総務省 学術雑誌『情報通信政策研究』 第６巻第１号
Journal of Information and Communications Policy Vol.6 No.1

特別寄稿

AI エージェント、サイバネティック・アバター、自然人の間のトラスト

中川裕志 [1]（理化学研究所・革新知能統合研究センター）

要　旨

社会における AI 技術の重要な応用分野として、インターネットにおいて自然人の代理的振舞いをする AI エージェント（AI Agent）あるいはサイバネティック・アバター（CA）が重要視され始めている。これらは、メタバースの構成要素として位置付けることができる。自然人本人と AI Agent や CA との間のトラスト、AI Agent や CA とインターネットを介