問われる。その際、資格、点検、免許などの既存の仕組みが、何のトラストを担保しようとしていたのかという機能分析が必要になる。トラストの対象が、意図や意思を持つと擬制できる人間だけでなく、（開発者が背後にいるとはいえ）定義上は意思を擬制し難いAIが増えてきたときに、トラストをどのように分担するのか。または、AIを信頼するに値する存在にし得るには何が必要なのかの検討が必要となる。こういったことを進めるためには、国内に留まらず、グローバルな官民の協力・連携が非常に重要だ。市民の参画も必要である。

目標を策定する上で、グローバルにマルチステークホルダーを巻き込んでいかなければならない点は、ご指摘の通り難しいが、だからこそ、世界経済フォーラムのようなグローバルな官民連携プラットフォームが注目されていると思われる。そういうところで大まかな流れを作り、法令や省令を作れる政府と、技術開発をしている民間をうまく繋ぎながら、進めていくことが必要である。

Q：情報過負荷を技術で解消することは基本的に成功しないという感覚を持っており、だからこそトラストが必要だと考える。トラストは隅から隅まで分かるということを放棄した思考停止の一形態であるという定義から考えていくと、トラストの再構築が何を指すのかは必ずしも明らかではない。失われたのはトラストではなく安心であり、安心を再構築するための方策を考える上で、トラストをどのように定義していったらよいのかと考えるのが素直であろう。

A：技術によって情報過負荷が抜本的に解決することはないが、改善しようと努力してきた。これは現代において始まったのではなく、少なくとも中世後期以降から行われてきており、現代になって初めて直面した問題ではない。歴史に学ぶことは多いのではないか、というのが言いたかったことだ。

トラストは複雑性の縮減だとルーマンが言った。いろいろなものが視野に入ってきて捌ききれないという課題の解決に貢献するのがトラストだという点は同意する。判断の元となった情報を再検証したり、後で確認したりすることができるという建前の下、ある意味で思考停止し、目の前の自分がやりたいことに集中するために何かを視界から外すという仕組みの一つ。

こうしたトラストの側面が、私たちに与える影響を考えることも必要である。思考停止していることによって何かを獲得はしているが、逆に何が視界から外れていったのか。視界に入るものが技術によって

バイアスがかかっている可能性もある。こうした点を意識した上で研究活動や技術開発を進めていく必要があるだろう。

トラストの基本構成要素は、未来におけるポジティブな予測や期待と、何か害悪を受けるかもしれないと思いつつリスクを引き受ける意思であると考える。他方、安心は自分の振る舞いによって引き起こされることが予測可能な範囲に収まっている状態であり、従って、トラストの基本的構成要素における、未来における予測や期待がうまく機能していない状態と言える。

Q：Notice and Consentは最終手段である。個人がデータによって差別されることを防ぐということがプライバシーの目的であり、Notice and Consentをすればよいというわけではない。個人のデータ差別を防ぐためには何が必要か、あるいはそれを凌駕するための大きな目的は何かを先立って検討する必要がある。 EUの一般データ保護規則（GDPR）でも、6つある目的のうち、Notice and Consentは最後に言及されるものであり、日本はこれに頼りすぎる傾向が強く、そもそも出発点を間違えている。また、技術は社会や人間のあるべき姿を実現するためにあるという点に関して、あるべき姿の定義そのものが非常に不安定である。そのあるべき姿を議論するところからスタートしなければトラストの定義にはたどり着かないのではないか。

A：ご指摘に同意する。パーソナルデータの適正な取扱いを制度に落とし込むときに、ある意味で問題が矮小化され、Notice and Consentのみを行えば問題ないという考えが主流になってしまった。同時に、日本の個人情報保護法制においては、データによる差別やデータに関するフェアネスの問題は枠外に置かれている。差別の話は、アメリカやEUと比べてあまり議論されておらず、これから合意形成をしなければならない段階である。つまり、日本では、ゴールベースであるべき姿を考えようとしても、そのゴールが曖昧であると思う。

Q：図2–12–6が示唆することは、現行法によるガバナンスは国や機関のお墨付きに拠るのに対し、第四次産業革命のガバナンスは自主的な管理に向かうということでよいか。

A：近代国家論は、基本的に公私二分論になっている。しかし、現実には民間企業や大学などの中間団体の協力なくしてガバナンスは行き渡っていなかった。これがさらに前景化してきており、アメリカのビックテックの自主的な取組みなどがないとうまく運ばない。こうした点で、ガバナンスは中央集権ではなくネットワーク化している。国民と国家が対峙するだけでなく、中間団体に当たるさまざまな組織を介して処理していく必要性が認識されるようになってきた。中間団体のガバナンスに対する存在感が増してきたことが、第四次産業革命時代のガバナンスの背景にある。

Q：信託、DFFT、ガバナンスの関係をどのように考えればよいか。

A：大屋先生の講演に倣うと、信託はイングランド法に基づいて、本人、信託を託される人、その託された人から利益を受ける人、という三者関係によって構成されている。先の質問への回答で指摘したように、データ流通やデータ移転に関して、必ずしも政府が管理しきれない現代社会において、マルチステークホルダーでガバナンスの仕組みを再構築しようとするのがDFFTの発想である。

Q：トラスト、AI、ガバナンスの関係はどのように考えればよいか。

A：これらの関係も複雑に絡み合っており、ガバナンスによってトラストが促進されたり毀損されたりすることもあれば、AIによって一部代替することができたりもする。個人的には、トラストを担っている人間や制度といった主体から、機械に、どのようにトラストの機能を移行するのかを考えている。

Q：トラストがガバナンスに置き換えられた場合、ガバナンスを担う主体をトラストできるかという問題がある。政府と中間団体の利益をメインにルールを作ってしまい、きれいな言葉にだまされて、市民のデータが吸い上げられることを危惧している。どのようなガバナンスでその不安を解消されようとしているのか。

A：これからトラストとガバナンスを再構築する上で、さまざまな中間団体を巻き込みながらネットワークを作る必要がある。そこで重要なのは、ネットワークのレジリエンスである。ネットワークのどこかの

ポイントで、例えば、市民からのトラストが低下したとき、それが全体に波及して、ネットワーク全てのトラストが低下しないように、トラストのアンカーポイントを複数設定する。そにアンカーポイントを政府が独占するのではなく、市民社会側に担ってもらうことで対応するのがよいのではないか。

Q：個人をトラストできるかという問題をどう考えるか。嘘をついたり、詐欺をしたり、悪事を働く者もいる。中国のようなスコアリング社会は、個人を信頼しないことを前提に成立している。一方、日本や欧州のような民主主義国家で、個人をトラストできるかどうかを社会にどうやって組み込んでいけばよいか。

A：個人をトラストできないから、それを補う法制度がある。詐欺罪や損害賠償制度はその例。別の例で言うと、「近代的個人」は、自己決定できる合理的な存在であることが建前とされてきたが、消費者保護法などは、人間のだまされやすさや情報の非対称性を鑑みて、どのようにそれをケアするのかという観点から整備されてきた。法制度の建前を若干修正しつつ、信頼できない、誤りやすい、もしくは目先の利益に釣られやすい個人でも自己決定できるような配慮が組まれてきた。マルチステークホルダーのネットワークを考える上で、個人という要素が実態としては強くなってきており、そこをどのようにガバナンスするのかは、これからの課題である。

コラム1

## トラストのガバナンス

2021年には、トラストのガバナンスやAI（人工知能）・CPS（Cyber–Physical Systems）のガバナンスを取り上げた白書・提言（下記①～④）が立て続けに出された。

①Society5.0における新たなガバナンスモデル検討会（事務局：経済産業省商務情報政策局情報経済課）：「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2月） https://www.meti.go.jp/press/2021/07/20210730005/20210730005.html
②世界経済フォーラム第四次産業革命日本センター：「ホワイトペーパー：Rebuilding Trust and Governance: Towards Data Free Flow with Trust（DFFT）」（2021年3月） https://jp.weforum.org/whitepapers/rebuilding–trust–and–governance–towards–data–free–flow–with–trust–dfft
③Trusted Web推進協議会（事務局：内閣官房デジタル市場競争本部）：「Trusted Webホワイトペーパー ver.1」（2021年3月） https://www.kantei.go.jp/jp/singi/digitalmarket/trusted_web/index.html
④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：「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） https://www.jdla.org/download/sg01–report/

②では、「ガバナンス」を「ある主体（含む人、組織、システム）が、制度・規範やシステムなどの手段によって、ある価値を実現するために、他の主体（含む人、組織、システム）の行動を規律、もしくは方向づけること」と定義している。また、④では、「AIガバナンス」を「一企業や組織内におけるAI に関する安全性や信頼性確保の原則の整備や、開発や利活用における管理の実施」と定義している。なお、②③には「トラスト」の定義も示されている（表3–1–1の最後の2項目）。

デジタル社会においてうまく機能するトラストの仕組みを設計・維持していくため、上記①～④に示されているような考え方は参考になる。具体的に①では、環境やシステムの変化に対して、目指すゴールも変化していくことを踏まえ、「事前にルールや手続が固定されたガバナンスではなく、企業・法規制・インフラ・市場・社会規範といったさまざまなガバナンスシステムにおいて、環境・リスク分析、ゴール設定、システムデザイン、運用、評価、改善といったサイクルを、マルチステークホルダーで継続的かつ高速に回転させていくアジャイル・ガバナンスの実践」が提言されている。②でも、「マルチステークホルダーによって、ガバナンスプロセスをアジャイ

ルに実行していくことが不可欠」とされている。

ここでのガバナンスは、国や権力者が強制統治するようなものではなく、マルチステークホルダー間での合意・相互チェックのもとで、環境変化などに応じて適正なものに適宜見直されて維持されていくものと理解される。②では、トラストアンカーとなるものが社会に受け入れられ、トラストチェーンが形成されるために、トラストガバナンスが役割を果たすことが示唆されている。③においても、マルチステークホルダーによるガバナンスによって、トラストを裏付ける経路や連鎖を分散協業して支えるための機能が提案されている。④では、日本の産業界でよく見られるB2B2Cのように長いサプライチェーンの中で、ガバナンスエコシステムを作っていくことの必要性が示されている。

## 2.13 山口 真一[32]「ソーシャルメディアにおけるトラスト問題」

専門は経済学における計量経済学というデータ分析手法である。その手法を使って、ソーシャルメディア上のフェイクニュースや誹謗中傷、ネット炎上といった諸課題、あるいは、ソーシャルマーケティングや情報社会の新しいビジネスモデルについて実証研究している。本日は、ソーシャルメディアの諸課題とトラストの関係についてお話しする。

### 自由な情報流通とトラスト

ソーシャルメディアが社会に普及して、さまざまな恩恵を我々にもたらしてくれた。大きなものの一つに、非対面・対多数のコミュニケーションを可能として、誰もが世界に発信可能になったということがある。それまで不特定多数への発信は著名人とか一部のメディアしかできなかったことを考えると、これはまさに革命的な出来事であり、人類総メディア時代が到来したといえる。

ソーシャルメディアによる情報の自由な流通が、経済・社会システムを大きく変えつつある。ソーシャルメディアを通して議論が活発になって、時には政治的な動きにつながることもある。インターネットが普及した黎明期、特に2000年前後は、インターネットについてポジティブな意見が目立った。ところが、2000年代後半あたりから暗転する。例えば、ネットでの情報の選択的接触と極端化が社会の分断をもたらすとか、選択的接触は従来のメディアでもあったわけだが、インターネットは強いということが指摘された。

私の研究でも、ソーシャルメディア上では極端な意見が多く発信されるというようなバイアスがあり、社会の意見分布とは大きく異なる意見分布となっていることが分かった。こういったことから、フェイクニュースとか、誹謗中傷・ヘイトとか、社会の分断とか、情報の偏りとか、ありとあらゆる点においてソーシャルメディアやネットの負の側面が非常に多く語られるようになった。まさにこの負の側面というのが、今回のこのテーマ、トラストに密接に関わっている。

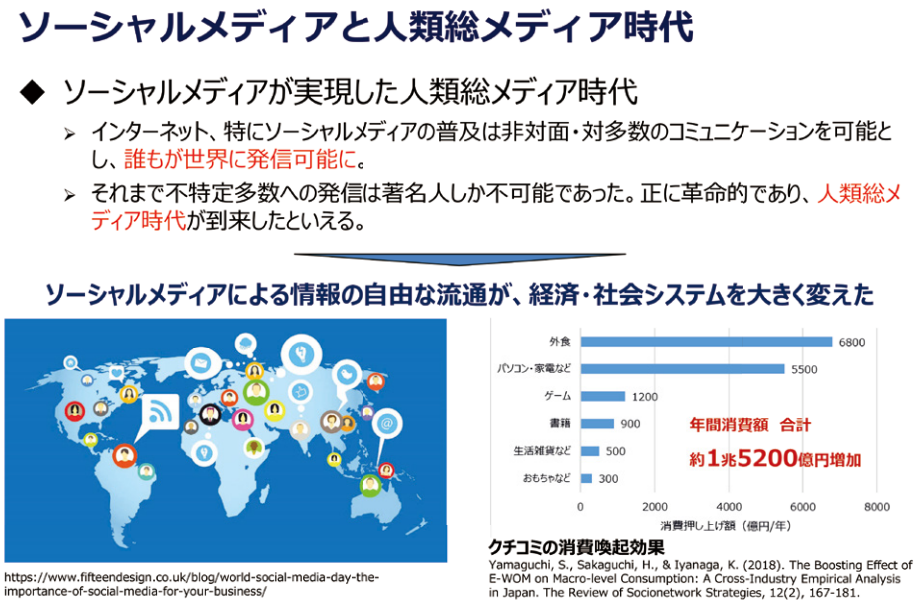


**図2-13-1　　ソーシャルメディアと人類総メディア時代**

32　国際大学グローバル・コミュニケーション・センター 准教授
https://www.glocom.ac.jp/researcher/301
東京大学 客員連携研究員、早稲田大学ビジネススクール 兼任講師、シエンプレ株式会社 顧問、など

**フェイクニュースと日本**

まず、大きなトピックとしてフェイクニュースを取り上げたい。それ以前からフェイクニュースは社会にあったが、フェイクニュースが極めて大きく注目された出来事が2016年のアメリカ大統領選挙だった。アメリカ大統領選挙以降も、イギリスのEU離脱やフランスの選挙といった政治的なイベントのたびにフェイクニュースが広がった。それだけではなく、ミャンマーのロヒンギャの弾圧のときに軍部が流したフェイクニュース、インドやメキシコで起こったフェイクニュースが原因となった殺人事件などが複数回起こっている。直近の2020年のアメリカ大統領選挙でもフェイクニュースの拡散が確認され、問題は全く収束していない。 Googleトレンドでも「Disinformation」は増えているし、「フェイクニュース」という言葉も微増とか横ばいである。フェイクニュースは社会に根づいていると言える。

フェイクニュースは欧米の印象が強いが、日本でも実はもう既に我々の身近にある。シエンプレ デジタルクライシス総合研究所の調査では、年間2,615件の疑義言説がソーシャルメディア上で拡散されたと報告された。1日平均7.2回の疑義言説が広まっていることになる。2020年は新型コロナウイルス関連のものが非常に多く拡散された。

政治でも、与党に有利なもの、野党に有利なもの、それぞれさまざまなものが拡散された。 Googleと実施しているInnovation Nippon2020[33]というプロジェクトでは、2019年度からはフェイクニュースについて研究をしている。2020年度は、新型コロナウイルス関連で10件、国内政治関連で10件、計20件のフェイクニュースについて人々の行動を調査、分析した。フェイクニュースへの接触率に関して分析した。新型コロナウイルス10件のフェイクニュースに1件以上接触した人が45.2%いた。全体を見ると、今回20件のフェイクニュースに関して、51.7%の人が1件以上接触している。つまり、2人に1人以上はフェイクニュースに接触していると言える。特に多かった新型コロナ関連では、ネットをよく使っている10代の接触率が50%以上と高いが、中高年以上でも接触率が40%以上と低くなく、年代にかかわらず接触していたことになる。

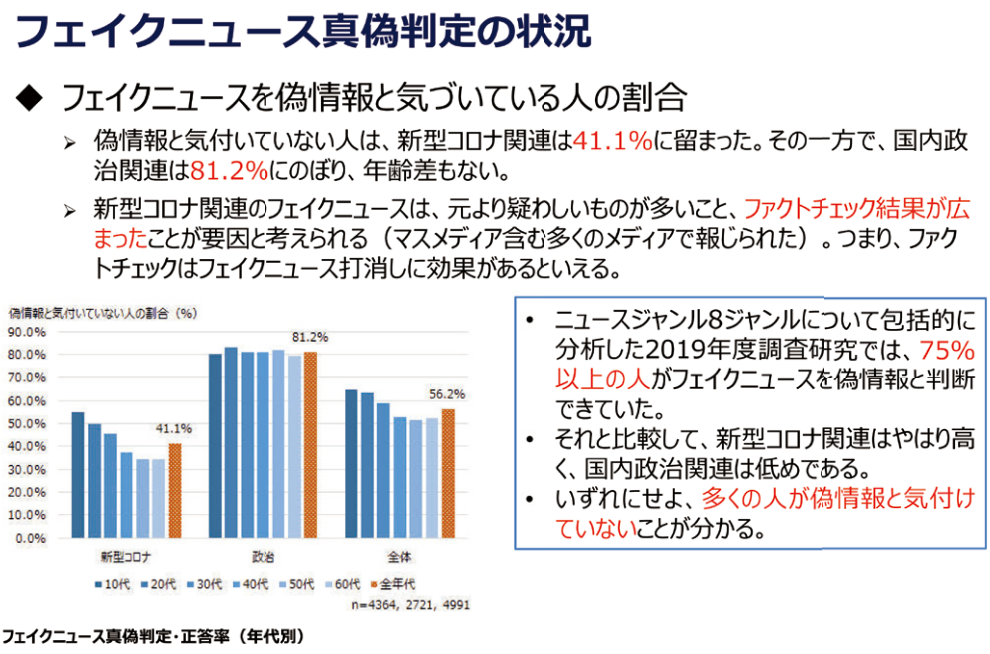


**図2-13-2　　フェイクニュース真偽判定の状況**

33 Innovation Nipponは、国際大学GLOCOMが、グーグル合同会社のサポートを受けて2013年に立ち上げた研究プロジェクトである。
http://www.innovation-nippon.jp/

さらに、接触した人はそれを信じているのか、偽情報と判断できているのかということを調査したところ、新型コロナウイルス関連のフェイクニュースについては、41.1％の人が偽情報と気づいていなかった（図2–13–2）。つまり、だまされていたわけである。さらに、政治のフェイクニュースになると81.2％の人がだまされている。5人に4人以上は少なくとも偽情報と気づけていないということが言える。

差がついた要因は2つ考えられる。一つは、新型コロナウイルス関連のフェイクニュースというのはもとより疑わしいものが多いことである。もう一つが、ファクトチェック結果がマスメディアを含む多くのメディアで取り上げられ、広まったことと考えられる。ファクトチェックが広まった結果、新型コロナは偽情報と気づく人が多くなった。一方で、政治のファクトチェックはテレビではあまり取り上げられなかった。このことから、ファクトチェックはある程度効果があるということが言える。

フェイクニュースの真偽判定に影響を与えている属性やリテラシーをロジットモデルで分析した。まず、リテラシーでは、情報リテラシーがフェイクニュース耐性に大きく貢献している。情報リテラシーとは、筆者の意見が入った文章かどうか分かるとか、文章から確実に言えることが何か分かるとか、読解力や国語力に近い能力である。そういう能力の高い人はフェイクニュースに対して耐性が高く、だまされにくいと言える。

属性については、まず、ネット歴の短い人とか自己評価の高い人はだまされやすいということが分かった。ネット歴の長い人はインターネット上の情報は玉石混交であるということが経験上よく分かっているが、ネット歴の短い人は、そういうことを知らずに、マスメディアと同じような感覚で接触してしまうため、だまされやすい。

また、ソーシャルメディアで情報やニュースに接触することは、フェイクニュース耐性を高めているということが分かった。多様な情報源に接触している方がフェイクニュースにだまされにくい。一方、ソーシャルメディアやメールに対して信頼度が高い人はフェイクニュースを信じやすいという結果になった。メディアによる情報の信頼性の違いに関しては教育が重要と考えられる。

最後に、マスメディアへの不満や自分の生活への不満が高いと偽情報と判断しづらい傾向が見られた。特に、マスメディアに不満が高いと、国内政治関連のフェイクニュースの判断能力を著しく低下させると言える。

フェイクニュースを接触した後の拡散行動を調査したら意外な結果が出た。図2–13–3の左側には、フェイクニュース接触後に偽情報と気づかずに拡散する行為をまとめた。「家族・友人・知り合いに直接話した」つまり、直接話すという行為が最も多い。メッセージアプリで伝えたのが次に多い。 TwitterやFacebookなど

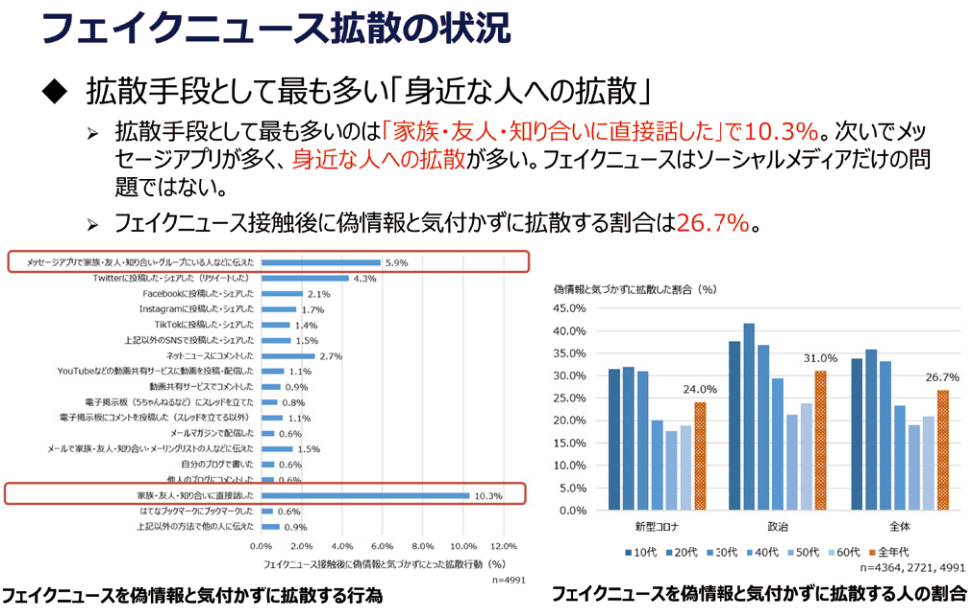


**図2–13–3　　フェイクニュース拡散の状況**

のSNSでの拡散もあるが、それよりも多いのは直接の会話とかメッセージアプリだった。ソーシャルメディアだけの問題というふうに矮小化してしまうと全体を見逃してしまう可能性がある。

同じ図の右側をご覧いただきたい。フェイクニュース接触後に偽情報と気づかずに拡散する割合は26.7%、4人に1人以上は、だまされた上で、拡散している。自分はだまされないとか、自分は拡散しない、と思うのではなく、身近なものであると考えるのが重要である。

拡散人数まで調査した結果を分析した。フェイクニュース20件に対して、ソーシャルメディア上で1万人以上に拡散したスーパースプレッダーは全体で0.62%しかいないが、拡散した人数でいうと95%以上を占めるということが分かった。結局、1%以下の人が95%ぐらいに拡散している。ワクチン関連の真偽不明情報、疑義言説を分析したレポート[34]によると、「Disinformation Dozen」と呼ばれる、12名のアカウントがワクチン関連のデマ投稿の65%を作成・拡散している。極めて大きな情報の偏りが存在する。

**フェイクニュースの社会的影響**

フェイクニュースは3つの社会的影響を持つ。1つ目が生活・経済の混乱である。「ライオンが動物園から逃げた」というフェイクニュースは生活を大きく混乱させる。また、ワクチンなど医療・健康系のフェイクニュースは命や健康に影響を与えるし、株価が暴落することもある。

2つ目が社会の分断の加速である。特に、政治的フェイクニュースの場合、誤った情報をもとに人々が対立してしまう。民主主義の議論には「共通の前提」が必要だが、その前提さえも変えてしまい、対話が困難になる。その結果、分断が加速する。

3つ目は情報の価値そのものの毀損である。インターネット上に例えば1%フェイクがあるとなったとき、我々はインターネットにある全ての情報を疑って見なければならない。これはインターネットに限らない。1%のフェイクが他の情報の信頼性を大きく失わせる。ソーシャルメディアは自由な情報の流通を実現して、これが大きな価値を生むと期待されていたわけだが、そのトラストが失われてしまう。さらに、フェイクニュースから生じた社会的影響を打ち消すには、社会が多くのコストを支払わなければいけない。

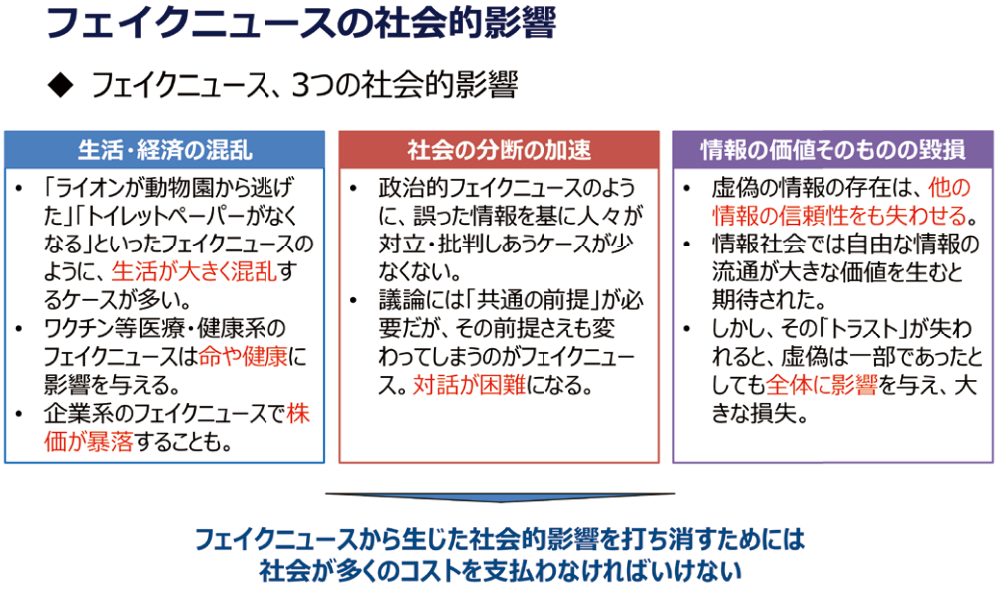


**図2-13-4　　フェイクニュースの社会的影響**

34 Center for Countering Digital Hate Briefing Note, “THE DISINFORMATION DOZEN”, https://252f2edd-1c8b-49f5-9bb2-cb57bb47e4ba.filesusr.com/ugd/f4d9b9_750e5af82aea4920a270b1c5a8b094c2.pdf

では、フェイクニュースに対するファクトチェックの効果に関する最近の研究成果を紹介する。ワクチンに関して、「菅首相が3月16日に打ったワクチンは偽物である」という話が広まったとき、それを肯定するツイートが圧倒的に多かった。1か月後にバズフィードジャパンがファクトチェックした。その後にそれについて言及したツイートの99%はファクトチェック結果を広めようとするものだった。また、ワクチンと不妊や流産するに関する言説に対して、河野太郎大臣がテレビで否定しブログで解説した。その結果、ツイート中で不妊を信じている人の割合というのは圧倒的に低下した。ファクトチェックの効果は少なからずあるということが言える。

**インターネット上の誹謗中傷の実態**

インターネット上の誹謗中傷という話をしたい。この話をするときに頭に浮かぶのは、2020年5月の木村花さんの事件であるが、誹謗中傷などがネット上に集まるネット炎上という現象は2020年には1,415件発生した。1日当たり約4件発生している計算になり、今日もどこかで誰かが燃えている、これがネット炎上の実態である。

さらに、新型コロナウイルスで自粛が進む中、2020年4月のネット炎上件数が前年同月比で3.4倍にも増えた。新型コロナウイルス関連の炎上もあるし、それ以外の炎上も結構あった。共通しているのが、新型コロナ禍においては通常では炎上しそうにないものまで燃えてしまっているということだ。

気がついてみたら、クラスターが発生した施設に対して誹謗中傷があふれたりとか、駄菓子屋に「コドモアツメルナ　オミセシメロ　マスクノムダ」というような怪文書を貼る人がいたり、新型コロナで「不寛容社会」が加速しているように見える。

その背景には2つあり、一つは、ソーシャルメディア利用時間が増える中で、不快に感じる情報と接する機会が増え、かつ、批判や誹謗中傷を書き込むが頻度が高まった。もう一つが、社会全体が不安に包まれるとネット炎上件数は増加するという現象がみられる。人々は、不安・ストレスを抱えているときに、悪者を見つけて批判するということで不安や無力感とかを解消しようというような動きが出るということが指摘されている。

実は炎上には良い影響もある。それは、企業などの強者の不正行為に対して消費者という弱者の声が通りやすくなった。例えば、ある企業の「高額解除料問題」では、認知症の方に大量のオプションを付けてタブレット端末の高額契約を結んでいて、家族が解除を要求したら20万円の解除料を取られたという話が大炎上した。犯罪ではないが企業の体質を問題とした。他にも、お節料理の通販に関しては、これまで泣き寝入りするしかなかった消費者が声を上げることができるようになって、それが大きく社会のうねりに拡大した。

しかし、悪い影響もある。ミクロ的には、炎上対象になった人の心理的負担増加とか社会生活への影響、企業であれば株価下落が起こる。よりマクロ的に考えると、人類総メディア時代になって、誰もが発信できるようになったが、気が付いたら、ソーシャルメディア上では、政治やジェンダーの話がしにくい、センシティブな話題ほどしにくい。攻撃的な人が自分を誹謗中傷するかもしれないからである。そうなると表現は萎縮せざるを得ない。その結果、人類総メディア時代の良い点がなくなり、不寛容な社会になる。

**ネットには極端な意見が表出しやすい**

表現の萎縮は2つの効果をもたらす。まず、企業や人々は、批判されにくい中庸的なサービス・意見しか展開できなくなる。長期的には、多様な好みを持つ消費者も、自分に合うものがなくなり、社会的公正が低下する。また、極端な意見の人のみがネットに残る。中庸な意見を持つ人は攻撃者が怖くてすぐ撤退してしまうが、極端な強い思いを持つ人は、逆側の極端な人から攻撃されても撤退しない。その結果、インターネット上は極端な意見ばかりになってしまう。谷型の意見分布になる。以前の調査で、ネットユーザーの75%がインターネットは攻撃的な人が多いと思うという結果が出ている。まさにそういう言論空間になってしまった。

実際には書き込む人は全体から見るとごくわずかである。年間1,400件くらいある炎上1件当たり、書き込んでいる人はネットユーザーの0.0015%、つまり7万人に1人しか、実際に書いてないということが分かっている。この話は有識者には研究する前から知られていて、例えば5ちゃんねる（旧2ちゃんねる）の元管理人

のひろゆきさんは、2ちゃんねるの炎上の主犯は大体5人以下という話をしていた。

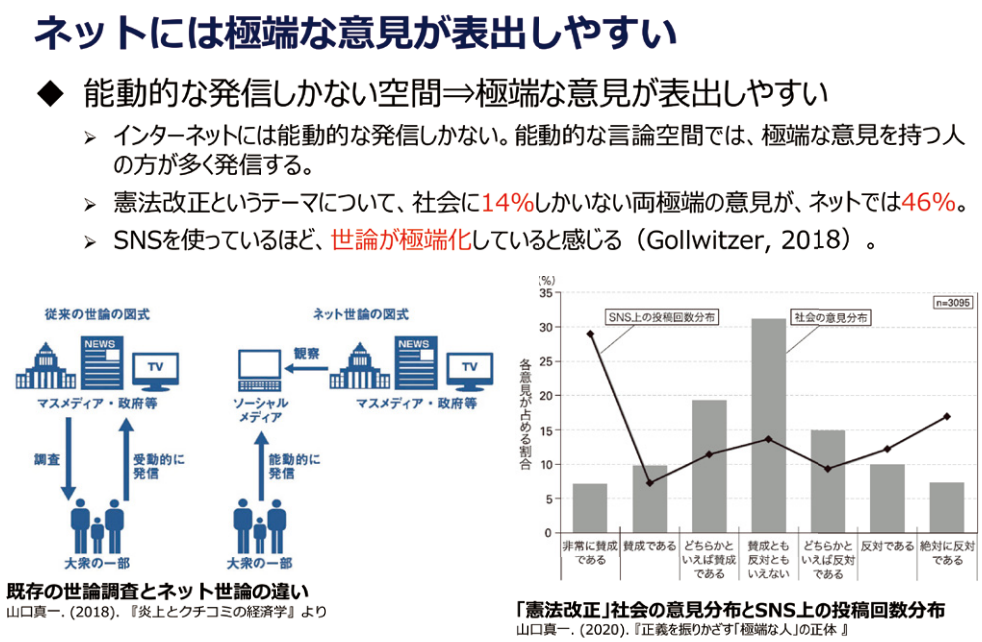


**図2–13–5　　ネットには極端な意見が表出しやすい**

世論調査とソーシャルメディアで決定的に違うところは、ソーシャルメディアは能動的な発信しかないのに対して、世論調査は聞かれたから答えるという形の受動的な発信（回答）しかない点である。図2–13–5の「憲法改正」について調査した結果をご覧いただきたい。「非常に賛成である」から「絶対に反対である」までの7段階で調査した。棒グラフは世論調査における社会の意見分布、受動的な発信である。山型の意見分布になっている。一方で、ソーシャルメディア上で投稿した回数を分析すると、最も多くなったのが「非常に賛成である」人の投稿で、次に多くなったのが「絶対に反対である」人の投稿だった。意見分布ではそれぞれ約7%で合わせても14%しかいなかった。ソーシャルメディア上の投稿回数では46%と約半分を占めている。

炎上に書きこむのは、特別な人ではなく一般的な人であることが分かっている。炎上参加者の肩書分布を調査すると、主任・係長クラス以上は31%、一般社員30%、個人事業主・店主9%、無職・主婦・バイト・学生30%である。炎上非参加者では、主任・係長クラス以上は18%しかいないので、役職を持っている人は炎上に書き込みやすい人である。世帯年収も高く、男性であるという特徴もあるが、特別な人ではない。さらに、その動機を調査したところ、「間違っていることをしているのが許せなかったから」とか、「その人・企業に失望したから」といったような、自分は正しくて相手は間違っているという正義感から書き込みをしているということが分かった。しかしながら、ここでいう正義感は社会的正義ではなく、一人一人の価値観における正義感である。リンチと変わらない。以上がフェイクニュースと誹謗中傷の話である。

**ソーシャルメディアにおけるトラスト問題への社会的対処**

社会的対処と情報社会の未来について簡単にお話ししたい。まず、インターネットの実名制には効果がないといわれている。韓国で導入したことがある。その後の研究で、一般的な書き込みは減少したけども、誹謗中傷的な掲示物の割合は変化しなかったということが分かった。自分が正しいと思っているので、実名制が導入されても特別書き込みをやめることはない。違憲判決が出て2012年には廃止された。

よくフェイクニュースや誹謗中傷には法規制をもっとしろという意見が出る。調査すると約75%の人がそう

いう意見を言うが、これにはリスクがある。Slippery Slope（滑り落ちる坂）という問題がある。類似した行為が連鎖的に行われ、だんだんと道徳的に許容できない行為がなされる現象のことを指す。法を施行してすぐのときにはうまく運用されるかもしれないが、10年、20年後、30年後、法律がどんどん拡大解釈されていって、やがて政権に批判的な情報を手当たり次第に取り締まるという可能性もゼロではない。社会全体に影響を及ぼすような法律は、20年後、30年後を考え、慎重に検討する必要がある。

プラットフォーム事業者を取り締まればいいのか。ドイツのネットワーク執行法では、侮辱などの違法な内容があるとユーザーから報告された場合、直ちに違法性を審査し、違法なものは24時間以内に削除する義務を、大手プラットフォーム事業者に課している。対応が不十分な場合、最高5,000万ユーロまでの過料が科せられる。これにも2つ問題点が指摘されている。多額の過料を逃れるためにプラットフォーム事業者が、自由な言論の場という立場よりも違法性を安易に判断して、過剰に削除する可能性が指摘されている。また、プラットフォームの多くは海外のプラットフォームであるから、他の国の一企業のAIやスタッフが違法性を検証して勝手に言論を削除してもいいのかという危険性もある。表現の自由の観点からリスクがあると言える。

では、どうすればいいのか。被害者に寄り添う法律が重要だろう。現在は、被害に遭った方が圧倒的に不利である。匿名の攻撃者を特定するための発信者情報開示請求を簡略化しようという動きが「プロバイダ責任制限法」の改正につながり、2021年秋に施行される予定である。これにより2回必要だった手続きが1回でできるようになった。

重要なのは自主的な対応の促進と透明性の確保である。プラットフォーム事業者には、原則として自由な言論の場を提供するという立場は変えない。同時に、規約違反を理由に対処するという姿勢、これは今後も維持することが求められる。 Donald TrumpがTwitterを追放されたのも利用規約違反だった。同時に客観的な検証を可能にするための透明性の確保は重要である。どういった基準でどう対処したか、年間どれくらいあるか、というようなローカルな情報が分かるのが望ましい。忘れてはいけないことは、フェイクニュースは社会にずっとあったものである。規制して根絶するのではなく、社会的影響をいかに弱めるかということが重要である。実際、プラットフォーム事業者は、例えばアーキテクチャーを工夫するなど、いろいろな対応をしている。ブロック、ミュート、返信制限とか、非表示ワードを設定できるようにするとか、リシンク機能といって誹謗中傷を投稿しようとしたらアラートが出るといった機能を実装している。

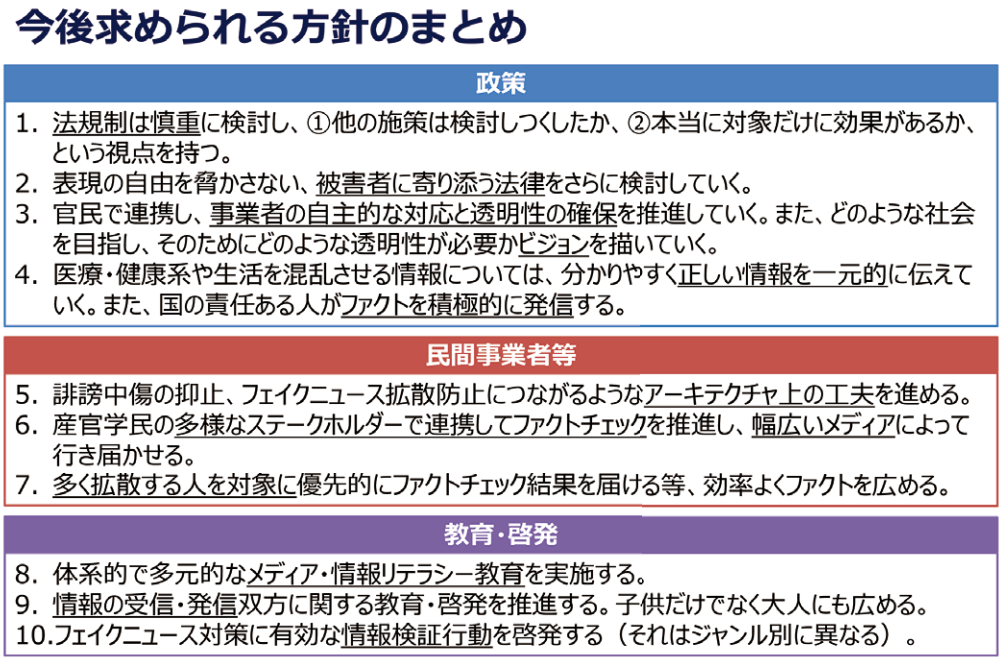


**図2-13-6　　今後求められる方針のまとめ**

我々生活者の側も考える必要がある。それは、誰でも誹謗中傷の加害者になり得るし、誰でもフェイクニュースを拡散する可能性があるということを忘れないということである。発信の際にも、ネット上の言葉遣いも良識に従うとか、発信するときに一呼吸置くとか、真偽不明な場合には拡散しないということを意識しておく必要がある。また、受信の側では、情報は偏っているということを忘れないとか、エコーチェンバー、フィルターバブルによって自分の見たい情報になっているかもしれない、そして、受信している情報にはデマも含まれると認識することが重要である。子供だけではなくて中高年以上も含めた人類総メディア時代だからこそ、人類全員がこれを認識する必要がある（図2–13–6）。

**情報社会のこれから**

情報社会の未来の話でしめくくりたい。より中長期的なマクロ的な視点で、近代化の歴史の中でネットの課題と情報社会を考えてみる。産業革命が1760年ころに起こって経済の発展が加速した。図2–13–7の右のグラフをご覧いただきたい。西欧における1人当たりGDPは産業革命から少し経って崖のようになっている。まさに産業革命の効果でそれまでとは異なるペースで経済が発展した。パラダイムシフトが起こって新しい時代が誕生した。産業社会においては、物の豊かさや富を築くこと、所有することが重視された。



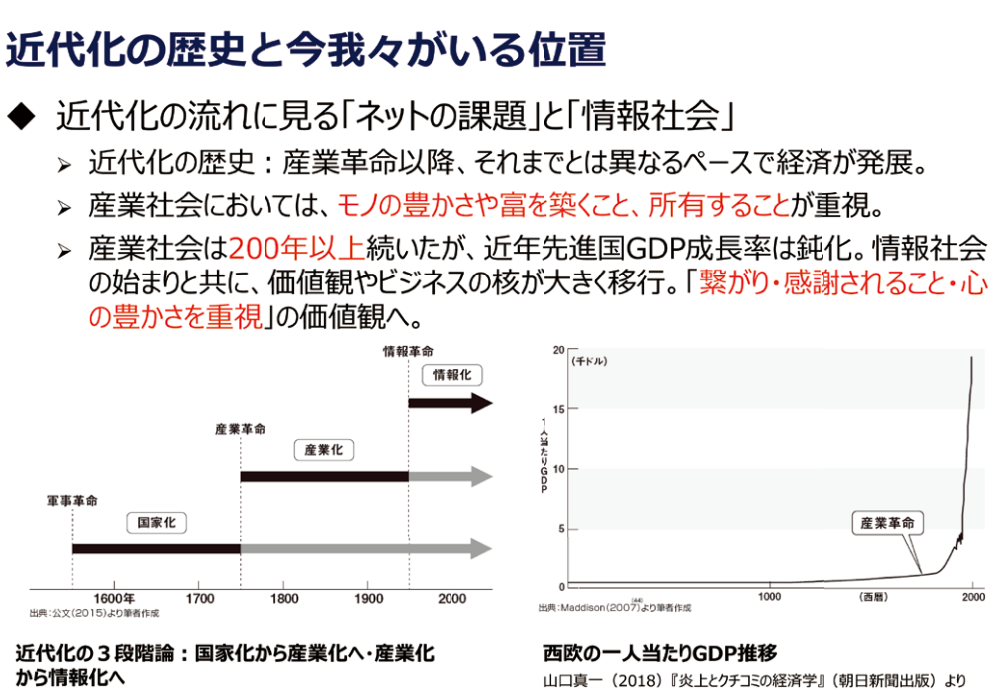


**図2–13–7　　近代化の歴史と今我々がいる位置**

産業社会は200年以上続いたが、近年、先進国のGDP成長率は鈍化している。情報社会が始まり価値観やビジネスの核が大きく移行してきた。所有することが物の豊かさで重視されていたのが、つながりや感謝されることや心の豊かさなどを重視する価値観になった。価値観が徐々に変化しビジネスも大きく変わっている。新しい時代、情報社会の誕生である。産業社会が200年以上続いた。情報社会もこれから200年続く可能性が高い。

インターネットが普及して数十年しかたってない。世界のデータ生産量もまだまだ指数関数的に増加している最中、まさに黎明期の動きである。時代の黎明期というのは多くの問題が発生する。産業社会もさまざまな問題を解決して、今のこの豊かな社会がある。情報社会も人々自身が問題を解決しながら、大いに発展し

ていくだろう。

私が今特に考えているのはソーシャルメディアである。ソーシャルメディアは、情報社会の変革の始まりに誕生したサービスにすぎない。今後、人工知能やデジタルツイン技術などにより、人の内面をデジタルで共有することが普及すれば、言語や画像でコミュニケーションを取るより圧倒的に濃いつながりが広がると予想できる。問題を乗り越えることが人類の進化である。ネット社会の未来は暗いというスタンスはとらない。

一つ忘れてはいけないのが、経済の自由で産業社会が発展してきたというのと同じように、情報社会においても表現の自由というものを保障したまま発展してくということが重要であるということである。そして、誰もが自由に発信できる時代だからこそ、他者を尊重する、自分がされて嫌なことを相手にしないというような当たり前の道徳を忘れないことが、今後、人々には求められていると考えている次第である。

**【主な質疑応答】**

Q：「今後求められる方針のまとめ」にある民間事業者などに期待する項目（5～7）に、プラットフォーマーはコストをかけるか疑問である。

A：企業側はしっかり進めていると私は理解している。理由は批判が大量につくからである。法規制に対して先手を打っている。サービスの差別化にもなる。しかし十分ではないので項目3「事業者の自主的な対応と透明性の確保」が必要である。

Q：情報リテラシー教育のアイデアとして、人の受信力はどうか。

A：面白い。情報の受信のところにはまだ力が入ってない。今、総務省の人とどう啓発すればいいかという話をしているところなので、ぜひ生かしていきたい。

Q：災害時になぜみんながリツイートするのか理由を調べたことがある。人の役に立ちたいという善意で、真実でないことを発信してしまうという人がすごく多かった。

A：社会学的にも利他的な動機からフェイクニュースを拡散してしまうことが多くなるといわれている。私の調査でも、拡散の動機に、怒り、不安に次いで、3番目に利他的な動機があった。なお真偽不明情報は拡散しないための教育が重要である。

Q：人の行動の背後にある心理的な影響や、そういった心理的なものを背景にした対策みたいなものが必要と思うが、既に検討されているのか？

A：フェイクニュースの拡散は、怒り、不安、人の役に立ちたいという動機が多い。感情的になったときほど、それを忘れないでという教育・啓発は必要と思う。

C：対策を個人のリテラシーに任せるのでなく、熟議を促すとか、一緒に考えるというようなところがソーシャルメディアで強化されると良い。

Q：対立しているところにいくらファクトを示しても逆効果になる。一緒に考えるような形のメディアなりプラットフォームになる必要があるのではないか。

A：陰謀論を信じている人に、その逆を示すと余計強固に信じるという傾向があるといわれている。否定せず、根気よくコミュニケーションを続けるのが良いと医者も言っている。メディアとかソーシャルメディアも何かできると良いかもしれない。

Q：国際的な対応や研究について伺いたい。 EUは、データ保護では世界を主導しているが、今日お話しになったような点に関してはどう対応しようとしているのか？

A：欧米では、ネットの誹謗中傷とか炎上に関する研究は活発ではない。ネットいじめの研究が活発である。日本的なネット炎上の研究はアジア圏が主流である。そもそも、誰かが何か悪いことをしても、それを集団でたたくという文化があまりない。一方で、フェイクニュースについては、圧倒的に海外の研究が多い。社会学ソーシャルメディア研究、経済学など、分野を越えた研究がなされている。

Q：アメリカのカリフォルニアとテキサスで、選挙という文脈の中でフェイクを禁止するという話を聞いたがそれはどうか？

A：選挙という期間に限定して対策しようという動きは各国で既に出てきている。韓国の実名制は違憲ということで廃止となったが、実は2019年の選挙期間中は復活していた。日本も今後参照することがあり得るかもしれない。

## 2.14 尾藤 誠司[35]「医療におけるトラスト（1）」

**医療における安心・安全と信頼**

新型コロナウイルス感染症のパンデミックは、医療における「知る」ことや意思決定をすること、社会と専門的な発信情報との信頼関係において大きなインパクトを与えた。例えば、デマも含むあらゆる情報がSNSに拡散され、人々が右往左往している。専門家セクターに属する者は、この情報は信用すべきでないとか、もうこんなことが起きているからこうすべきだ、という認識が多少一致しながらある。しかし、人々が情報をどう受け取るかというと、サイエンスに常につきまとう真実の曖昧さを受け入れきれずに、拒否してしまったり受け入れすぎてしまったりする。この認識の差により、それぞれの社会のクラスターの中で生まれた共同幻想の陣地取り合戦のようなものが起きていて、新型コロナウイルス感染症における情報の錯綜、それに基づく勧告、推奨、正義の混乱を見せている。

このような状況で、安心・安全という言葉が最近もよく聞かれる。安心・安全の医療という言葉は、医療の世界では30年ほど前から声高にいわれているが、覚悟がない、無責任な言葉だと思っている。安心・安全は、社会そのものが安定しているときにあるものだ。例えば、車の運転中に青信号になったら、私と、私に害を及ぼす他の車との関係性や、その人たちの運転が下手か、私を殺そうとしているかなどと考えずに、ただただ青信号を突っ切る。これは安心しているからであり、すなわち、考えていない。

一方で医療には、未来の不確実性が必ずついてくる。しかも、個別性が非常に強い。個別性に基づく不確実性があってこそ、クライアントに対してなるべく利益を与えて不利益を最小にしていくのが、医療がやっていく一つの覚悟性だ。医療や、新型コロナウイルス感染症のように今まで社会に起きたことがないことに関しては、未来は分からないはずで、安心という状態はない。それではそこに何が必要かというと、信頼が必要だと思う。

予測できない未来は人を不安にさせるが、IF/THEN構文（もしこういう状況がそろったらきっとこうなる）のような確定的なことが分かれば、安心材料が増えていく。自らの不安をどのように手なずけながら危険な状況を歩いていけるか、というのが個人レベルでの心象風景だと思うが、ここに医療が付け込んでいるところがある。不安な未来、厄介な未来しかないのが医療の本質である中で、IF/THEN構文を用いて適当に絡めとるなどしている。もちろん医療者に悪意はないが、結果的にそういうことをしているのが医療というビジネスだと思う。

**「ディストピア」としての病院　―トラストを強制する場―**

病院は一種のディストピアであり、トラストを強制する場である。患者の健康に関する情報は専門家である医師の方が圧倒的に知っていて、それに対するアンサーも医師が圧倒的に持っているということを、患者は病院に入った途端に刷り込まれていく。

医師が提供して、患者がそれを信じてしまうものがいくつかある。一つは、患者の今の状況について。医師の言語ではそれを診断という。次に、今後の見通しについて。医師の言語では予後である。そして、厄介な未来を回避するために医師が提供できる選択肢について。さらには、その中でのベストな選択肢について。これらを無意識に、さもサイエンティフィックな感じで提供するのだが、そこにはサイエンス＝真実であるというマジックのようなものがある。さらには、サイエンスが持つ不確実性をうまく表現できないまま、例えば、「このままだと脳梗塞になりますよ」という話をする。しかし、血圧が高くてコレステロールが高い人に、このま

**35** 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
http://umakara.net/
https://note.com/bitoseiji/

まだと脳梗塞になりますと言うときには、10年のうちに脳梗塞になる確率が8%くらいのことを言っている。薬などをたくさん出せば8%が4%になる、という程度だが、これがサイエンスという衣を着ると、このままだと死んでしまうが薬を飲めば助かる、という魔術のようなものに取って代わっていく。こういうところが医療情報が持つ邪悪な部分だ。また、単に医療情報を発信するだけでなく、内科医である私がこのトーンで話をするので、そこにも罪作りな部分があると思う。

例えば、診断のときに「風邪ですね」と言うのだが、実は風邪を診断することはほとんどできない。なぜかというと、風邪を診断するための検査キットがない。咳が出て鼻が出て体がだるい人に、肺炎ではない、肺がんでもない、心不全でもない、COVID-19でもない、多分このまま治るだろう、というときに「風邪ですね」と言う。緻密を極めれば、前述のようなことを言うわけだが、そうすると患者はより混乱してしまう。ここで安心のロジックを用いて、本来ならば個別には将来が読めないが、ある程度そういうものだと錯覚させるように「風邪ですね」と言う。この安心のロジックに対して、患者が信頼するというからくりがある。

また、見通しとして「きっとよくなります」という言葉もよく使う。本来は不確実なものを、ある程度確実だろうと専門家の解釈の中で見通してこのような言葉を使う。「分かりません」というのが反対言葉だ。自分のエピソードとして、大学生の時にテニスボールが目に当たって目が見えなくなったことがあった。今は眼科学の知識があるので1か月後には治るものだったが、大学病院の助教授にまた見えるようになるか聞いたら、「分かりません」と言われてとても不安になった。ここで「大丈夫ですよ」とも言えるのだが、サイエンスとしてより正確な情報としては、「分かりません」なのだと思う。ここをどのように丸めてしまうかは、専門家にかかっている。「きっとよくなります」と言ってしまえる人の方が信頼は得られるが、こういうことを平気で言ってしまう人は、信頼を得ていくところに少々詐欺師的な要素があると思う。



**インフォームドコンセント（Informed Consent）**

帽子屋で帽子を買うときに、基本的に帽子屋には任せない。大体自分で分かっているからである。そして、帽子を買うときに、自分が持つ帽子に対する価値の方が、ディシジョンメイキングにおいて偉いと思っているからである。

一方で医療においては、医療者は患者のインフォームドコンセントを受領するプロセスを必ず踏むように言われている。インフォームドコンセントは、同意能力があることが確認され、専門家から説明を受け、その理解を確認し、理解が確認された上で当事者が最善と考えるチョイスを提示する、という患者の主体的な行為である。

しかし、インフォームドコンセントのプロセスをそのまま行っても大概うまくいかない。専門家としてAかBどっちにしますかと聞くと、大体、お任せしますと言われてしまう。お任せしますでは困るのだが、お任せしますでは困ると言う医師も、専門家としての覚悟性を放棄してしまっている可能性がある。

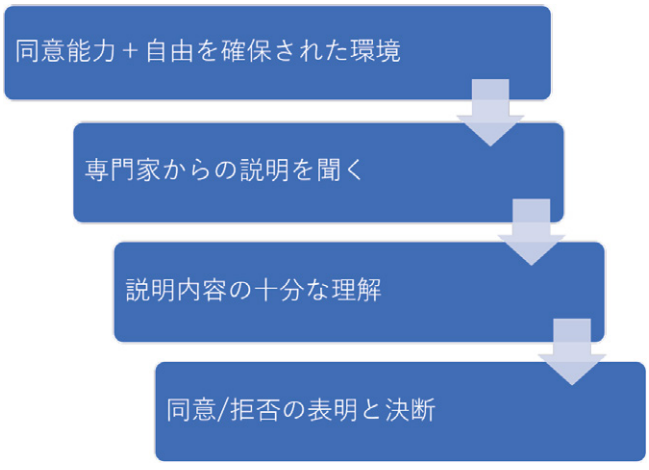


**図2-14-1　　インフォームドンセントの4つのプロセス**

医療の中でパターナリズムという言葉がよく使われる。パターナリズムとは、父親が子供に対して、お前はきっとこうした方がいいに決まっている、とおせっかいを焼く態度のようなもので、通常のサービスにおいては良くない態度といわれている。専門家が持っている価値をいかに当事者に押し付けずに、説明的関係を保ちながら当事者の考える選好を重視していくか、ということが重要視された時代が40年ほど前にあった。医療者はインフォームドコンセントのプロセスを型通りに進めることで、実は医療者が決断に関与しようとする姿勢そのものが、だんだんなくなってきたように思う。

**決断のプロセスにおける専門家への信頼**

情報を得てそれを理解し、決断する、という決断のプロセスは、インフォームドコンセントの4つのプロセスのうちの3番目と4番目だが、本当はこの理解と決断の間にもっといろいろなことがある。当事者である患者が専門家から、あなたの病気はこういうもので、このままいくと健康に対するこういうアウトカムが待っているだろう、こういう選択肢があるがどうするか、と聞かれて、患者がじゃあこうします、というのがインフォームドコンセントのプロセスだ。しかし、人間は自分にふりかかるリスクに関してそんなにシンプルに決断できない。理解した事実を自分のこととしてどう認識していくか。その上で、自分はこの医療に関する決断において何を叶えようとしているのか。さらには、叶えたいことに対して、自分が大切にしている他者と相談し、内省するなどし、そこには葛藤がある。この4つのプロセスをぐるぐると回っているのが通常の意思決定だと思う。

専門家は無意識にこのプロセスに関与しているのだが、インフォームドコンセントではこのプロセスはオミットされている。インフォームドコンセントは、患者は自律的な存在として立っていて、他者からの干渉を受けずに合理的な判断が可能であるという前提の意思決定のモデルであり、私はその前提に対して異議を唱えている。主体的な一人の人間が、他者との関係性を分断し情報だけで合理的に選択して決めるというプロセスにおいて、不安という感情や不確実性、自分に対するリスクが大きいとき、個人の輪郭はおそらくにじんでいく。このような中でいかにセルフを尊重していくか、自律性を尊重していくかというのが、トラストと関係していく上での意思決定のプロセスだと思う。

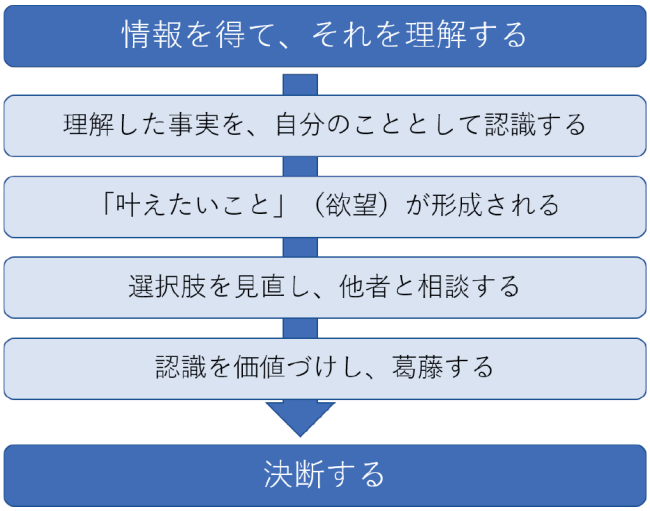


**図2-14-2　　私のからだに関する決断のプロセス（仮説）**

**プロフェッションと信頼**

当事者であるクライアントから専門家への信頼は、専門家が持つ能力と意図の2つの要素からなる。能力と意図がそろっているからこそ、集団や、クライアントである個人は、専門家を信頼する。プロフェッショナル集団が集団として一般的信頼を得るためには、いくつかの要件が必要といわれている[1]。

しかし、信頼は一般的信頼だけでは成り立たない。医師ならきっと私に悪いことはしないだろうという「一般的信頼」の上で、東京医療センターの内科医である尾藤誠司という医師に対する「人格的信頼」、さらには、自分がクライアントになったときに、東京医療センターの尾藤誠司がこの症状に対して大丈夫ですよと言うときの関係性も持った「関係的信頼」の三段階の信頼がある。

また、先ほどの「風邪ですね」「きっとよくなります」のような言葉が多用されることにより、クライアントにどっぷりと信頼されてしまう「酩酊」のような副作用もある。多少酔っ払っている部分があるからこそ、専門家から発せられたメッセージを無条件に受け取ってしまう。酔っ払わせているだろうなと思いながら、専門家として「大丈夫ですよ」と言う、こういうところが専門家の専門家たるテクニックであるし、覚悟なのかもしれない。

**近未来の当事者―専門家関係はどう変わるか？**

15年後の診察室は患者－医療者の二者関係ではなく、患者－医療者－情報技術の三者関係に間違いなくなるだろう。

今、患者は医療者を「情報のサイエンスを持っていて、能力があって、きっと正しいことを言っている医療者」として受け取っているが、情報技術も含めた三者関係になると、例えば、患者が登録している“Chroniccondition.com”というデータベースからこう言われたのだが、医療者と情報どちらを信頼していいのか分からない、という状態になりうる。例えば、患者が“Chroniccondition.com”から「あなたには高血圧があります。この血圧ステータスのままでいた場合、10年以内に脳卒中を発症する確率は2.76%です。」という情報を受け取ると、びっくりして医療者のところに行く。すると医療者が「大丈夫ですよ」などと言ってくれるのだが、そこは正確な情報、能力などと、医療者が専門家として持っている覚悟性とのバランスなのかなと思う。

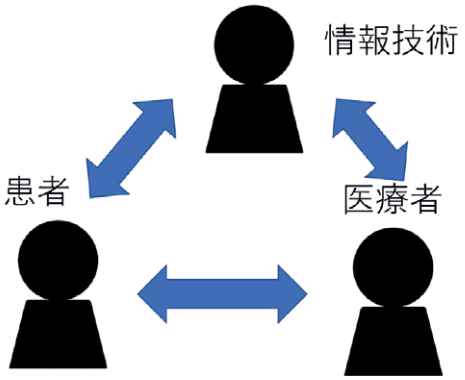


**図2-14-3　　15年後の診察室**

一方で、医療者と情報技術も裏で内通している。裏で内通しながら、患者－医療者間でどういうやり取りをしていくかがテーマとなる。私は、真実というものがあるとすると、真実の度合の怪しさも含め、医療者が専門家として患者にメッセージしていたものを情報技術にアウトソーシングしていく方が、専門家の持つ邪悪性のようなものを少なくできると思う。

患者と医療者の間にどのような関係性をもたらしていくべきか。私は最近、感情端末としての専門家と呼んでいる。情報端末と感情端末、この二者がある程度裏で内通しながら、クライアントである患者がそれぞれにアクセシビリティーを持って、どううまくやっていけるかがこれからのテーマと思っている。

**【主な質疑応答】**

Q：コンピューターサイエンティストのトラストの研究者の間では、リスクのあるところにトラストが必要といわれている。社会学者の山岸俊男も、社会的不確実性が高いところに信頼が必要で、不確実性がない、つまりリスクがないところに安心があると捉えていた[2]。

A：山岸の信頼の構造は共感を持って読んだ。安心は人を考えなくさせる魔術で、信頼は、怖さや不安を持ちながら信頼しているというのが私の解釈だ。

Q：患者、情報技術、医療者の三者関係でコミュニケーションするときに、インターネット上には怪しい情報もある中、患者が受け取る情報を医療者側がコントロールする必要があるか。

A：医療情報は玉石混交だが、専門家セクターが信頼するような、いわゆるサイエンティフィックなエビデンスに近い情報はどこにあるのかというのは、20年前から大きなテーマである。例えば日本医療評価機構が、ある程度信頼できる診療ガイドラインを掲載するなどの取り組みを行っている。しかし、専門家セクターの情報は一般論としてのエビデンスであり、患者が知りたいのは非常に個別的なことなので、限界がある。個別化された情報をPersonal Health Recordとしてクラウドに載せることが制度化されれば、医学的な妥当性が立証された個別化情報を患者に提供できると思うが、プライバシーやセキュリティーなど、非常にハードルの高い問題が存在すると思う。

また、その情報にアクセスできたところで、解釈の問題がある。例えば相対リスクという疫学用語があるが、相対リスクが半分になるという言葉を医療者が使うとき、死ぬ確率が半分になるということもあれば、コレステロールを高いままにしておくと1万人中80人脳梗塞になるのが、薬を飲むと40人になるということもある。これも半分になっているが、助かるのはせいぜい1万人中40人である。このような解釈の部分は埋められないままでいる。

さらに、正しい情報にアクセスするためにはバイアスがかかってはいけないのだが、自分が心配なこと、助かりたいことが書かれている情報を追い求めるのは避けられないので、多くの場合失敗してきている。

Q：トラストは高ければ高いほど良いと思っていたが、酩酊もあるとすると、患者－医療者間のあるべき信頼関係とはどのようなものか。例えば、専門家に対するトラストの要素である意図と能力のうち、能力の要素を高く持っているなど。

A：意図が入っているかいないかが、コンフィデンスとトラストの違いである。専門家は、クライアントを良くしようという意図が強く、コンフィデンス部分よりトラスト部分が強い。そこに悪意があるかどうかはそれほど難しい問題ではなく、悪意がある専門家は、専門家集団の中で自浄すべきと思う。問題は、パターナリスティックな人間は善意で言っているということだ。善意で、勉強しろ、この塾に入れ、この薬を飲め、と言うわけだが、その善意は必ずしも当事者の叶えたいことではない可能性が高い。当事者にとって叶えたいことに基づいた善意であるというのが、トラストを受ける側のテクニックであり義務であるが、メディカル、特に専門家セクターになればなるほど、その部分が欠落していくというねじれの構造にある。専門家セクターになればなるほど、どれくらい死なないか、どれくらい脳卒中にならないかという、サイエンスとして妥当なことが良いことであるというロジックが成立するからである。しかし、私が母親を見送ったとき、母親が死なないことや脳卒中にならないこと、血糖値がコントロールされていることより、人としての尊厳がいかに保たれ、平和に亡くなってくれるかを望んでいた。しかし医師と話すと、死なないことを善意で押してくる。しかもありがたい感じで言ってくれるので、無下に断るわけにもいかない。さらに、その善意がキラキラしていると信頼してしまう。お医者さんがあんなに一生懸命言ってくれるのだから、やはり任せなければとなってしまう。

患者－医療者の二者関係の中では、どうしても酩酊のマジックが増幅し、酔っ払うほど信頼しすぎて自律的な意思決定を曇らせてしまう。そこに、少しクールダウンしてくれるような、さらに、この専門家は患者に対して悪意があるわけではないことも分かってもらうような第三者セクターがあるといいのではないかと思う。

Q：専門性を要することはよく分からないため、酩酊して医療者に頼りたいと思ってしまうが、患者当事者も十分に理解をしていなければならず、セカンドオピニオンを聞くべきかなど重要な判断も必要である。患者にも、クリティカルシンキングという、日本人には今までなかった習性が求められるか。

A：先程のようなガイドラインをしっかり読めることや、受け取った情報がエビデンスに基づいているか、そのエビデンスがどれくらい自分に恩恵をもたらしてくれるかについて、当事者はクリティカルな思考を持つべきであるという啓発活動が20年ほど前から行われているが、少なくともヘルスケアの中では失敗してきている。正論ではあるが、やはりそんなに物事を合理的に考えられないし、不安があるからこそ、その不安をかき消すために情報に入っていく。不安をかき消したいという感情をなくしてクリティカルに情報を見るのは非常に難しいと思う。

意思決定のときに大事なことは、とにかく安心しないことだ。安心・安全文化の一番の問題は、安心・安全を押し付ける主体者が、「考える」ということをやめさせようとすることである。いかに当事者に不安であり続けてもらうか。そして、当事者が、私は不安である、何が不安なのか、いつの時点でどうなるのが不安なのか、こういうところが言語化され、それを聞いてもらいたい人に理解してもらいながら、より安全な状況を作っていくために対話するというやり方が現実的にはうまくいくのではないか。

Q：安心も実は言語化されていない。逆に、不安を当事者が言語化することによって一つ階段を上れるように思う。

A：安心が言語化されていないというのは大事なことだ。誰の、いつの、どのような安心を目指そうとしているのか、その状況においてそれぞれの関係者がどう考えているのかを言語化していくことから、不安に向き合っていく思考はスタートすると思う。

Q：安全工学、安全システム学の先生から、安全装置を自動化してしまうと、当事者が今安全かどうかを認識する能力がなくなると聞いた。医療の世界でも、慣れてしまわない方が、患者も医師もどちらも安全をキープできるのではないか。

A：その通りだと思う。

Q：医療におけるトラストは、医師免許などの第三者認証の形で、制度として専門家の資格を認めることが大きいと思う。情報技術の発展により医師の役割や関係性が変わっていく中で、専門家の資格認証に変化はあるか。

A：既に変わってきている。医学部の卒前教育においても、今までは医学的知識だけだったが、現在は、病院実習に入るときの実技の面接で、しっかりと対話ができるかが必要になっている。また、合衆国などにおいては、コミュニケーション、態度領域の部分が、コンピーテンシーとして医師免許を発行する要件になっている。医師になってからは、専門家として正当な意図を持って患者に向き合っているか常に評価する仕組みとして、2020年の厚労省のガイドラインから、初期臨床研修の評価の対象がそちらにシフトしている。

Q：個人の専門家への信頼と、制度やシステムへの信頼の2つのレベルがあるのではないか。COVIDワクチンを受けない人たちは、個人の意思よりも制度全体、そしてサイエンスへの信頼がない。いくら一人の専門家が頑張っても、システムへの信頼がないと、まず病院に行かないということもあると思う。

専門家と患者のトラスト要素のうちの能力は、専門の能力だけでなく、人間と人間の理解の能力、エモーショナルインテリジェンスと言えるかもしれないが、患者の立場、患者のリアリティ、現実を理解できないと、信頼を成立することは難しいのではないか。

患者－医療者－情報技術の三者関係において、医療者が選択、例えば診断をするときに、後ろにAIの技術があると、医療者だけでなく後ろにある技術にも信頼があるかどうかが大事ではないか。

また、私はイギリス人として、安心も大事だが、自分の生活や体のオートノミー、コントロールがより大事である。これにはかなり文化の背景もあると思う。

A：全面的に同意見だ。信頼にはまず一般的信頼があって、そして専門家は人なので人格的信頼があり、

その上に関係的信頼と三段階に分かれている。一般的信頼は、例えば日本の医師や政治家に対する信頼、人格的信頼は尾藤という医師に対する信頼、関係的信頼は本日受診し薬を飲む上で話をしている尾藤に対する信頼である。現場では、シェアードディシジョンメイキングができることが重要で、その前提条件としての関係的信頼という立て付けになっている。関係的信頼をどう作っていくかというところに、どこまで本当のことを言うか、逆にどこまでちょっとだけ嘘をつくかということも含まれる。

よく社会学でいわれるのは、特に日本においては、イギリスなどに比べて一般的信頼が非常に低い。例えばイギリスのGPは非常に信頼されているが、日本の開業医は全然信頼されておらず、日本の政治家も、一般的信頼が非常に低い。その一つのからくりが、安心の強要だと思う。一般的信頼専門家集団やガバメント集団は、私たちに任せておけば安心だということを刷り込もうとする。すなわち、社会は安定しているから、個別の人民は考える必要などないというメッセージを送りすぎる。専門家にしても国を統治する人間にしても、いかに自分たちも怖いのか、いかに自分たちも不確実なことを不確実だと認識できているのか、そのために医師なら医師集団、医療者なら医療者集団で、この先どうなるか分からないから、こういう仕組みを皆で何とか作ろうとしています、みたいなことをやっていくのが、一般的信頼を得ていく上で非常に重要だと思う。

Q：一般的信頼で医師と政治家を並べたところが興味深い。医師は、一般人が分からないような専門知識をもつ専門家であり、医師免許の形できちんと保証されたところで得られる信頼だと思う。政治家は、専門知識というより透明性や説明責任によって信頼を得ているように思う。

A：専門家のトラストの要素である能力と意図の比重の差ではないか。能力で政治家を見ている人はあまりいないと思う。ちゃんとした意図を持って正直にメッセージを発信し、正しいところに導いてくれようとしているのかという、意図の比重が大きいのではないか。一方で医師は、まともな人だろうという前提で、ちゃんと勉強しているか、ちゃんと注射が打てるかなどの能力部分が信頼の大きな要素となっているのではないか。

その中で、状況安定性、すなわち安心に対するメッセージの出し方によって信頼が変わってくるのは、政治家も医師も共通していると思う。法律でも教育でも統治でも、プロフェッショナルといわれる人たちが集団としてちゃんとしているのかを考えるときに、安定し続けて変化をしない頑強な存在であると分からせようとすればするほど、一般的信頼は薄れるのではないか。

Q：人間としての尊厳を求める患者と、生きる、死なない、大きな病気にならないということを第一義に考える医療者との立場の違いがあるときに、情報技術はそこに対して貢献できないと思う。一歩離れてそういうことを助言できるような倫理端末、生き方端末のような役割があると良いと思う。

A：実は、東京医療センターで倫理サポートチームを10年ぐらいやっている。患者や患者家族が延命治療をやめてほしいと言う一方で、担当医療チームが助かるかもしれないから頑張りましょうと言ったときに、一対一の関係性だと、どちらが正しいかと勝負のようになってしまう。ここに対話の筋道を支援するチームが入ると、ここでのアジェンダはどちらが正しいという話ではなく、患者にとって一番いいディシジョンメイキングは何かという話のはずだ、と主体者からアジェンダに目を向けられるようになる。このように、「人が問題なのではなく、問題が問題である」と考えることをナラティブと言う。対立する価値となると、どうしてもその価値を持った人間のアジェンダになってしまうところを、いかに対話のテーブルに乗せていくか。物語の登場人物として患者や家族、担当医療チームがいて、ベストチョイスに向かっているというように物語を作っていくようなやり方ができるのが、患者－医療者の二者関係に倫理チームが割って入る意義だと考えている。

Q：これからの診察室が患者－医療者－情報技術の三者関係に変化していく中で、コミュニケーションの仕方が改善されていくと思うが、このあたりの理解を深めた研究の出口としてどのようなアウトプットが考えられるか。

A：人工知能も含めた情報時代には、専門家に対する、第三者による意思決定や価値の翻訳などのサービ

スができるのではないか。シェアードディシジョンメイキングの本来的な目的は、情報をシェアし、認識をシェアし、価値をシェアし、そして欲望をシェアするという、多面的なシェアである。情報が正確な曖昧さを持った情報にどんどん精緻され、当事者の立場と専門家の立場が明瞭になればなるほど、当事者が大切にしているのはこういうところだと思うので、担当の医療者に、医療の専門的知識を持った第三者が少し手を添えてみます、というような第三者的なサービスが成り立つと思う。

Q：EHRやPHRが医療全体のトラストにつながる方向に行くのか疑問がある。診療報酬に最適化しているのが医療機関だと思うが、そういったところがむしろEHRを難しくしているのではないか。

A：電子カルテの情報は、今は大体経営分析に用いられており、日本は2周遅れぐらいだ。電子カルテの情報から、病院のクオリティーを一般の方々に透明性を持って発信する取り組みを日本医療評価機構と行っている。技術的に難しいところもあるが、ようやくそういうフェーズに来たという実感はある。

Q：DPCのデータを社会に還元するようなエコシステムを作るには、法制度で強制するしかないのではないか。海外でそういった動きはあるか。

A：イギリスやアメリカには、質の高い医療をやっているところにお金を付けるPay for Performanceという仕組みがある。イギリスには医療システムを統治できる行政組織が、アメリカには統一された電子カルテがあるので、DRGやDPCのような一般的なハードなデータ以外にも、会話のテキストデータなどのソフトなデータも含めて取り組んでおり、前例は蓄積されている。

Q：トラストを上げるために患者との意思決定をより丁寧に行わなければならないとすると、時間、労力が必要になり、医療現場は嫌がるのではないか。

A：イエスでもありノーでもある。シェアードディシジョンメイキングのときに何を伝えようとしているか、何を知ろうとしているかについてしっかりと意識することで、時間も労力も節約できると思う。個別に情報提供する必要のない一般的な情報と、この患者がどのような生活をしているか、この仕事をしている患者に対して医師が考えているお勧めは何か、などの生身の情報を切り分けることだ。さらに、何に向かって情報提供しているかをしっかりと見据えていくことだ。

また、当事者から専門家が教えてもらう情報が重要である。これは逆トラストで非常に大事だと思っている。トラストは、クライアントが専門家をトラストするだけでなく、専門家がクライアントをトラストしなければ成立しない。特に認知症や自閉症の診療のときに、専門家が患者を見下しているようだと絶対にうまくいかない。当事者は非常に敏感にそこをキャッチするので、私を信頼していないなと思う。そうすると、インビテーションの部分から始まらない。

いかに専門家側がライフストーリーを持った当事者をトラストし、当事者が妥当なディシジョンメイキングをしていく上で、自分はこの人から何を教えてもらおうとしているのかを踏まえていくことが専門家としてのテクニックだ。現時点では、そこは情報技術というよりは、専門家の大きな役割の一つだと思う。

## 2.15 山本 ベバリーアン[36]「医療におけるトラスト（2）」

「医療におけるトラスト」の2回目として、医療とAIとトラストについて、主にステークホルダーとしては患者・市民からのトラストを話したい。さまざまな学術論文や報告書などで、ヘルスケア領域にAIが導入される際に、トラストは必要な条件、望ましい条件であるといわれており、ガイドラインなどでは、透明性（Transparency）、確実性（Reliability）、説明責任（Accountability）の3つが、トラストの形成要件として示唆されている。しかし、私はこうした説明では医療のトラストに関する問題を十分に概念化できていないのではないかと考えており、本日は問題提起とそれに関する議論を行ないたい。

なお、今日の発表における「トラスト」と「信頼」は英語の”Trust”と同義語として扱う。

**トラストの簡単な定義と分析**

トラストの定義でまず一般的に出るのは人間の間のトラスト（Interpersonal Trust）である。ヘルスケア領域では、特に医療従事者へのトラストについては、善行（Benevolence）と正直（Integrity）と能力（Ability）の3つの要因が大事ではないかとよくいわれる[1]。

信頼の分類方法で重要なものとして、一般的な信頼（Generalized Trust）と特定的な信頼（Particularized Trust）がある。社会の中の信頼が低い、つまり一般的な信頼が低い場合は、自分の仲間集団に対する信頼、つまり特定的な信頼が高くなる傾向があるといわれており、例えば論文などではイタリアがそういった例でよく挙がる。一般的な信頼は、自分の仲間ではなく他の人間を信頼できることであり、AIとヘルスケアにおいては特定的な信頼よりも一般的な信頼の形成が大事ではないかと考える。

一般的な信頼に関連する要因について、主に日本を研究対象とした論文[2]に基づいて説明する。1つ目は年齢である。年齢が上がるとともに一般的信頼が高まる傾向があるが、コホート効果[37]もある。例えば日本の場合は、戦前と戦争中に生まれた人たちの一般的信頼は低い、高度成長期に生まれた人たちは高い、といった事例が報告されている。2つ目は性別であり、女性の方が低く、男性が高いという傾向があるが、日本の場合は、女性が社会に進出して仕事を持つようになることで、女性と男性で一般的信頼が同レベルになってきている。他に、高等教育を受けた期間が長い、豊かな経済的状況、といった環境で一般的信頼が形成されやすく、また、日本の場合、田舎よりも都市に住む方が一般的な信頼が高いと報告されている。最後に、マイノリティー集団意識が高い場合は一般的な信頼が低く、逆に前述のように特定的な信頼が高くなる。日本の社会におけるトラスト形成においても、マイノリティーは誰かを考えるのは大事である。そして、その社会における汚職の程度も大事な要因で、汚職が多い、もしくは多いと見られている国は一般的な信頼が低くなる。

一般的な信頼の対象としては、人間ないしそのグループが対象となる研究が多いが、本セミナーシリーズで取り上げられている、テクノロジーや技術のような物へのトラストも大事である。ヘルスケアの場合、医療制度への信頼は重要なポイントで、信頼されていない場合、自分を診断・治療する医者を信頼していたとしても全般的にトラストが低くなるため問題になる。そして、先ほど述べた通り、一般的な信頼が高いといわれる社

**36** 大阪大学 人間科学研究科 人間科学専攻 教授
https://researchmap.jp/beverleyyamamoto
NPO法人HAEJ（遺伝性血管性浮腫患者会）理事長
https://haej.org/about
「ヘルスケアにおけるAIの利益をすべての人々にもたらすための市民と専門家の関与による持続可能なプラットフォームの設計」AIDE ProjectのHP
https://aide.osaka.jp/
https://en.aide.osaka.jp/

**37** 暴露された環境の違いにより、ある特性について世代間で差が生じること

会においても、平等に誰でもトラストするわけはない。仲間とみなされる集団（In Group）に対してはトラストが高く、その外の集団（Out Group）に見られる人たちに対するトラストが低くなるという状況が、どの社会でもある程度見られる。例えば日本でも、日本人と見られる人は仲間、日本人と見られない私のような外国人は外の集団になりやすい。

続いて、日本の社会における一般的信頼について見る。私のイギリス人としての印象は、日本の社会は一般的信頼が高い。経済学者であるFrancis Fukuyamaのトラストに関する有名な本[3]では、日本とドイツが一般的信頼の高い国の例として挙げられ、アメリカよりも高いと指摘されている。一方、用いたデータや研究の対象者により結果が変わってくるようで、山岸らの研究によると、日本の一般的信頼はアメリカよりも低いと述べられている[4]。また、時系列での変化を見ると、1990年から2013年の間に一般的信頼が低下しているという結果が、35年間の日本人の国民性調査の分析から報告されており[2]、経済的な失われた10年、20年や、リーマンショックの影響が大きいのではないかと考察されている。格差社会という意識が強くなったのも影響しているのではないかと思う。

AIヘルスケアを考えるときも、こうした一般的信頼の影響は無視できない。



**健康・医療におけるAIの開発・実装とトラスト形成**

ヘルスケア・医療といってもかなり多様な領域であり、例えば福祉・介護分野と医療診断では、AI利用における課題やステークホルダーが大きく異なるため、違いを認識しながら考えないといけない。また、IoTにより遠隔モニタリングなどの医療従事者介入なしでのヘルスケア領域が拡大しつつあり、これら技術についてもAIがバックグラウンドにあることが多いことから、課題となる。

医療におけるAIとトラストを考える上で留意すべき点として、医療分野で特に医療従事者に対して期待される思いやり（Compassion）、信頼（Trust）、共感（Empathy）といった要素が、AIが入ることでその役割や性質が変わる可能性があるという点が挙げられている[5]。これを図式化したものが図2–15–1である。

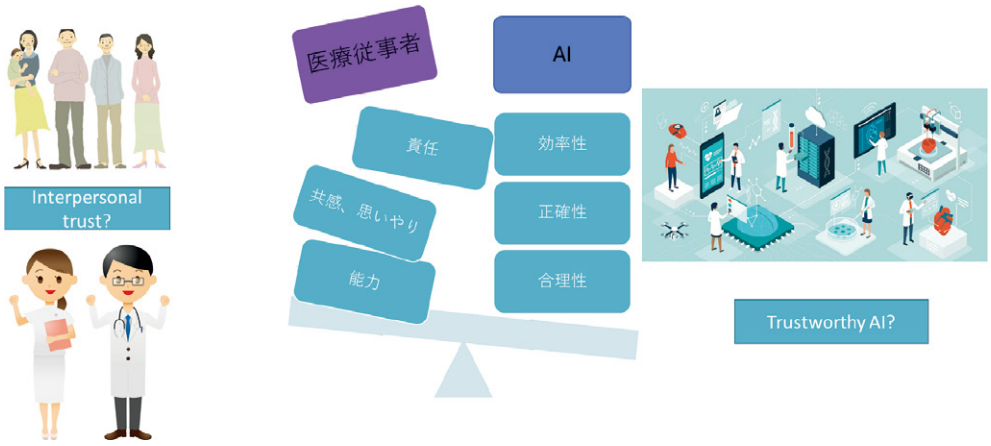


**図2–15–1　医療AIが医療にもたらす混乱と信頼に関わる課題**

市民・患者と医療従事者の間の信頼には、責任、感情、思いやり、能力といった要素が関係するのに対して、AIの方は効率性や正確性、合理性といった特徴によりTrustworthy AIが規定されると考えられる。Trustworthy AIが開発、導入されることで、これまで重要であった市民・患者と医療従事者の関係性が変わってくるのか、トラストがどうなるか、医療従事者の能力が低下してしまうのかといったさまざまな課題が出てきて、患者・市民側に懸念や不安を抱かせる可能性がある。本セミナーシリーズの中でも取り上げられている通り、Trustworthy AIとはどのようなものかという点も大きな課題で、ヘルスケア分野でもEUのガイドラ

インをベースとした議論が行なわれている。

一方、AI がTrustworthyかどうかと、市民や患者がトラストするかどうかは別の課題ではないかと考える。つまり、似ている概念である信頼（Trust）と信頼性（Trustworthy）を分けて考えた方がいいのではないかと考えている。 Gilleらの論文[6]では、信頼性それ自体が信頼関係をもたらすと限らない、従って、信頼とAIの議論においては、特性としての信頼性だけではなく信頼関係を構築するプロセス全体に焦点を当てるべきであると述べられている。要するに、関係性の構築を考えると、そのTrustとTrustworthyの両方の概念を考えなくてはいけないと思う。ここで「Will trustworthy AI be trusted?」が市民・患者グループにとってビッグクエスチョンとなる。

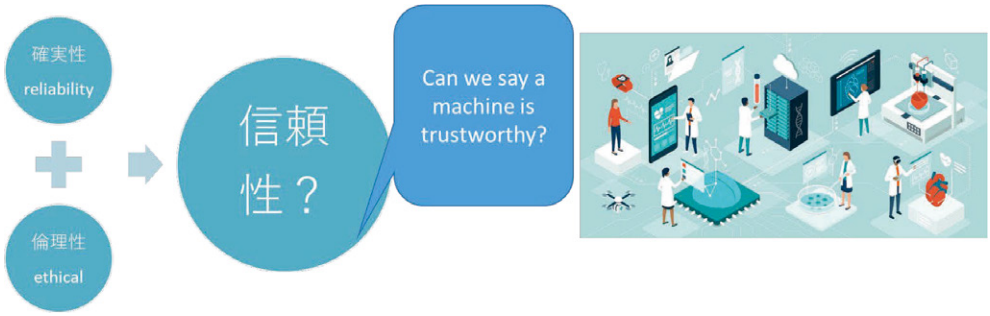


**図2–15–2　　信頼（Trust）vs 信頼性（Trustworthy）**

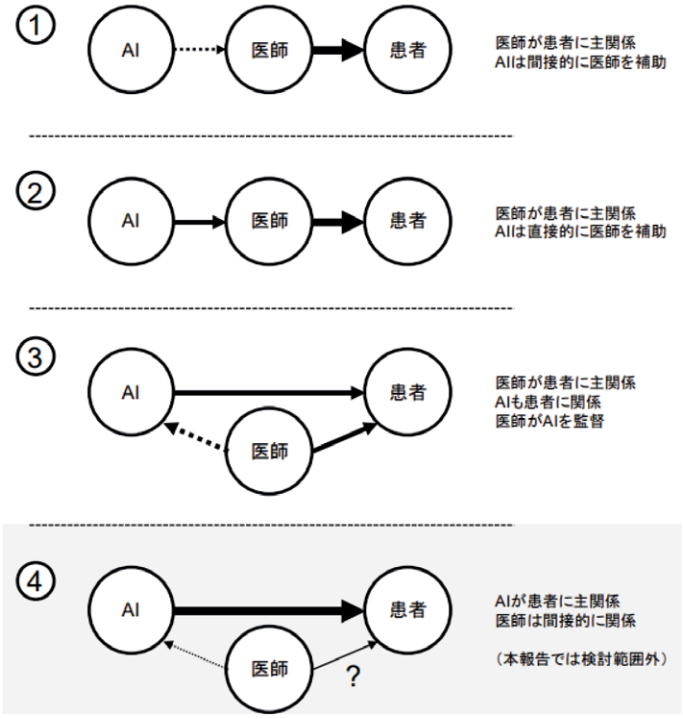


**図2–15–3　　AI・医療従事者・患者の関係の在り方**

また、患者、医療従事者とAIの関係の在り方もトラスト形成を考える上で重要なファクターである。医療従事者介入なしのサービスも開発されつつあるが、医療機関という現場で見てみると、今のところ医療従事者が介入しており、患者が直接AIに直面することはない。つまり、図2–15–3のモデル1や2の状況で、人間の医師がループに入っている。一方、モデル3や4の場合は医師、医療従事者の役割が薄くなる。こうした

モデルが実際に入ってくるかどうかはまだ分からないが、トラスト形成の在り方がモデルによって変わってくるはずである。

1や2のような医療従事者が間に入るモデルであれば、医療従事者のAIへのトラストやコンフィデンスが大事になると思う。また、人間がその意思決定のサイクルに入る、すなわち Humans in the Decision Loop が非常に大事である。東大の青木らの研究では、AIを使用して作成されているケアプランについて、ケアプランナーが意思決定に入っている場合には、信頼する人たちの割合が8.95%高くなると報告されている[7]。

もう一つ大事な概念として、今の医療ではインフォームドコンセント（Informed Consent）が基本となっているのに対し、AIが入って患者・市民にとっても医療領域が変化する場合、エデュケーティドコンセント（Educated Consent）が必要なのではないかとGilleらの論文[6]では主張している。エデュケーティドコンセントの形成過程としては、患者・市民がその教育の対象となるだけではなく、学びのプロセスに入って新しい知識を得る過程で関係性が構築されることが大事と考えられている。

これまで挙げた、市民・患者が医療AIを信頼するかどうかに関わる要因を図2–15–4にまとめた。AIの場合は、医療従事者や病院、医療制度についてのトラストだけでなく、技術へのトラストも加わってくる。また、一般的信頼に関わる要因である年齢、性別、教育、経済的状況なども関係する。研究における一つの問題点として、医療と市民・患者の関与についての研究の中では、経済的に豊かな人たちがステークホルダーの対象になりがちである、つまり一般的信頼が高い層を対象としていることは考慮すべきである。

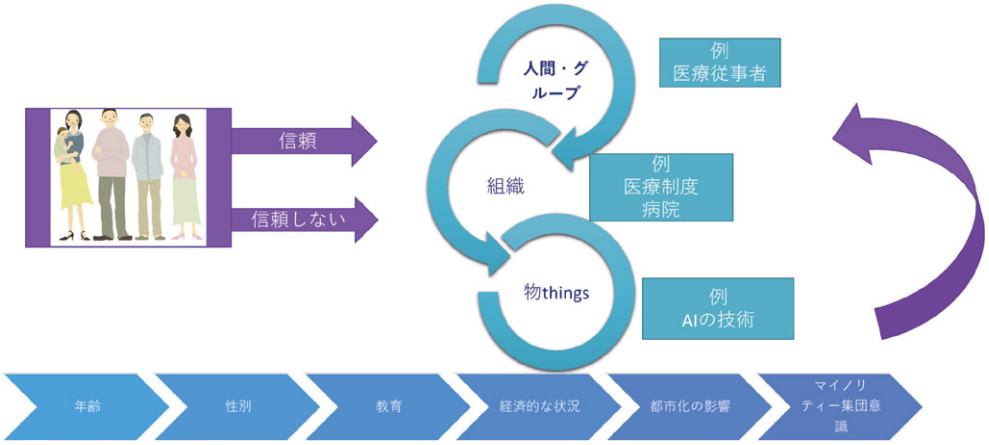


**図2–15–4　医療AIへの信頼に関する要因[34]**

ディストラスト（Distrust：不信）もトラスト形成において考慮すべきである。前提としてトラストが良いことであると決めがちであるが、ディストラストは必ずしも悪いこととか理解できないことではない。「The Wisdom of Distrust」[8]という本の中で、トラスト・ディストラストは両方とも、客観的な社会経済的な文化政治的な要件を反映した経験が蓄積した結果ではないか、といった検証を行なっている。すなわち、両者ともその集団が直面しがちな条件に対する合理的な反応であると考えられる。そうすると、合理的な反応であるディストラストが生じた背景となる要因をどうにかしないといけない、というのが課題となる。Ruha Benjaminの「People's Science」[9]の中で、似たような理論が出てくる。この文献では、アメリカのアフリカ系アメリカ人を例に、特にヘルスケアに対するディストラストを生み出した背景について述べている。多くの

38　出典：PMDA報告書
https://www.pmda.go.jp/files/000224080.pdf

文献において、タスキギー梅毒実験[39]といった歴史的な事例が代表的な要因としてよく引用されるのに対して、Benjaminは、それよりも今の制度の信頼性（Trustworthiness）にフォーカスしなければ、ヘルスケアシステムが日常的に直面している、ディストラストが生まれ続けている社会的状況から目に背けることになる、と主張している。日本の社会でも同様にディストラストを持っているマイノリティーがいると考えられ、どうすればよいかという課題も出てくる。

医療システムやヘルスケアへのディストラストが合理的に生じているということを踏まえた上で、AI医療に対する信頼を形成するための方法論として、小さな一歩かもしれないが、市民・患者参加型の方法論を探す必要があるのではないか、というのが私の研究の着眼点である。

**医療におけるステークホルダー関与（Stakeholder Engagement）とトラスト**

日本においても、患者と医師の対話モデルが、旧来のパターナリズムから意思決定を共有するモデルへ、さらに最近では医師が提供する情報や患者の収集した情報に基づき意思決定を行なうモデルへと移行しつつある。モデルが移行した背景として、一つは民主的な観点からの要請、すなわち”Nothing about us without us”という言葉で代表されるように、自分の体に関われる医療については、自分が意志決定に入るべきという考え方である。また、医療を実践し説明責任を有する立場からも望ましく、さらに実際に良い成果や成功率の向上といった効果も見られている。最近では、アメリカやイギリスなどにおいて、市民・患者参画が医学研究に対するファンディングの必要条件になってきている。

そして、情報社会の変化により、患者が自分で情報を収集できるようになったのも大事な要因である。自分で収集した情報をもとに患者が意思決定に参加できるようになった一方で、エコーチェンバーができてしまうという欠点もある。そのため、トラスト形成においても、参加型でダイアログを重視するのが望ましいと考える。

こうした流れの中で、PPI（Patient and Public Involvement）という市民と患者が参画する手法が海外で広がってきており、国内でも2018年よりAMEDが促進している。 PPIには参加（Participation）からエンゲージメント（Engagement）、参画（Involvement）までさまざまなレベルがある。日本はPPIを実施するためのインフラがまだ整っているとは言えない状況であるが、その中でも少しずつ進めている。

**ステークホルダー関与による医療AIにおける信頼構築　～「AIDE[40]プロジェクト」の事例～**

大阪大学とオックスフォード大学の共同で進めているAIDEプロジェクトは、PPIによりAIを利用したヘルスケアについての信頼構築を試みるものである。プロジェクトの狙いとしては、ステークホルダーの関与によるトラスト形成、つまりAIが開発された後にトラスト形成を図るよりも、開発レベルからその実装までのプロセス全体に市民と患者が関わる方が、そのプロセスを通してトラストが形成されるのではないかというものである。

図2–15–5は、AIDEプロジェクトのアウトラインである。まず、WP（Work Package）2でPPIP（市民と患者の参画パネル）を構築して、定期的に会合を行なう。続いてWP3として、専門家で構成したアドバイザリーボードのアドバイスを受けて、スコーピング分析レビューと追加分析を行なう。そこには患者は直接入らないが、報告は行なう。その報告を受けて、WP4では主に医療従事者やAI研究者がオックスフォード側と阪大側合同でフリーディスカッションを行ない、ステークホルダーは誰かといった課題が出ている。秋に実施予定しているのが、WP5のステークホルダーのフォーカスグループで、患者・市民や医者、看護師、病院の

**39** 正式名称はTuskegee Study of Untreated Syphilis in the Negro Male。アメリカのアラバマ州のアフリカ系アメリカ人比率が高い町であったタスキギーにおいて、1932～1972年まで実施された梅毒の臨床研究。治療を受けない状態での梅毒の症状の長期観察が目的であったが、対象であった黒人男性たちに目的が知らされていなかったこと、研究期間中に治療法が確立したにもかかわらず治療を行なわなかったことなど、医療倫理的に大きな問題を有する人体実験であった。

**40** Artificial Intelligence in Healthcare for All: Designing a Platform for Sustainable Stakeholder Engagement の頭字語

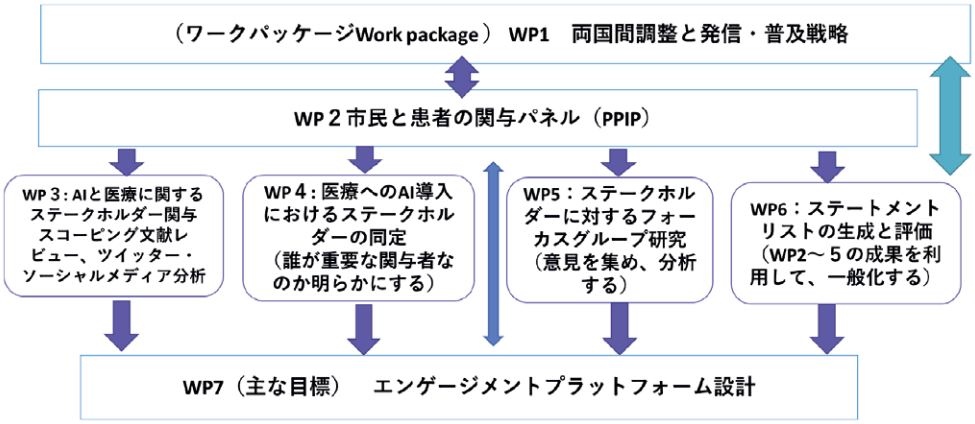


**図2-15-5　　AIDEプロジェクトの7つのワーク・パッケージ**



事務も含めたステークホルダーやオックスフォード・阪大両方のPPIPからデータを収集する。それをもとに質問を作ってPPIPに再び確認してもらい、WP6へのインプットとする。こうした取り組みを通じた最終的な私たちの目標は、PPIPがプラットフォームの一部として含まれるエンゲージメント・プラットフォームの設計である。

プロジェクトの重要なポイントは実践的なデータ収集であるが、基本的にPPIPが全プロセスに関わる。内閣府SIPのAIホスピタルとも連携を取りながら進めており、今のところは双方にとって有益な連携となっている。そして、AIDEプロジェクトでは、ほぼ毎週プロジェクトの情報やAIホスピタルの情報を流すことで、一般市民の認識や関心を高める取り組みを行なっている。閲覧数も結構多い。 PPIPは2か月ぐらい募集して、構成（一般市民、患者、患者の家族）、男女比、年齢ともバランス良く11人集まった。最初に行なったトレーニングは、大阪大学側とオックスフォード大学側でインフラが全然違うためやり方が異なるが、阪大側ではまず大阪大学AIホスピタルプロジェクトについて川崎良先生が紹介し、その後メンバーから課題についてもっと知りたいという希望があり、東大の井上悠輔先生にお願いして、医療AIの倫理的・社会的な課題についてワークショップを実施した。その後、医療AIへの期待や懸念を特定、グルーピングするワークショップを2回実施した。参画メンバーは活動に対し非常に積極的で、自分が代表を務める患者会に情報を持っていきたいとい

**表2-15-1　　各国のPPIPで挙がった医療AIへの期待と懸念（上位5項目）**

| | 日本 | 英国 |
|---|---|---|
| 期待 | •改善された病院管理<br>•ケアの質の向上<br>•関係の改善<br>•＊コスト削減<br>•＊患者体験の改善 | •精度の向上<br>•診断と治療の速度の向上<br>•＊健康状態の監視能力と高度な診断<br>•＊個別化及び患者中心の医療とケア<br>•医療従事者の時間の解放 |
| 懸念 | •＊従来のヘルスケアのあり方からの変化（改悪）<br>•＊自律性の制限<br>•＊技術的課題と説明責任<br>•新たな格差<br>•データ管理 | •＊バイアスとデータの　代表性<br>•＊同意とデータ利用<br>•規制の必要性<br>•除外のリスク<br>•＊信頼性と透明性の課題 |

＊その国の PPIP でのみ話題になった項目

う患者もいた。オープンなダイアログを通して認知度や関心が高まり、医療AIが自分から遠いものでなく近いものだと感じるようになってきているのではないかと考える。

PPIPの中で挙げられた医療AIに対する期待や懸念は、表2–15–1のようにオックスフォード大と阪大で異なっていた。懸念の中で注目しているのは、日本側で出てきたヘルスケアの在り方そのものも変わってくるという懸念である。具体的には、医療の本質が変化することへの心配や、開発されたAIシステムへの不安といった点が挙がった。対話を重ねながらトラストを形成する過程で、市民・患者にとってより望ましいAIを開発することも可能になると考える。

**まとめ ～医療AIにおける市民・患者のトラスト形成に向けて～**

医療AIのトラストは、個人・集団レベル、組織レベル、ものレベル（AI技術）の3つのレベルで考える必要がある。また、信頼（Trust）と信頼性（Trustworthiness）という概念を区別することが大事であり、Trusteeに信頼性があることがTrustorにとっての信頼に必ずしも繋がらない、という視点が大事ではないかと考える。そして、ステークホルダーのグループの一つである患者は、実際には非常に多様なグループであり、病気の多様性に加え社会的な経験、教育の違いなども考慮しないといけない。そして、良い人間関係とPPIというフレームワーク、そして医療AIの本質を通して信頼構築に取り組む必要があるのではないかと考える。

**【主な質疑応答】**

Q：図2–15–3の医療AIのタイプ分けについて。1や2のように患者と直接対面しない医療AIは、患者と直接接する医師がAIをトラストするかどうかで導入が決まるのではないかと感じている。一方、3や4のような患者が直接対面する医療AIの導入については、現時点では法規制の問題もあるが、患者がトラストするかどうかが本当は重要ではないかと考えている。 PPIを通して目指しているのはタイプ3や4の医療AIに対する信頼構築なのか。

A：AIDEプロジェクトが対象とするのはタイプ1と2である。医師が医療AIをトラストするのであれば患者はそのままで受け入れるというのはパターナリズムであり、AIも完全ではないことからあまり良いトラストとは言えない。また医療AIの効果的な導入には実際の患者目線が必要と考える。例えば、一般市民を対象とした研究で、アルゴリズムによるハイリスク診断（生活に大きく影響を及ぼすような診断）は信頼しないという報告があるが、本当の患者でずっと未診断だった病気がアルゴリズムによりやっと診断がついたという状況であれば、受け止めが変わるはずである。

一方、タイプ3と4のような状況は、今の病院においてはあまり想像できない。海外では「医療従事者の方は患者と対面する質が高い時間がほしいのに対し、患者は病院に行かなくていい状況を望んでいる」といった報告もあり、患者に歓迎される面も大きいのではないかと思う。

Q：Medical Chat Applicationのように、自分でデータを入力して、医師なしでアドバイスが返ってくるといったサービスにおける信頼について、どのように考えているか。

A：ヘルスケアAIやステークホルダーエンゲージメントを扱った研究において、そもそも患者を対象にする論文はまだ少なく、また主に医用画像を扱う研究が多いため、チャットアプリケーションといったものを対象とした事例については、十分に知見がない状況である。ウェブ上の情報提供や口コミは大事だと思うので、今後考えていきたい。医療AIは技術の適用領域が非常に幅広く、病気や患者も多様であるため、現在はいくつかの技術に焦点を当て、そのステークホルダーや課題について検討しているところである。

Q：PPIの適用範囲について。タイプ3や4のような状況について、イギリスではBabylon Health社のようにAIが市民・患者と対面するヘルスケアサービスが既にあり、市民側も想像ができ議論もしやすいと思うが、日本のようにそういったサービスが実装されていない場合、PPIPにおいて市民を巻き込んだ議論が可能なのか。

A：可能だと考えるが、どういったトレーニングを行なうか、どういった人をPPIPに入れるかが重要となる。AIDEプロジェクトにおいて、当初、日本側の市民・患者にとってはタイプ1や2の医療AIについても現実的なものとして想像できないという状況であった。その状況に合わせて、PPIPのフィードバックを得ながらトレーニングを設定した結果、今では一般の市民・患者よりもAIについて知識がつき、PPIPで議論を行なうだけでなく周囲に活動を紹介するようなメンバーも出てきている。プロジェクトを通して、市民・患者がヘルスケアAIについて考えるためのリソースを提供できているのではと考えており、将来的にはスケーリングアップしていきたい。

Q：科学者の視点では、AIはあくまで機械・システムであり、効率性、正確性、合理性といった点に着目するが、市民や患者はAIを擬人化してもう少しエモーショナルに接するのではないか。

A：指摘の通りであると思う。特に介護の領域では、チャットボットのような技術でエモーショナルな関係性ができるといった報告もある。人間は関係性を作ることで安心する傾向があり、対面するAIとの関係性をユーザーがどう把握するかも大事である。

C：医師に対する患者の信頼を形成するためには、まずは双方向の共感、つまり、患者から医師への共感だけでなく医師が共感してくれていると患者が感じることが必要ではないか。

Q：社会福祉の分野では、これまで人間が担ってきた共感や思いやりをもったAIの開発に対する期待があるが、医療の現場ではその点についてどう思われているのか。

A：PPIPの中で挙がった患者・市民グループが懸念する点は、医師側の専門性低下や診断力低下、それに伴う医者への信頼低下である。つまり、思いやりや共感への期待以上に、医師や医療AIの能力に対する信頼性の方が大きな要因であると言える。一方、今のところはAIではなく医師が患者と対面するため、医師の思いやりや共感が重要であると考える。

Q：トラストとディストラストの関係性はどのように考えているか。対義語なのか。

A：私もまだ結論が出せていない。患者が医療従事者の指示やアドバイスを受け入れるためにトラストが必要という考え方もあるが、それだけではパターナリズム、盲目的なトラストBlind Trustとなってしまうことから、クリティカルシンキングやセカンドオピニオンが重要なのは確かである。そうした考え方とトラスト、ディストラストの関係性についてもう少し考えを深めたい。

Q：意思決定における信頼の役割について。心理学における精緻化見込みモデルによると、自分が中身についてよく知っているメッセージについては、受けるか受けないかの意思決定を自分でできるのに対し、よく知らない場合は、中身ではなくメッセージを発した人に対する信頼によって受け入れるかどうかを決める。このモデルに基づくと、PPIの効果は、信頼の形成というよりは、参画を通して理解することにより自分で意思決定ができるようになることではないか。

A：興味深い見方だと思う。精緻化見込みモデルにおける信頼は人間の間の信頼を意味すると思うが、医師と患者が一緒に意思決定する場合、診断の中身に対する理解や人間の間の信頼に加えて一般的な信頼、特に医療制度への信頼が重要である。 PPIは、一般的な信頼の形成を通して市民・患者の意思決定に寄与できると考えている。

C：一般的信頼と社会関係資本（Social Capital）の関係について。一般的信頼の解説の中で、社会関係資本＝人々が共同で働く能力として定義し、一般的信頼があると社会関係資本が高くなるという説明があったが、個人レベルでの捉え方であると考える。地域レベルで考えると、一般的信頼は社会関係資本の一つと見るべきではないか。

# 3 ｜ 俯瞰ワークショップ

15回の俯瞰セミナーシリーズ（第2章）で把握できたさまざまな分野におけるトラスト研究の動向をもとに、それらの俯瞰的な整理をCRDSにて試みた。そして、その俯瞰的整理の結果を確認しつつ、まとめた内容の妥当性やトラスト研究の課題などを議論する俯瞰ワークショップを開催した（開催概要は付録2参照）。そこでの論点として以下の5つを設定して議論した。

【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？
【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

ワークショップには、セミナーシリーズの講師にコメンテーターとして参加していただいた。そして、トラスト研究動向の俯瞰的整理の内容や今後の方向性・重要課題についての論点をCRDSから発表し、コメンテーターから意見をいただきつつ議論を進めた。

以下にその結果をまとめた。まず3.1節にはトラスト研究動向の俯瞰的整理を中心にまとめた。この内容はワークショップでCRDSから発表したものをベースとしているが、当日の議論などの結果を踏まえて一部アップデートしたものとなっている。次に3.2節にはワークショップで出された意見や示唆をまとめた。その一部は既に3.1節の整理に反映済みだが、3.1節には盛り込めていない観点やより掘り下げた意見などもあり、3.1節を補足・補強する内容となっている。

## 3.1 俯瞰的整理

### 3.1.1 トラスト研究の変遷

人文・社会科学分野におけるトラスト研究の変遷を図3–1–1にまとめた[1]。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。

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・ギデンズ）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。

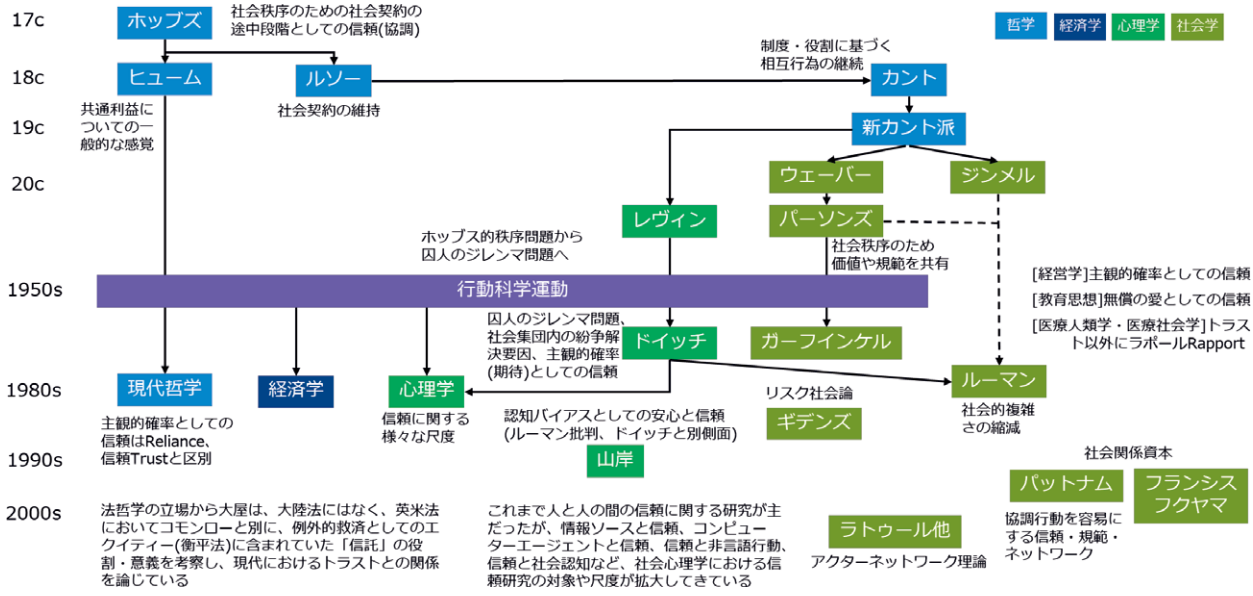


**図3–1–1　　トラスト研究の変遷（1）人文・社会科学分野**

次に、情報科学分野におけるトラスト研究の変遷を図3–1–2にまとめた[2]。情報科学分野の研究として、「トラスト」という言葉を使って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、セキュリティーやプライバシーの研究に関わる取り組みが一

1　2.1節の図2–1–3をベースに、セミナーシリーズで紹介された関連する話題や研究者を追記した。

2　2.6節の表2–6–1の情報をベースに、セミナーシリーズで紹介された関連する話題を含めてプロットした。

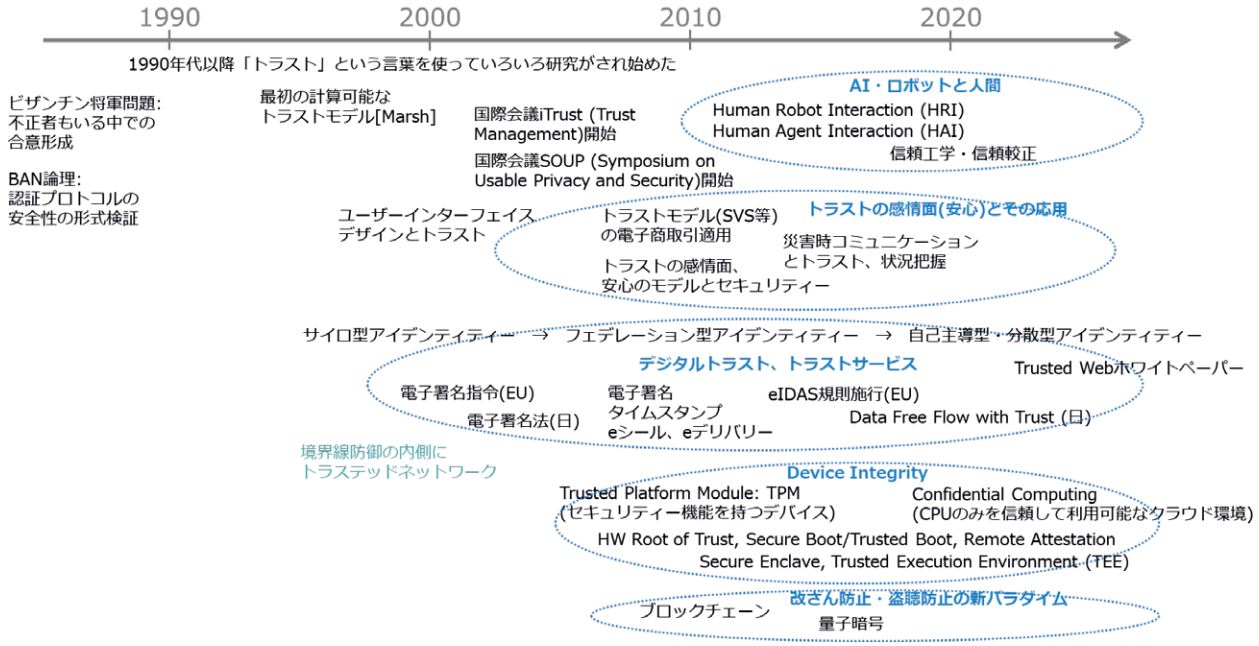


**図3-1-2　　トラスト研究の変遷（2）情報科学分野**

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ー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技術も注目されている。

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3-1-3にまとめた[3]。

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ー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・セキュリティー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

3　2.7節の図2-7-2を中心に、セミナーシリーズで紹介された関連する話題を追記した。

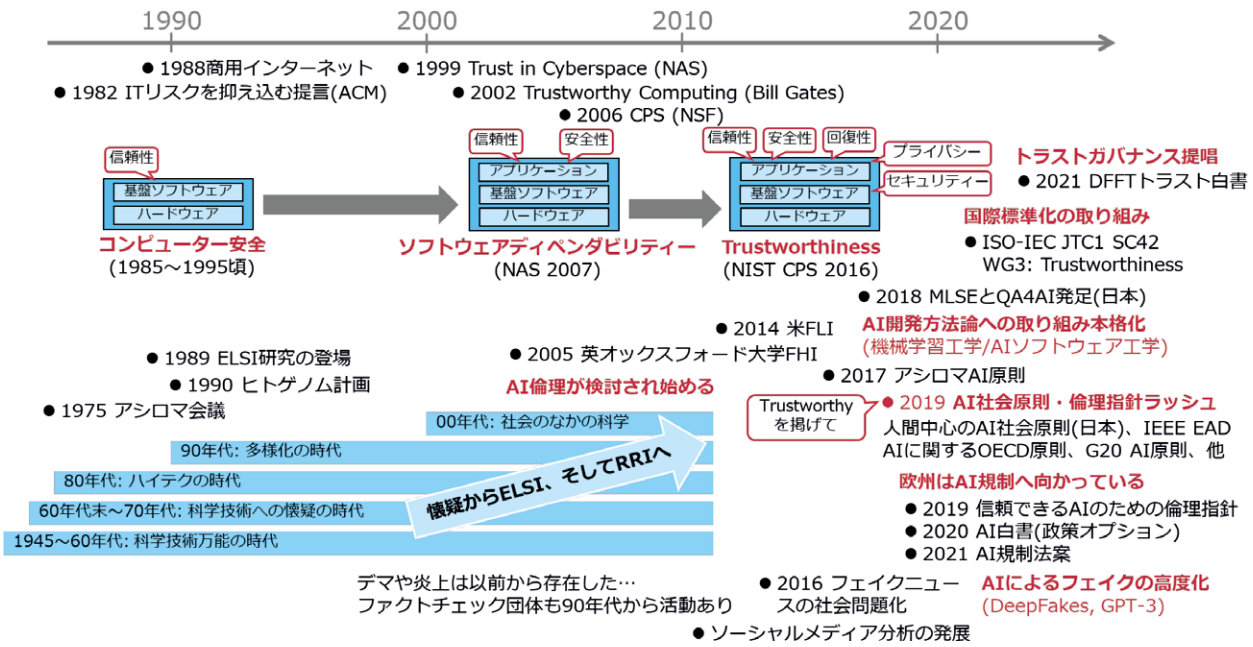


**図3-1-3　　トラスト研究の変遷（3）科学技術のリスク面**

イクニュースが社会問題化し、DeepFakes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。

## 3.1.2 トラスト（信頼）の定義・役割

以上のようにトラストはさまざまな分野で研究されてきている。その中では表3-1-1に示すようにトラストの定義もさまざまである。トラストをどう定義するか自体が研究課題でもあるので本稿で決めつけはしないが、以降の議論のベースとするため、表3-1-1の定義やセミナーシリーズでの講演・議論からほぼ共通的と思われる理解を図3-1-4に示す。

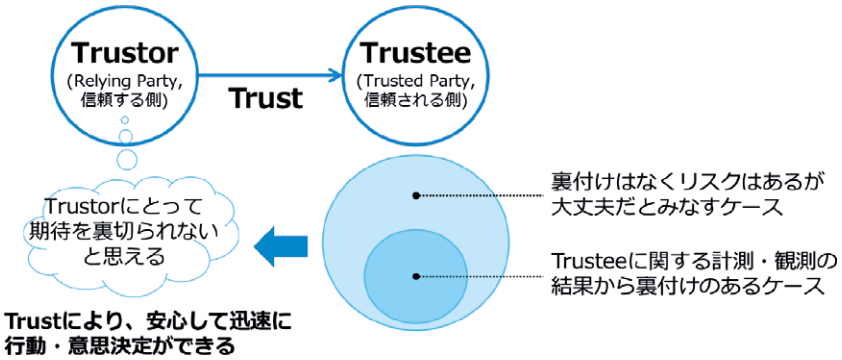


**図3-1-4　　トラストの構成要素と基本的役割**

信頼する側はTrustor（あるいはRelying Party）、信頼される側はTrustee（あるいはTrusted Party）と呼ばれる。 TrustorがTrusteeをトラストする（信頼する）というのは、Trusteeは期待を裏切らないと、Trustorにとって思える状態と考えられる。そのとき、TrustorはTrusteeがこれから行うであろう行為の可能性について、Trusteeに関する計測・観測の結果から裏付けを持って予想しているケースだけでなく、そのよ

うな裏付けはなくリスクはあるのだが大丈夫だとみなしているケースを含む。つまり、リスクはあるのだが、裏切られないと思っているのがトラストしている状態である。逆にトラストできないと、さまざまなリスクケースを考えることになり、人間の思考能力を超え、行動・意思決定がなかなかできなくなる。トラストにより、安心して迅速に行動・意思決定ができるようになる。

**表3-1-1　　トラスト（信頼）のさまざまな定義**

| 分野 | 定義 |
|---|---|
| 社会学 | （信頼は）欠けている情報を内的に保証された安全性に置き換えるのであり、（自分に）利用可能な情報を超えて、行動の期待を一般化することにより、社会の複雑さを減少させる［Luhmann 1979］ |
| | （信頼は）情報不足を内的に保証された確かさで補いながら、手持ちの情報を過剰に利用し、行動予期を一般化することで、社会的な複雑性を縮減するもの、未来における他人の振る舞いによる利益を見越して、未来における他人の振舞い（裏切り）による害が生じうることを認識しつつも、現在において決定を行なうもの［Luhmann 1968?］ |
| | （信頼とは）自然的秩序および道徳的社会秩序の存在に対する期待［Barber 1983］ |
| | （信頼は）他者の誠実さや愛あるいは抽象的な原理への信念を表すような、人やシステムが一群の結果や出来事を実際にもたらすという確信［Giddens 1990］ |
| 哲学 | AがBはCすると信頼するのは、（1）AはBがCすると期待し、（2）このAの期待（1）が、Bが自分の関心を叶えようという動機に基づいているというAの信念か知識に基づいているとき［Hardin 1991］ |
| | AがBはCすると信頼するのは、（1）AはBに重要事Cを任せ、（2）Aはどのようにを扱うのかのコントロールをある程度Bに許し、（3）AはBがCを扱うことができると確信しており、（4）Aは自分に対するBの善意に確信を持っているか、少なくともBの悪意や無関心を予期しないとき［Baier 1986］ |
| | AがBはCすると信頼するのは、（1）AがBの善意に対する楽観的態度を持ち、（2）AはBの予期される行動Cに対するBの能力に対する楽観的態度を持ち、（3）Aは自分が頼りにすることを認識することによって直接BがCするように動機付けられると信じているとき［Jones 1996］ |
| | AがBはCすると信頼するのは、（1）AはCの配慮にあたってBがある社会的規範に内的にコミットしていると期待し、（2）Aは「BがCの配慮にあたってAによって想定されている社会的規範を認識し、また、その規範が何を要求しているかを理解することができる」と確信しており、（3）AはBが自分に課せられた規範に従って行為することができると信じているとき［Mullin 2005］ |
| 経営学 | （信頼は）他者の意図や行動についてのポジティブな期待に基づきリスクを受け入れる意図を含む心理状態である［Rousseau 1998］ |
| 経済学 | （信頼は）1人ないし複数の行為者が特定の行為を遂行するという一定のレベルの主観的確率であり、彼らの行為をチェックすることができる以前に（あるいはチェックすることができる能力とは独立に）、その行為が自分自身の行為に影響を与える状況で形成される［Gambetta 1988］ |
| 社会心理学 | 信頼は、相手の行動によって自分の「身」が危険にさらされる状態で、相手がそのような行動をとらないだろうと期待すること［山岸 1998］ |
| 動物行動学 | 信頼は、他者の誠実さまたは協力への依存、あるいは少なくとも他者があなたを欺かないという期待［de Waal 2009］ |
| 人間工学 | 不確実な状態の中で、相手が自分のゴール達成に協力してくれるという信念［Lee 2004］ |
| 信頼工学 | 信頼 = AIの性能に対する人間による主観的期待値 |
| 情報科学 | 事実を確認しない状態で、相手先が期待したとおりに振る舞うと信じる度合い［Trusted Web推進協議会 2021］ |
| | ある主体が持つある価値観に基づいて、他の主体（含む人、組織、システム）がポジティブな行動をし、かつ／あるいは、特定のネガティブな行動をしないことへの期待［世界経済フォーラム第四次産業革命日本センター 2021］ |

また、トラストの機能を理解するために、関連する他の用語との関係を考えることは参考になる。図3-1-5には、セミナーシリーズの講演の中で、関連する他の用語との関係に言及した部分を抜粋した（それぞれの詳細は該当節を参照）。

(a) 2.3節

| | |
|---|---|
| ・安心 (Assurance)<br>・信頼 (Trust) | 自分を裏切ると、相手自身が損をするから<br>相手が「いい人」だと思うから(＝相手の人間性)、<br>相手が自分に好意を持っていると思うから(=相手との関係性) |

(b) 2.6節

| | |
|---|---|
| ・Security | 心的な脅威（故意による） |
| ・Safety | 身体な脅威（故意でない） |
| ・Reliability | 物理的な脅威（故意でない） |
| ・Trust | 心的な脅威（故意＋故意でない） |
| ・安心 | Emotional part of Trust/Dependability |

(c) 2.4節

| | 知識・能力 | 内容 | 対応 |
|---|---|---|---|
| 代理 agency | 委託者＞受託者 | 不定 | 透明性と監督 |
| 権威 authority | 委託者＜受託者 | 定型的 | 資格認定と敬譲 |
| 信託 trust | 委託者＜受託者 | 不定 | 信認関係としての規律 |

(d) 2.11節

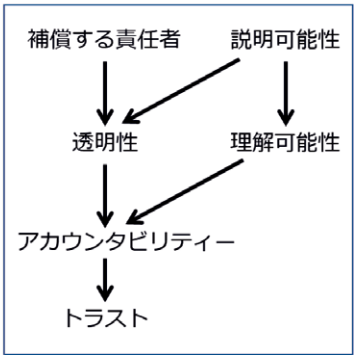


**図3-1-5　　トラストと関連する用語との関係（一つの見方として参考に）**

## 3.1.3 デジタル社会におけるトラスト問題

第１章において、旧来のトラストは身近な人たち同士の信頼関係が中心だったが、デジタル化の進展に伴い、バーチャルな人間関係の広がり、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などにより、トラストを巡る環境変化が起きていることを述べた（図1-2）。そこで、ここでは、そのような変化によってトラストに関わる問題が生じつつある、あるいは、今後起こると思われる、代表的なケースを挙げる。

**【ケース１】医療意思決定におけるトラスト**

医療における意思決定は、従来、患者と医療者の二者関係で行われていた。つまり、患者は医療者の能力・知識に頼りながら、意思決定を行うものだった。しかし、デジタル化の進展により、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。つまり、図3-1-6に示すように、患者と医療者と情報技術（AI）の三者関係の中で医療意思決定が行われるようになってきた[4]。このような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在り方を見いだしていくことが必要になっている。

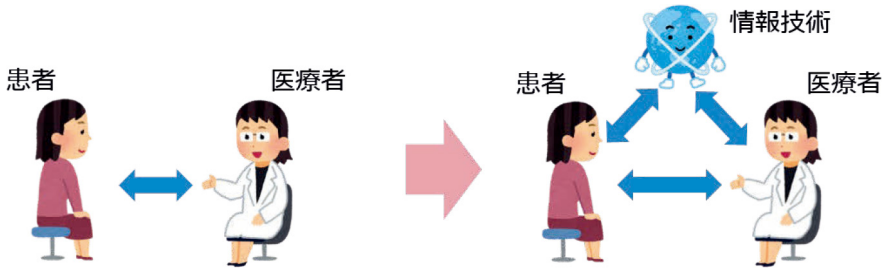


**図3-1-6　　医療意思決定のケース**

4　三者の関係は、2.15節で図2-15-3に示されたようなバリエーションがあり得る。

**【ケース２】自律走行システム**

AI技術を活用することで、固定環境だけでなく、さまざまな状況に応じて運行し得る自律走行システムが、今後、車やドローンなどに広がると思われ（図3–1–7）る。しかし、AI技術を用いた状況認識や走行制御は100%の精度が得られるものではない。その誤りやブラックボックス性によって事故が発生したとき、その原因の解明や責任の所在がどうなるかの懸念がある。社会は、このような自律走行システムをトラストし、受容できるかが問題となる。


図3–1–7　自律走行システムのケース

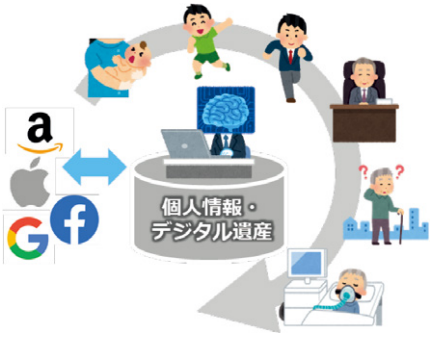


図3–1–8　パーソナルAIエージェントのケース

**【ケース３】パーソナルAIエージェント**

パーソナルAIエージェントは、各個人の個人情報を預かり、さまざまな外部サービスに対して個人情報の提供を含むやり取りを代行する、各ユーザーにカスタマイズされたAIシステムである（図3–1–8）。お薬手帳・母子手帳のような特定用途や、生涯のさまざまな時点での個人情報管理を代行するだけでなく、ユーザーの死後にデジタル遺産の管理を任せられることもあり得る。AI技術はブラックボックス性があり、パーソナルAIエージェントが常に個人ユーザーの意図・期待の通りに振る舞うと100%の保証をすることはできない。このようなパーソナルAIエージェントに、大切な個人情報を委ねることができるのか、また、期待に反する事態が起きた場合に責任の所在はどうなるか、といった懸念が生じる。

**【ケース４】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てしまった。例えばDeepFakesというソフトウェアでは、敵対的生成ネットワーク技術（GAN）によって、本物と見紛うフェイク動画を簡単に作成でき、政治家に思い通りの発言をさせたフェイク動画が作られて政治干渉に使われたり、フェイクポルノ動画が作られて個人攻撃・棄損に使われたりといった問題が既に起きている。また、GPT–3という深層学習による言語生成技術では、人間が書いたような自然なフェイク文章を生成できる。 GANの技術を使えば、なりすまし音声も作ることができてしまう。

これらによって簡単に人をだますことができてしまい、裁判などでの証拠の信頼性も揺らいでしまう。このような技術が使われる以前からデマやフェイクは作られていたが、ソーシャルメディアの普及によって簡単に大規模拡散されてしまうために、大きな社会問題を生んでいる（図3–1–9）。一方、フェイクの法的規制を強くするならば、表現の自由、言論の自由が妨げられる恐れも生じる。


図3–1–9　メディアにおける
フェイク拡散のケース


図3–1–10　人を評価する
AIシステムのケース

**【ケース5】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システムの応用がされつつある（図3–1–10）。また、中国では国民のさまざまな行動履歴を追跡し、個人の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。適切にAI技術を使うことで、評価者による揺れを抑え、公平で客観的な評価がなされるという期待がある一方で、学習データに偏りや差別的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点が指摘されている。また、透明性を確保することがトラストにつながる反面、評価アルゴリズムに過剰適合して行動する人々を生み得るという懸念もある。

**【ケース6】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広がっている。また、さまざまな分野で、シェアリングビジネスや、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（図3–1–11）。

既にさまざまなビジネスが広がっている状況だが、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など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。実際に取引を行った相手を互いに評価し合う相互評価スコアはある程度は有効であるものの、それだけでは不十分だという問題も起きている。


図3–1–11　仮想世界のトラストに基づく
取引のケース

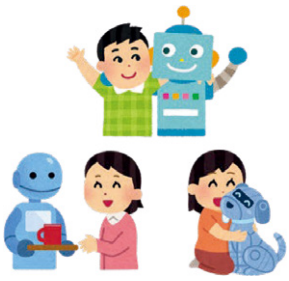

図3–1–12　スマートロボットのケース

**【ケース7】スマートロボット**

親近感を持てる外観、会話を含むインタラクション能力を備えたスマートロボットが、家庭や店舗のような生活空間で人と共存し、さまざまな面で人を支援するようになると期待されている（図3–1–12）。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、そうでないと分かると失望す

るといったことが起きやすい。その一方で、身近なロボットに、過度な親近感・依存感を持ってしまうタイプの人もいる。

### 3.1.4 TrustorとTrusteeに関わる要因とその変化

図3–1–4に示したトラストの構成要素と基本的役割をベースとして、トラストに関わる要因やその変化として注目される点を挙げる。

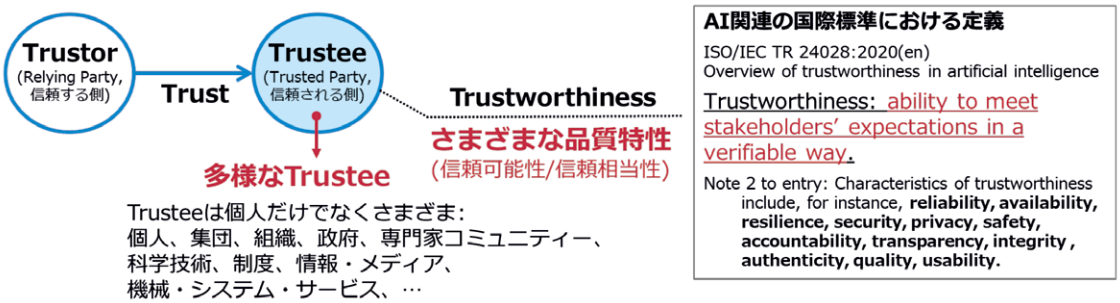


**図3–1–13　　Trusteeに関わる要因の多様化**

1点目としてTrusteeの多様性が挙げられる。Trusteeになるものは、個人だけでなく、集団、組織、政府、専門家コミュニティーのような人が集まって形作られるものも該当するし、さらには、科学技術、制度、情報・メディア、機械・システム・サービスのような人が作ったものも該当する（図3–1–13）。

Trusteeに関わるTrustworthinessという概念がある。Trustworthinessの定義も議論があるが、ここでは、Trusteeについてトラストするに値するかを判断する際に考慮される品質特性をTrustworthinessと呼ぶことにする（図3–1–13）。日本語では「信頼性」と訳されることが多い用語だが、IT分野ではReliabilityやDependabilityも「信頼性」と訳されているので、それらと区別できる「信頼可能性」「信頼相当性」という訳が提案された[5]。

Trusteeの多様化に伴い、Trustworthinessは人に関わる属性だけでなく、機械・システム・サービスなどに関わる属性も含む。例えば、AI関連の国際標準においてTrustworthinessは「検証可能な方法でステークホルダーの期待に応えることができること」と定義され、Trustworthinessの要素として、Reliability、Availability、Resilience、Security、Privacy、Safety、Accountability、Transparency、Integrity、Authenticity、Quality、Usabilityなどが例示されている。

2点目として着目するのはTrusteeの複合化である。図3–1–14の例では、Trustorであるユーザーから見てメインのTrusteeは「システム／サービス」であるが、単にその「システム／サービス」をトラストするかだけでなく、それに携わるベンダー、運用者、開発者、保守者などをトラストするかも複合的に絡む。すなわち、トラストするか／しないか、という一つのケースにおいて、その相手（Trustee）は一つではなく、実際には、複数の相手が複合的に絡んでいることが多々あるということである。

5　本ワークショップにおいて大屋雄裕（慶應義塾大学法学部教授）から提案された。

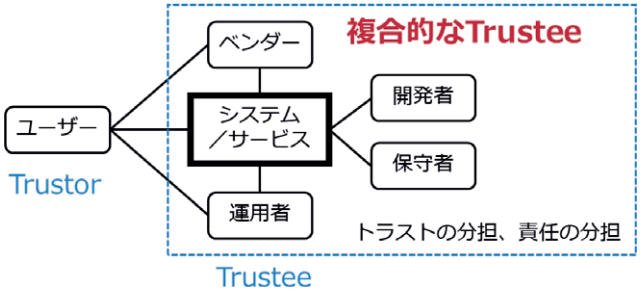


**図3-1-14　　Trusteeの複合化の例**

3点目として、Trusteeが機械・システム・サービスなどのケースにおいて、そこで使われている技術の複雑化・自律化がトラストに与えている影響にも目を向けておきたい。その技術の複雑化・自律化は、Trusteeのブラックボックス化、その動作の予測困難化を招くので、Trustorはトラストしにくくなる。以前であれば、長期間の実績・経験を重ねることでトラストを得ることができたのかもしれないが、今日は技術発展が速いことから、実績・経験を重ねるのに十分な期間の確保が難しくなっている。

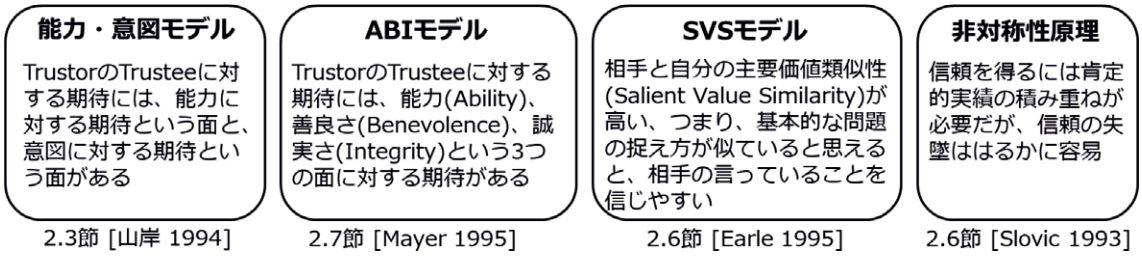


**図3-1-15　　トラストに関するモデル**

4点目として、トラストするか否かは、TrusteeのTrustworthinessだけで決まるものではなく、最終的にはTrustorの主観に左右されるものだという点を、改めて踏まえておきたい。図3-1-15にセミナーシリーズの中で言及されたトラストに関するモデルを挙げたが（詳細は該当節を参照）、特に能力・意図モデル、ABIモデル、SVSモデルはTrustorの主観に依存する面が示されている。すなわち、図3-1-4で裏付けのあるケースと裏付けのないケースがどのような状況ならTrusteeは期待を裏切らないと思えるかは、Trusteeの能力や意図（あるいは能力や善良さや誠実さ）に対するTrustorの主観や、Trusteeとの主要価値類似性に関するTrustorの主観に左右される。


**図3-1-16　　トラストが綻ぶと起こる事態**

3.1.2節で述べたように、トラストがうまく機能していると、ふだんはトラストを特に意識することなく、安心して迅速に行動・意思決定できる。しかし、上述の1点目から3点目で述べたようなTrusteeの多様化・複合化、技術の複雑化・自律化によるブラックボックス化・予測困難化が進むと、4点目として挙げたようなTrustorの主観を左右するさまざまな不確かさが増してしまう。

その結果、不確かさが増してトラストが綻ぶと、例えば図3–1–16のような事態が起こる。考え尽くせないようなさまざまな心配・可能性が気がかりとなり、不信感で眠れなくなるとか、逆に、もう深く考えるのはやめて（ある意味で常に思考停止）、頼りきり、任せきりになるとかが起きてしまう。また、リスクは減っていないにもかかわらず、Trusteeと親密な会話を持つことで信じやすくなることから、悪意を持った他者からだまされやすいという側面もある。頼りきり、任せきりで、常に思考停止という状態は好ましいものではなく、トラストすることは無条件に良いことだというわけでない。

### 3.1.5 取り組みの状況・方向性

第1章で述べたデジタル社会におけるトラスト形成に関わる問題意識や、3.1.4節で触れたトラストの綻びが引き起こす問題、より具体的には3.1.3節に示したトラスト問題の7つのケースなどに対して、対策につながる技術的な取り組みが進められている。

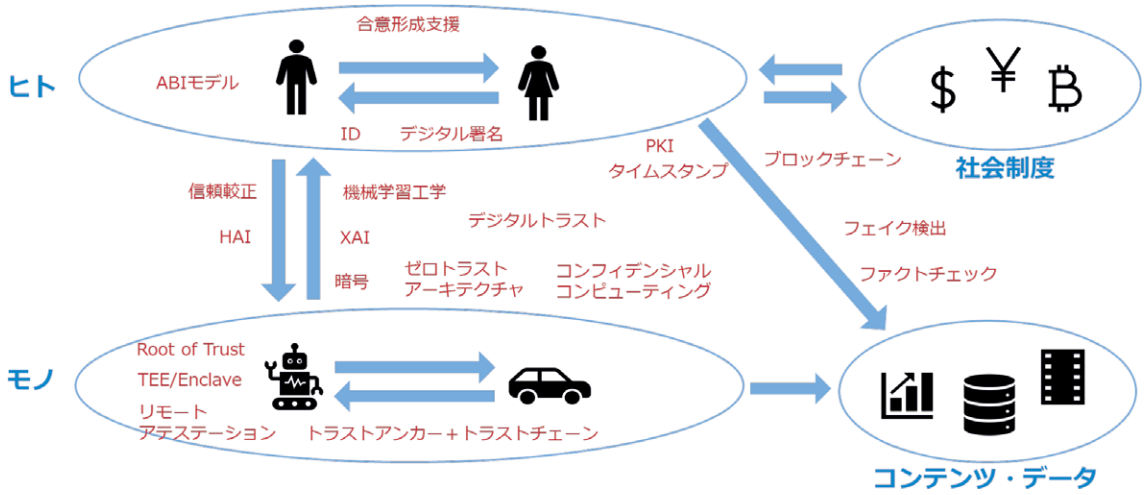


**図3–1–17　　技術的な取り組みの状況**

図3–1–17には、TrustorやTrusteeになるものを、ヒト、モノ、コンテンツ・データ、社会制度に大きくカテゴライズした上で、第2章（俯瞰セミナーシリーズ）で言及された主な技術をプロットした（現状の技術的な取り組みを網羅できているわけではない）。

このような取り組みとして、ある程度まとまった技術分野となっているものを、表3–1–2に示した。これらは情報科学系の技術分野であるが、人文・社会科学系の研究からこのような面の知見が得られるのではないかと期待されることのいくつかを表3–1–3に挙げた。

**表3-1-2　　トラストと関わりの深い情報科学系の主な技術群**

| 分野 | 取り組まれている技術の例 |
|---|---|
| デジタルトラスト | ゼロトラスト環境におけるあらゆるヒト・モノの真性性保証の仕組み：デジタル証明、Hardware Root of Trust、Secure Boot/Trusted Boot、Remote Attestation、Secure Enclave、Trusted Execution Environment（TEE）、Confidential Computing、デジタル署名、タイムスタンプ、eシール、eデリバリー、自己主権型アイデンティティー、ブロックチェーン、… |
| 信頼されるAI | 機械学習応用システムの安全性・信頼性を確保する開発方法論・技術体系（機械学習工学）：機械学習品質マネジメントガイドライン、機械学習テスティング手法、説明可能AI（XAI）、公平性配慮機械学習、プライバシー配慮機械学習、Safe Learning、運用時の機械学習品質維持手法、フェイク検知・ファクトチェック、… |
| HAI（Human Agent Interaction） | 信頼工学・信頼較正、適応ギャップ、メディアの等式、ナッジ、意図スタンス、… |

**表3-1-3　　人文・社会科学系の知見への期待（例）**

| 分野 | 関わり得ると考えられる知見の例 |
|---|---|
| トラストに関する基礎的理解 | 社会変化・デジタル環境の中でトラストに関わる要因とトラスト形成の関係評価、トラストが壊れる要因・状況の解明、人間と人間の間のトラスト関係から人間と人工物の間のトラスト関係への対象拡大、トラストに関する基礎的理解に基づくトラスト向上策の提案、脳科学・発達科学との連携、… |
| トラストを支える制度設計 | 技術開発状況とその社会受容の状況や人々のリテラシーの状況などを踏まえたトラスト形成の課題抽出、適した制度設計（ハードロー、ソフトロー、保険制度など）、… |
| トラスト関連技術の社会受容や定着 | トラストガバナンスフレームワーク、ガバナンスエコシステム、先端技術・複雑技術の社会受容に関わる要因とトラストの関係、科学技術のリスク・不安に関わる分析・評価、トラスト判断のため教育施策・リテラシー向上施策、… |



これまで述べてきたように、さまざまな分野でトラストに関わる研究が行われている。しかし、それらの間の知見の共有や連携はまだ非常に少ないように思える。3.1.3節に示したトラスト問題のケースを見ても、情報科学系の技術開発だけでは解決できず、人文・社会科学系の分析・知見を生かすことが不可欠である。そこで、情報科学系と人文・社会科学系を合わせた総合知としての取り組みの方向性を考えていきたい。

JST CRDSでは、その方向性についての検討を深め、その戦略提言を策定していきたいと考えているが、現時点では、大まかなイメージを図3-1-18に示す。デジタル社会のさまざまなケースでトラストがうまく機能するようにするためには、技術開発だけでなく、法制度・保険などを含む制度設計や、経験の共有やリテラシー向上も含む教育・啓発を組み合わせていく必要があると考える。それらの手段を組み合わせて、トラストがうまく機能するように枠組み全体を設計・管理する取り組みを、ここでは「トラストガバナンスフレームワーク」と呼ぶ。

図3-1-18の下半分には、このような枠組みを実現するには、総合知による取り組みが必要であることを示した。二重の円は図3-1-4に対応している。内側の円は「Trusteeに関する計測・観測の結果から裏付けのあるケース」であるが、このケースを拡大するために情報科学系の技術開発は有効である。一方、その外側の「裏付けがなくリスクはあるが大丈夫だとみなすケース」は、人間の主観、思考・心理面に依存する部分であるため、人文学系の研究によってその理解が深められる。また、これらの状況を踏まえた適切な制度設計を行うことは、社会科学系が担うものである。情報科学系、人文学系、社会科学系の研究が必ずしも、この3通りに分かれるわけではないが、3つの面から総合的な枠組みを考えていく総合知が求められるのは間違いない。

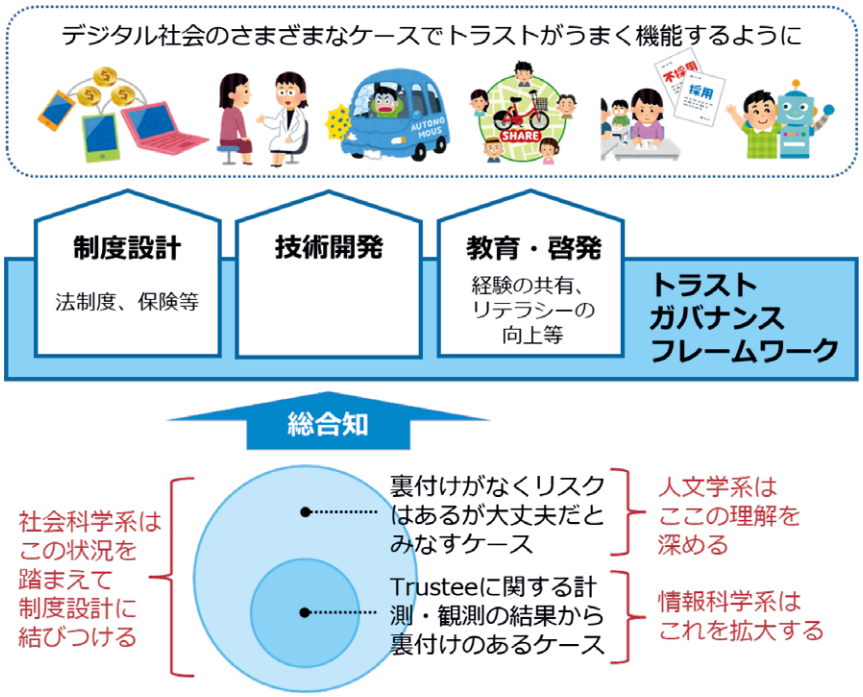


**図3-1-18　　総合知による取り組みの方向性（イメージ）**

## 3.2 総合討議

本節にはワークショップにおける総合討議において出された意見や示唆をまとめた。その一部は3.1節に取り込んだが、そこにまだ取り込めていない観点や、より掘り下げた意見などが含まれている。

### 3.2.1 トラストについての多面的な理解

・トラストの意味分類として、サービスが期待通りに動くというトラスト、情報が本物であるというトラスト、相手の人間が本物であるというトラストがある（コラム2参照）。
・Trusteeの特性が客観的に条件を満たすかと、それを受けてTrustorの内的心理状態としてトラストが成立するかは分けて考えるべき。
・TrustorがTrusteeをトラストするという一方向だけでなく、双方向のトラスト関係も考えるべき。
・IT分野では、ReliabilityやDependabilityの訳語として「信頼性」が使われることが多い。それらと区別できるように、Trustworthinessの訳語として「信頼可能性」「信頼相当性」を使ってはどうか。
・トラストは人間が日常的に無意識に行っているという面がある。日常的には、人に対するトラスト、社会に対するトラスト、技術に対するトラストなどを一つ一つ分けて意識はしていない。リスクが見え始めると、トラストを意識するようになる。今日、技術発展にはリスクもあるということから、トラストが意識されるようになった。
・トラスト状態は、間違ってトラストしているというケースもあり得る。そもそも何が真実なのかは分からないとも言えるので、結局は誰を信じるかと、Chain of Trustがポイントになるのではないか。
・トラストしている状態が常に良いというわけではない。トラストによって安心して意思決定できるという面がある一方、オートパイロットからマニュアル運転に切り替えるように、適切に不信を持ってベリファイすることも必要である。
・カメラによって被写体が写真にどう撮られるかは、カメラの設定や撮り方によって変わるので、写真であっても世界に存在したことの一つの表現に過ぎない。しかし、それでも存在しないものは写らないだろうというトラストを我々は持っている。その意味で、ファクト自体がトラストによって構成されたものだと考えられる。
・ワクチンの安全性に関してCDCが言ったら一応はファクトだと捉えようというような例が挙げられるが、社会的なガバナンスシステム（一種の権力）によってトラストが支えられるという面はある。
・社会制度を背景としてトラストが生まれることもある。
・1996年の国民生活白書に初めて「安心安全」が使われ、それ以降ずっと行政で使われている。安全と安心はかなり違うものなのにセットで使いすぎる。どちらの話なのかを明確にすべきという指摘がしばしばされている。

コラム2

## トラストの3側面

俯瞰ワークショップにおいて、大屋雄裕（慶應義塾大学教授）からトラストには異なる3つの側面があるという指摘があった。今後の検討において重要なポイントになると考え、ワークショップ後も意見交換を行い、以下のような3側面として整理した。

① 対象真正性：本人・本物であるか？
② 内容真実性：内容が事実・真実であるか？
③ 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？

これらの側面のそれぞれについて、信じることができるかを確認・判断することは、状況によって必ずしも容易でない。また、デジタル化の進展によって、その確認・判断がさらに難しくなっているようなケースも生じている。

対象真正性については、印鑑・サイン、身分証・鑑定書、デジタル署名、さまざまな認証技術などが手掛かり・よりどころとして用いられている。しかし、それらの偽造・偽装やなりすましも絶えない。

内容真実性については、そもそも事実・真実は絶対的なものが一つ定まるというようなものではない。事実は、観測装置・手段によって観測結果にノイズやバラツキが生じるし、人の五感・感性も人それぞれである。また、学説は真だと信じられたものが、後に覆されることがある。さまざまな手掛かりから一定の確度を持って社会的共通認識が形成されたものが、事実・真実に相当するものと扱われ得るということなのかもしれない。デマやフェイクが拡散されやすくなり、それらに惑わされるリスクが高まっている。

予想・対応可能性については、対象の振る舞いとして想定されるものが事前に説明され得るならば、ある程度安心できる。人の振る舞いやタスクであれば契約・資格認定など、機械・システムであれば仕様書などがその材料になる。しかし、ブラックボックスAIや大規模複雑システムのように、事前に説明されないとか、説明されても理解することが難しいといったタイプのものについては、そのような形での安心は得られない。

このような3つの側面からデジタル社会におけるトラストの在り方を考えていくことは、今後の重要な研究課題だと考えている。ここでは課題の指摘にとどめ、取り組みの方向性・可能性については、機会を改めて論じることにしたい。

### 3.2.2 トラストの変化要因およびトラスト研究の変遷

・ビッグデータ時代になって、さまざまな情報が取得され、関係も複雑化したことで、旧来のトラストの捉え方ができなくなっている。
・経済学では合理性だけを考える傾向があったが、その考えでは囚人のジレンマなど、トラストさせることが難しい状況が問題になる。合理性だけでなく、心理面をどう織り込んでいくかが課題となってきた。
・昔のITは善意の下で作られていて信用できていたが、コンピューターウイルスが出てきたことで、作る側の悪意や技術の悪用の可能性が意識されるようになった。故障だけでなく悪意や攻撃もあるということで、ITにおけるリスクの捉え方が変わった。最近のAIにおいてもリスクベースの考え方がされている。リスクとトラストは表裏一体だと思える。
・セキュリティーを担保できるようになったとき、その立場を悪用して権力に変えるようなことをすると、ユーザーがコントロールできないことになってしまう。
・ナッジでも知られるアメリカの憲法学者Cass Sunstein（キャス・サンスティーン）は、我々が意思決定する事項が多くなりすぎて、全てを自主的・自律的に判断することはもう無理だと言っている。選択しないことを選択すると「Choosing Not to Choose」という本を書いた。どうしてもという自分の興味関心領域を守り、そこに自分の時間を使う。何をトラストに任せて、何はトラストしないかというメタ選択が人生の中心になってくると、サンスティーンは指摘している。
・ベリファイ可能なTrustworthinessに対して、それを高める技術と、それを伝える技術がある。 Device Integrityやリモートアテステーションなどのベリファイ技術が伝える技術の例で、現代はベリファイにコストがかからなくなってきた。
・ビジネス視点では、情報サービスなどにおけるトラスト獲得の競争が起きているという面もある。
・社会課題解決がマルチステークホルダーに関わり、バーチャルな関係でのトラストを考えることが必要なケースが増えている（例えば個人情報管理における第三者提供）。マルチステークホルダーのトラストやインセンティブ設計、それを支える制度設計などが重要課題になってくる。

### 3.2.3 トラスト問題のケース・事例

・パーソナルAIエージェントのケースは、広く捉えれば意思決定支援サービスあるいはレコメンドシステムの問題と考えられる。ただし、従来のレコメンドシステムの多くはサービス提供側にいるため、サービス提供側の意図・都合が反映されることがあるのに対して、パーソナルAIエージェントはユーザー側にいるのでそれを回避できる。
・パーソナルAIエージェントはユーザー側に置く意思決定支援だが、必ずしもユーザーのコントロール通りにならないという点に問題がある。
・トラストする際、Trustorは「悪い状況にならない」と必ず思うとは限らない。医療シーンでは、悪い状況になるとしても納得して受け入れるケースがある。
・自動走行のケースについて、誰がTrustorかというと、乗っている人や通行人など、関係者がいろいろ考えられる。
・メディアのケースと仮想世界の取引のケースには、アイデンティフィケーション、真正性保証の問題が絡んでいる。
・メディアのケース、フェイク問題では、拡散力の変化・拡大という要因がある。また、マクロ的な社会への影響として、社会不安（震災・感染症など）や社会を二分する出来事（選挙など）のときに、前提条件を変えてしまい、社会の分断を招き、民主主義を揺るがす。フェイクが一部存在するというだけで、全体を疑わなければならなくなるという問題も起きる。全ての情報の価値を棄損し、人々のコストを大幅に

上げることになる。
・必ずしも信頼してもらわなくてもよいはず。トラスト問題については、どういう場面で信頼がないと、社会的にどんな大きな問題が起きるかを考えるべき。
・各ケースをトラストの基本構造（図3–1–4）と対応付けて整理してはどうか。

### 3.2.4 意思決定の主体性に関わる問題と日本人のメンタリティー

・レコメンデーションは、選択肢を増やしてくれて、それを選んでも選ばなくてもよい。選ばなければ何も起こらない。その一方で、選択肢や情報を絞り込んで、だいたいこれでいいでしょうと薦めるタイプや、デフォルトでは実行されてしまうタイプもある。ランキングするタイプも、リテラシーの高い人でないと、絞り込まれたのと同じになる。このような選択肢を絞ってしまうタイプは要注意である。Trustworthinessのガバナンスの問題と考えることもできる。
・トラストの確立によって、患者の意思決定の自律性が迫害されるという懸念を感じている。コロナ問題でも、あまり考えずに怖いという感覚でゼロリスクっぽい方に行ってしまう人もいる。リスクとベネフィットのバランスをよく考えることなく、安易なトラストによって、自分らしく生きるという意思が薄れていくという問題がある。医療・ヘルスケア分野では特に、医療者のような専門家がパワーを持っていることになり、このような問題を生みやすいことを懸念している。
・医療者と患者のShared Decision–makingでは、医療者のコミットメントという観点が重要になる。医療者は押し気味にレコメンドするか、引き気味に伝えるか、というオプショントークの中にコミットメントが見える。
・日本人はリスク社会でパターナリズムを求める（主体性なきパターナリズム）。信頼は意思決定やリスクテイクに関わるが、リスクテイクしたくない人たちにとって、レコメンデーションやナッジがどう働くか。（最近はパターナリズム的な社会から自分でリスクテイクする方向に少しずつだが変わりつつあるようにも見えるが）
・西洋諸国の感覚では、患者は自分なりに収集した情報や考え方を持っているものであって、医者を一度トラストして受け入れたからといって、その後も医者の言うままになるわけではない。患者がトラストしたことで主体性・自律性がなくなるというのは、空気を読んで我を抑える日本人の問題である。信頼の意味も異なってくるはずである。
・ゼロリスクを求める一般市民・生活者という見方は必ずしも適切ではない。通常は安心や信頼を意識せずに生活していて、リスクが出てくると意識するようになる。それで、生活者が「安心できない」「リスクがゼロでない」と言うときは、多くの場合、実は国・自治体を信頼できないと思っていて、その代わりにそういった言い方になりやすい。

### 3.2.5 トラスト研究としてのさまざまな論点

・TrustorがTrusteeをトラストできるかにはレベルがあるのではないか。「徳」レベル、「意図」レベル、「機能・行動」レベル、「結果」レベルのような整理が考えられる。
・安全、安心、信頼の関係・構造を考えつつ、物理的なリスクの軽減と心理的な安心の関係や、リスク状況においてTrustorの安心がどう生まれるかといった点も整理したい。
・近代において、あらゆる人はある分野で専門家であり、別の分野では素人であるという社会である。そういう社会で専門家に対する信頼をどう考えるか。ギデンズはそういう専門家とリスクや信頼の関係を論じた。
・ブラックボックス問題は専門家システムの問題に通じるところがある。しかし、AIの問題は、専門家の責

任に帰着されるとは限らない面もある。例えば、運用中にブラックボックスに新たな恣意性を持たせ得るとか、単体AIがトラストできたとしても多数のAIが連携するとブラックボックス化してトラストできないとか、人間には追跡できない処理速度で起きる問題とかがある。アクターネットワークも考える必要があるかもしれない。

- AI技術は100%の保証は無理なので、その可能性を高めることには取り組みつつも、保険や免責のようなものも考えていく必要がある。
- ゼロリスクではないAIを使っていくとき、どう受け入れていくか。技術側からは、問題が起きたときに修正しやすい技術にしていくことで、これはTrustworthinessの一つの流れかもしれない。一方、それを使っていく人間の方からすると、皆の共通認識としての正しい世界・正義のようなものに照らして、その範囲に収まっている限り、不完全でも使っていこうという考え方ができる。欧州ではそこが定められていると思うが、日本では議論がされていない。これを技術系だけで考えるのは無理で、人文・社会系に期待したい。
- 社会制度がトラストの背景にもなるというとき、ある社会制度が内生的に出現したケースと外生的に与えられたケースを分けて捉え、その上で制度の安定性がどのように保たれているかという分析は、経済学的な観点からも興味深い。
- 個人、家族・少人数グループ、組織（会社など）、国家というようなレベルを考え、それぞれの間のトラストを別に分けて考えた方が良いかもしれない。
- TrustorがTrusteeをトラストできるかは、最終的には主観的判断になるが、Trustorが判断するために十分な情報を提供することが、制度では大切になると思える（例えば医師免許）。例えば、賠償請求のようなケースでは、「通常の一般人が合理的に意思判断するために必要な情報が伝えられているか」「合理的な意思判断を妨害するような特殊事情（長時間の拘束など）がなかったか」といった点が法的には重要になる。これはTrustworthinessのガバナンスシステムと位置付けられるし、このようなガバナンスシステムの存在が当事者に知られていて、信頼もされているということが求められる。
- Trustworthinessのベリファイにおいて、膨大な量の関連コンテンツや行動履歴データの検索・解析は人間には無理なので、AI技術の活用の可能性がある。 AI技術を使うことで、リアルタイム会計検査のような意味合いのことが、Trustworthinessのベリファイでもできるかもしれない。
- 目指すものの一つについて「トラストガバナンスフレームワーク」という言葉を使っているが、ある狭い意味でのトラストに関する「トラストフレームワーク」（ガバナンスも含む）は20年以上前から取り組まれて実装されている。別の言葉を考えた方が良いかもしれない。
- 従来の「トラストフレームワーク」は、ポリシーからトラストモデルを作り、トラストフレームワークで維持していくものである。トラストドメインの中での管理と、マルチドメインにまたがる管理があり、それらの間でポリシーの整合を図るのがとても難しい。「トラストガバナンスフレームワーク」でも同様の難しさがあるのではないか。これまでの経験・知見を生かせないかと思う。
- トラストが生まれたり、不信の状態がトラストに変わったりと、トラスト関係が時間的に変化していくことにも目を向けるべき。また、継続的な信頼と刹那的な信頼（詐欺など）とを区別して議論したい。トラストした結果を評価して、次にトラストするかに継承していくという流れもある。マネジメントサイクルの概念を取り入れるのがよい。
- トラスト関係の推移・時間変化は興味深く重要な観点だが、法律家は扱わない。法律家が扱うのは、個人の主観による個人差の部分は考えず、客観的なTrustworthinessを扱う。このTrustworthinessのガバナンスの問題と、トラストの推移のような観点は分けて考えるべき。

### 3.2.6 「総合知」による研究推進における課題

・分野横断・学際的研究の意義が理解されるようになりつつあるものの、論文にならないと業績にならない。業績評価・処遇の面が伴っていないため、結局、取り組む研究者が増えないという問題を聞いている。
・人文・社会科学分野では、各分野の研究論文が重視される傾向がある。共同で本を書くことはできても、学際的研究で査読論文を書くことは難しい。「総合知」に関わる学際的研究で論文が書けたとしても、そのような研究者のポストがあまりないため、キャリアパスが描けない。
・欧州ではRRI（Responsible Research Innovation）が掲げられ、人文・社会系が社会課題解決に参加するという文化ができつつある（まだ参画している人は少ないが）。日本では全然少ないのは、ポストがないという問題が大きい。研究費よりもポストを用意しないと、学際的研究は必要であっても進まない。また、学際的研究の成果を評価するようなシステムに変えていかなくては、研究者は入ってこない。
・一部の大学ではそのような研究のための組織ができ、以前に比べると、ポストを作ろうという動きは出てきている。しかし、そのポストが時限的なので、その期間に業績を上げるためには論文を書かねばならないため、論文を書きやすいテーマに向かいやすいというわけで、問題は根深い。
・論文を書きやすいテーマを押さえつつ、新しい方向性も少し加えていくといった手法を取らざるを得ないというのが実態だろう。
・理系の側で分離融合の大型プロジェクトを立てて、そこに文系も含めて人をたくさん入れて進めて実績を上げるというのが一つの手だと思う。人文・社会科学の側から大型プロジェクトを立てるのは難しいだろう。
・特にAI研究では、人文・社会系の知見の必要性を感じている人たちは多い。説明可能AIにしろ、フェイク問題にしろ、人間の側に目を向けないといけない。
・社会的な課題に対する研究ターゲット設定では、情報系の価値観だけでなく、人文・社会系の価値観からの貢献を大きく期待したいと思っている。
・それは正しい方向性だが、現状はまだギャップが大きい。人文・社会系は問題の捉え方やスケール感が情報系と異なり、大きすぎる問題や個別すぎる問題を取り上げて、すり合わせが難しいことが多い。ギャップを埋めていくために、研究者間の信頼関係も必要である。
・社会科学は社会を観察する科学であって、問題を解決することには関心がない。社会を運営する側に関わってしまうと独立性が確保できなくなる。ただし、法学・行政学や社会心理学は実学に近く、工学と組んで問題解決をすることに関心を持ちやすい。
・文理連携の形として、例えば、文献を読んで論点を整理するサーベイや、周辺も含めて視野を広げるための取り組みは、人文系が得意と思う。
・IT系のプラクティカルな取り組みと、人文系の関連研究を丁寧に調べる取り組みとで、時間的な感覚も異なる。共同プロジェクトをいますぐ一緒にやるのも理想だが、今回のセミナーシリーズやワークショップのような場は素晴らしく、このような機会を繰り返していくことはとしても有意義だと思う。 IT系は近視眼的な目標設定・取り組みになりがちなところもあり、人文系との交流から、もっと本質的なところを押さえるようにすべきと思う。
・人文・社会系も変わっていく必要があると感じている。人文・社会系とひとくくりでいうが、人文・社会系の中の横のつながりは弱く、各分野は孤立しているようにも思う。スコープを広げていく必要があると感じるところもある。今回のセミナーシリーズワークショップでは他分野のことを知ることができ、このような機会は非常に重要と感じた。
・情報系は言葉の使い方が浅いと感じる。使いやすい言葉を使って、それが広まる。飾りとして「トラスト」という言葉を使っているようなところがないか。人文系は言葉に敏感なので、連携してよく考えていくことが望ましい。

# 参考文献リスト


**第1章**

［1］ 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）.

［2］ EY総合研究所, 報告書「人工知能が経営にもたらす創造と破壊」(2015年9月).

［3］ 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）.

［4］ 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).

**第2章**

**2.1節**

［1］ 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).

**2.2節**

［1］ 上出寛子,「社会心理学における信頼」,『信頼を考える：リヴァイアサンから人工知能まで』(小山虎編著, 勁草書房, 2018年）, pp. 137-156（6章).

［2］ Carl I. Hovland and Walter Weiss, “The Influence of Source Credibility on Communication Effectiveness,” *Public Opinion Quarterly*, Vol. 15, No. 4, pp. 635–650 (1951). DOR: 10.1086/266350

［3］ Anne Marthe van der Bles, et al., “The Effects of Communicating Uncertainty on Public Trust in Facts and Numbers,” *Proceedings of the National Academy of Sciences*, Vol. 117, No. 14, pp. 7672–7683 (2020). DOI: 10.1073/pnas.1913678117

［4］ Judee K. Burgoon, Thomas Birk, and Michael Pfau, “Nonverbal Behaviors, Persuasion, and Credibility,” *Human Communication Research*, Vol. 17, No. 1, pp. 140–169 (1990). DOI: 10.1111/j.1468–2958.1990.tb00229.x

［5］ Susan T. Fiske, Amy J.C. Cuddy, and Peter Glick, “Universal Dimensions of Social Cognition: Warmth and Competence,” *Trends in Cognitive Sciences*, Vol. 11, No. 2, pp. 77–83 (2007). DOI: 10.1016/j.tics.2006.11.005

［6］ Solomon E. Asch, “Forming Impressions of Personality,” *The Journal of Abnormal and Social Psychology*, Vol. 41, No. 3, pp. 258–290 (1946). DOI: 10.1037/h0055756

［7］ Seymour Rosenberg, Carnot Nelson, and P. S. Vivekananthan, “A Multidimensional Approach to the Structure of Personality Impressions,” *Journal of Personality and Social Psychology*, Vol. 9, No. 4, pp. 283–294 (1968). DOI: 10.1037/h0026086

［8］ Susan T. Fiske and Cydney Dupree, “Gaining Audiences’ Trust and Respect about Science,” *Proceedings of the National Academy of Sciences*, Vol. 111, Supplement 4, pp. 13593–13597 (2014). DOI: 10.1073/pnas.1317505111

［9］ Philipp Kulms and Stefan Kopp, “A Social Cognition Perspective on Human–Computer Trust: The Effect of Perceived Warmth and Competence on Trust in Decision–Making with Computers,” *Frontiers in Digital Humanities*, Vol. 15, No. 14 (2018). DOI: 10.3389/fdigh.2018.00014

［10］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).

［11］ ニクラス・ルーマン,『信頼：社会的な複雑性の縮減メカニズム』(勁草書房, 1986).
［12］ Bernard Barber, *The Logic and Limits of Trust* (Rutgers University Press, 1983).
［13］ Angelo Romano, Daniel Balliet, Toshio Yamagishi and James H. Liu, "Parochial Trust and Cooperation across 17 Societies," *Proceedings of the National Academy of Sciences*, Vol. 114, No. 48, pp. 12702–12707 (2017). DOI: 10.1073/pnas.1712921114
［14］ Oriel FeldmanHall et al., "Stimulus Generalization as a Mechanism for Learning to Trust," *Proceedings of the National Academy of Sciences*, Vol. 115, No. 7, pp. E1690–E1697 (2018). DOI: 10.1073/pnas.1715227115
［15］ Partnership on AI, "Human–AI Collaboration: Key Insights from a Multidisciplinary Review of Trust Literature" (2019). https://partnershiponai.org/paper/human–ai–collaboration–trust–literature–review–key–insights–and–bibliography/
［16］ Elinor Ostrom, "A General Framework for Analyzing Sustainability of Social–Ecological Systems," *Science*, Vol. 325, No. 5939, pp. 419–422 (2009). DOI: 10.1126/science.1172133
［17］ Edmond Awad, "The Moral Machine Experiment," *Nature*, No. 563, pp. 59–64 (2018). DOI: 10.1038/s41586–018–0637–6

**2.3節**

［1］ Toshio Yamagishi and Midori Yamagishi, "Trust and commitment in the United States and Japan", *Motivation and Emotion* Vol. 18, No. 2, pp. 129–166 (1994).
［2］ George A. Akerlof, "The Market for "Lemons": Quality Uncertainty and the Market Mechanism", *The Quarterly Journal of Economics*, Vol. 84, Issue 3, pp. 488–500 (1970).
［3］ Ernst Fehr and Bettina Rockenbach, "Detrimental effects of sanctions on human altruism", *Nature*, Vol. 422, No. 6928, pp. 137–140 (2003). DOI: 10.1038/nature01474
［4］ Toshio Yamagishi, Karan S. Cook, Motoki Watanabe, "Uncertainty, Trust, and Commitment Formation in the United States and Japan". *American Journal of Sociology*, Vol. 104, Issue 1, pp. 165–194 (1998).
［5］ Ernst Fehr and Simon Gächter, "Altruistic punishment in humans", *Nature*, Vol. 415, No. 6868, pp. 137–140 (2002). DOI: 10.1038/415137a

**2.5節**

［1］ 神里達博,「情報技術におけるELSIの可能性：歴史的背景を中心に」,『情報管理』, Vol. 58, No. 12, pp. 875–886（2016).
［2］ Jean E. McEwen, et al., "The Ethical, Legal, and Social Implications Program of the National Human Genome Research Institute: reflections on an ongoing experiment", *Annual Review of Genomics and Human Genetics*, Vol. 15, pp. 481–505 (2014). DOI: 10.1146/annurev–genom–090413–025327
［3］ Robert K. Merton, "The Normative Structure of Science", *The Sociology of Science: Theoretical and Empirical Investigations* (Robert K. Merton (ed.), University of Chicago Press, 1973).
［4］ 村上陽一郎,『科学者とは何か』(新潮社, 1994).
［5］ Richard Owen, et al., *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society* (Wiley, 2013).
［6］ エルンスト・シューマッハー（著）, 小島慶三・酒井懋（訳）,『スモールイズビューティフル』(講

談社, 1986年).
［7］ ジョン・ザイマン（著）, 村上陽一郎・他（訳）,『縛られたプロメテウス―動的定常状態における科学』(シュプリンガー・フェアラーク東京, 1995).

**2.6節**

［1］ L. Lamport, R. Shostak, M. Pease, “The Byzantine Generals Problem”, *ACM Transactions on Programming Languages and Systems*, Vol. 4, No. 3, pp. 382–401（1982). DOI：10.1145/357172.357176
［2］ 佐藤一郎,「「ビザンチン将軍問題」とは何か」,『NII Today』, Vol. 69（2015). https://www.nii.ac.jp/today/69/4.html
［3］ Michael Burrows, Martn Abadi and Roger Needham, “A logic of authentication”, Digital Equipment Corporation SRC Report SRC–RR–39 (1989). https://www.hpl.hp.com/techreports/Compaq–DEC/SRC–RR–39.pdf
［4］ Michael Burrows, Martin Abadi and Roger Needham, “A logic of authentication”, *ACM Transactions on Computer Systems*, Vo. 8, Issue 1, pp.18–36 (1990). DOI: 10.1145/77648.77649
［5］ 長谷部浩二・岡田光弘,「BAN論理からProtocol Composition Logicへ: セキュリティプロトコルの論理的検証法」,『応用数理』, Vol. 17, No. 4（特集:数理的技法による情報セキュリティ）, pp. 311–322（2007). DOI: 10.11540/bjsiam.17.4_311
［6］ A. Whitten and D. Tygar, “Why Johnny Can’t Encrypt: A Usability Evaluation of PGP 5.0”, *Proceedings of the 9th USENIX Security Symposium*, pp.169–184 (1999).
［7］ L. J. Camp, "Design for Trust", *Trust, Reputation and Security: Theories and Practice* (RinoFalcone (ed), Springer–Verlang, 2003).
［8］ J. Riegelsberger, M. A. Sasse, and J. D. McCarthy, “Privacy and trust: Shiny happy people building trust?: photos on e–commerce websites and consumer trust”, *Proceedings of CHI2003*, Vol. 5, No. 1, pp. 121–128 (2003).
［9］ B. Friedman, P. H. Khan and D. C. Howe, “Trust online”, *Communications of the ACM*, Vol. 43, Issue 12, pp. 34–40 (2000).
［10］ S. P. Marsh, “Formalizing trust as computational concept”, Ph.D. thesis University of Stirling (1994).
［11］ R. Anderson, I. Shumailov, “Situational Awareness and Adversarial Machine Learning – Robots, Manners, and Stress” (2021). https://www.cl.cam.ac.uk/~rja14/Papers/situational–awareness2021.pdf
［12］ S. Xiao and I. Benbasat, “The formation of Trust and Distrust in Recommendation Agents in Repeated Interactions: A Process–Tracing Analysis”, *Proceedings of ICEC 2003*, pp. 287–290 (2003).
［13］ R. T. Stephens, “A framework for the identification of electronic commerce design elements that enable trust within the small hotel industry”, *Proceedings of ACMSE’04*, pp.309–314 (2004).
［14］ Tetsuro Kobayashi and Hitoshi Okada, “The Effects of Similarities to Previous Buyers on Trust and Intention to Buy from E–Commerce Stores: An Experimental Study Based on the SVS Model”, *IT Enabled Services* (Springer–Verlag Wien, 2013), Vol. 9783709114254, pp. 19–38 (Chapter 2). DOI: 10.1007/978–3–7091–1425–4_2
［15］ L. J. Hoffman, et al., “Trust beyond security: an expanded trust model”, *Communications*

*of the ACM*, Vol. 49, No.7, pp.94–101 (2006).
［16］ Y. Murayama, et. al., “Trust Issues in Disaster Communication”, *Proceedings of the 46th Hawaiian International Conference on System Sciences (HICSS–46)*, pp.335–342 (2013).
［17］ Y. Tanaka, Y. Sakamoto, and H. Honda, “The Impact of Posting URLs in Disaster–related Tweets on Rumor Spreading Behavior”, *Proceedings of the 47th Hawaii International Conference on System Sciences (HICCS–47)*, pp. 520–529 (2014).
［18］ 大竹清敬・他,「対災害SNS情報分析システムDISAANA」,『NICT NEWS』No. 452,pp. 8–9 (2015).
［19］ M. R. Endsley, “Toward a Theory of Situation Awareness in Dynamic Systems”, *Human Factors Journal*, Vol. 37, No. 1, pp. 32–64 (1995).
［20］ J. Harrald, T. Jefferson, “Shared Situational Awareness in Emergency Management Mitigation and Response”, *Proceedings of HICSS–40* (IEEE, Waikoloa, HI, USA), pp. 23–23 (2007). https://ieeexplore.ieee.org/document/4076416?reload=true&arnumber=4076416
［21］ T. Kanno, K. Furuta and Y. Kitahara, “A Model of Team Cognition based on Mutual Beliefs”, *Theoretical Issues in Ergonomics Science*, Vol. 14, No. 1, pp. 38–52 (2013). DOI:10.1080/1464536X.2011.573010
［22］ 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).
［23］ 吉川・白戸・藤井・竹村,「技術的安全と社会的安心」,『社会技術研究論文集』, Vol. 1, pp. 1–8 (2003).
［24］ M. Deutsh, “The effect of motivational orientation upon trust and suspicion”, *Human Relation*,No. 13,pp. 123–139 (1960).
［25］ M. Deutsh, *The resolution of conflict* (Yale University Press, 1973).
［26］ D. Gambetta, “Can we trust trust?”, *Trust: Making and Breaking Cooperative Relations* (electronic edition, Department of Sociology, University of Oxford), pp. 213–237 (chapter 13). Originally published from Basil Blackwell (1988), available online from the following site: http://www.sociology.ox.ac.uk/papers/ gambetta213–237.pdf
［27］ Paul Slovic, “Perceived risk, trust, and democracy”, *Risk Analysis*, Vol. 13, Issue 6, pp. 675–682 (1993). DOI: 10.1111/j.1539–6924.1993.tb01329.x
［28］ T. C. Earle and G. Cvetkovich, *Social trust: Toward a cosmopolitan society* (Westport, CT: Praeger Press, 1995).
［29］ 中谷内一也,『安全．でも, 安心できない—信頼をめぐる心理学』(筑摩書房, 2008).
［30］ J. D. Lewis, A. Weigert, “Trust as a Social Reality”, *Social Forces* (Oxford University Press), Vol. 63, No. 4, pp. 967–985 (1985). DOI: 10.2307/2578601
［31］ 日景奈津子・カールハウザー・村山優子,「情報セキュリティ技術に対する安心感構造に関する統計的検討」,『情報処理学会論文誌』, Vol. 48, No. 9, pp.3193–3203 (2007).
［32］ 西岡大・村山優子,「オンラインショッピング時の安心感における情報セキュリティ技術に関する安全とユーザ属性との関係」,『情報処理学会論文誌』, Vol. 55, No.9, pp. 2168–2176 (2014).
［33］ Richard E. Petty, John T. Cacioppo, *Attitudes and persuasion: Classic and contemporary approaches* (William C. Brown, 1981).
［34］ Pradip Lamsal, “Understanding Trust and Security”, Technical Report, Department of Computer Science, University of Helsinki, Finland (2001). http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=B347CB60E1EADBBB2E59A9A9EAB4C7A8?d

oi=10.1.1.17.7843
［35］Cheskin Research and Studio Archetype/Sapient, “eCommerce Trust Study”, Joint project (1999). https://zdocs.pub/doc/17report-ecomm-trust1999-4gpdl9onxe17
［36］Y. Murayama, N. Hikage, C. Hauser, B. Chakraborty and N. Segawa, “An Anshin Model for the Evaluation of the Sense of Security”, *Proceedings of the 39th Hawaii International Conference on System Science (HICSS 2006)*. DOI: 10.1109/HICSS.2006.46
［37］山本太郎・他,「インターネットにおける不安発生のモデル化とその検証について」,『コンピュータセキュリティシンポジウム2009（CSS2009）論文集』, pp. 1-6（2009）.

**2.8節**

［1］ 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).
［2］ NIST, “Zero Trust Architecture”, SP 800-207. https://csrc.nist.gov/publications/detail/sp/800-207/final
［3］ “Remote Attestation Procedures Architecture”. https://datatracker.ietf.org/doc/draft-ietf-rats-architecture/

**2.11節**

［1］ 崎村夏彦,『デジタルアイデンティティー：経営者が知らないサイバービジネスの核心』(日経BP, 2021).
［2］ European Commission HLEG (High-Level Expert Group on Artificial Intelligence), “Ethics guidelines for trustworthy AI” (2018). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
［3］ European Commission, “White Paper on Artificial Intelligence – a European approach to excellence and trust” (2020). https://digital-strategy.ec.europa.eu/en/consultations/white-paper-artificial-intelligence-european-approach-excellence-and-trust
［4］ European Commission, “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). https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206
［5］ European Commission HLEG, “Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment” (2020). https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
［6］ AIプロダクト品質保証コンソーシアム（QA4AIコンソーシアム）,「AIプロダクト品質保証ガイドライン」(初版2019年、最新版2020年). http://www.qa4ai.jp/download/

**2.12節**

［1］ Fox-Brewster, Tom. "Londoners give up eldest children in public Wi-Fi security horror show." The Guardian (29 Sep 2014). https://www.theguardian.com/technology/2014/sep/29/londoners-wi-fi-security-herod-clause
［2］ 工藤郁子,「にじいろの議　人々の『眠り』と『目覚め』 社会の信頼、再構築を」, 朝日新聞夕刊（2020年8月12日). https://www.asahi.com/articles/DA3S14584996.html
［3］ セオドア・M・ポーター（著）, 藤垣裕子（訳）,『数値と客観性－科学と社会における信頼の獲得』(み

すず書房, 2013).
［4］ ジェリー・Z・ミュラー（著）, 松本裕（訳）,『測りすぎ　なぜパフォーマンス評価は失敗するのか?』（みすず書房, 2019).
［5］ リチャード・ホーフスタッター（著）, 田村哲夫（訳）,『アメリカの反知性主義』,（みすず書房, 2003).
［6］ ロレイン・ダストン, ピーター・ギャリソン（著）, 瀬戸口明久・岡澤康浩・坂本邦暢・有賀暢迪（訳）,『客観性』(名古屋大学出版会, 2021).
［7］ 春山習,「主権と統治（1）」,『早稲田法学』, Vol. 94, No. 1, pp. 74–77（2018).
［8］ Bruce A. Ackerman, *The Civil Rights Revolution (We the People, Volume 3)*（Harvard University Press, 2014).
［9］ 山本龍彦,「「睡眠」の質と憲法：「国民主権」から基本法制定を考える」,『中央公論』, Vol. 132, No. 5, pp. 66–69（2018).
［10］"'Defeatism about Japan is now defeated': Read Abe's Davos speech in full" (2019). https://www.weforum.org/agenda/2019/01/abe-speech-transcript/
［11］"G7 Roadmap for Cooperation on Data Free Flow with Trust" (2021). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/986160/Annex_2__Roadmap_for_cooperation_on_Data_Free_Flow_with_Trust.pdf
［12］「WTO 電子商取引共同声明イニシアティブ：オーストラリア、日本及びシンガポールの閣僚による声明」(2021). https://www.mofa.go.jp/mofaj/files/100272386.pdf
［13］アン・ブレア（著）, 住本規子・廣田篤彦・正岡和恵（訳）,『情報爆発－初期近代ヨーロッパの情報管理術』(中央公論社, 2018).
［14］アルフレッド・W・クロスビー（著）, 小沢千重子（訳）,『数量化革命－ヨーロッパ覇権をもたらした世界観の誕生』(紀伊国屋書店, 2003).

**2.14 節**

［1］ 田中朋弘,「職業の倫理―専門職倫理に関する基礎的考察」.
［2］ 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).

**2.15 節**

［1］ Onur Asan, et al., "Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians", *Journal of Medical Internet Research*, Vol. 22, Issue 6 (2020). DOI: 10.2196/15154
［2］ W. Jagodzinski, et al., "General trust in a changing society: The development of interpersonal trust between 1978 and 2013 in Japan", *Bulletin of Data Analysis of Japanese Classification Society*, Vol. 8, Issue 1, pp. 25–46 (2019). DOI: 10.32146/bdajcs.8.25
［3］ Francis Fukuyama, *Trust: The Social virtues and the creation of prosperity*（New York: Free Press Paperbacks, 1995).
［4］ Toshio Yamagishi and Midori Yamagishi, "Trust and commitment in the United States and Japan", *Motivation and Emotion*, Vol. 18, No. 2, pp. 129–166 (1994).
［5］ A. Kerasidou, "Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare", *Bulletin of the World Health Organization*, Vol. 98, No. 4, pp. 245–250 (2020). DOI: 10.2471/BLT.19.237198.
［6］ F. Gille, A. Jobin and M. Ienca, "What we talk about when we talk about trust: Theory of trust for AI in healthcare", *Intelligence–Based Medicine* 1–2, 100001 (2020). DOI: 10.1016/j.ibmed.2020.100001

［7］ Naomi Aoki, “The importance of the assurance that “humans are still in the decision loop” for public trust in artificial intelligence: Evidence from an online experiment”, *Computers in Human Behavior*, Vol. 114, 106572（2021）. DOI: 10.1016/j.chb.2020.106572
［8］ John M. Johnson and Andrew Melnikov, *The wisdom of distrust: reflections on Ukrainian society and sociology*（Emerald Group Publishing Limited, 2009）.
［9］ Ruha Benjamin, *People’s Science: Bodies and Rights on the Stem Cell Frontier*（Stanford University Press, 2013）.

# 付録

## 付録1 俯瞰セミナーシリーズ開催概要

開催日程：2021年7月29日（木）～ 9月1日（水）全15回（表4–1の通り）
開催形態：オンライン（Zoomミーティングを使用）

**表4–1 セミナーシリーズの講師と日程**

| 回 | 開催日時 | 講演トピック | 講師 |
|---|---|---|---|
| 1 | 7月29日（木）<br>15：00～16：30 | ゼロトラストから考える<br>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）<br>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）<br>13：00～14：30 | 人文・社会系の<br>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）<br>15：30～17：00 | ソフトウェア品質保証における<br>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）<br>9：30～11：00 | ヒューマンエージェント<br>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）<br>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）<br>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）<br>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）<br>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）<br>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）<br>9：30～11：00 | 行動経済学・実験経済学と<br>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）<br>15：00～16：30 | ソーシャルメディアにおける<br>トラスト問題 | 山口 真一（国際大学GLOCOM 准教授） |
| 13 | 8月27日（金）<br>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）<br>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）<br>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）

## 付録2 俯瞰ワークショップ開催概要

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］
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］
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）

# 付録3 協力いただいた有識者の方々

セミナーシリーズ講師以外に、以下の方々にもインタビューや議論をさせていただいた。
（50音順、敬称略）

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）
野崎 和久　（同 シニアプロデューサー）他

**セミナー／ワークショップ企画・報告書編纂メンバー**

| | | | |
|---|---|---|---|
| 総括責任者 | 木村 康則 | 上席フェロー | （システム・情報科学技術ユニット） |
| チームリーダー | 福島 俊一 | フェロー | （システム・情報科学技術ユニット） |
| チームメンバー | 井上 眞梨 | フェロー | （システム・情報科学技術ユニット） |
| | 加納 寛之 | フェロー | （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 企画運営室企画グループ） |
| | 住田 朋久 | フェロー | （企画運営室） |
| | 高島 洋典 | フェロー | （システム・情報科学技術ユニット） |
| | 戸田 智美 | フェロー | （ライフサイエンス・臨床医学ユニット） |
| | 花田 文子 | フェロー | （企画運営室） |
| | 福井 章人 | フェロー | （システム・情報科学技術ユニット） |
| | 的場 正憲 | フェロー | （システム・情報科学技術ユニット） |
| | 宮薗 侑也 | フェロー | （ライフサイエンス・臨床医学ユニット） |
| | 茂木 強 | フェロー | （システム・情報科学技術ユニット） |
| | 山本 里枝子 | フェロー | （企画運営室） |
| | 若山 正人 | 上席フェロー | （システム・情報科学技術ユニット） |

**俯瞰セミナー＆ワークショップ報告書** **CRDS-FY2021-WR-05**

# 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～

令和 4 年 2 月　February 2021
ISBN 978-4-88890-771-2

**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/

ISBN 978-4-88890-771-2

FOR THE FUTURE OF
SCIENCE AND
SOCIETY


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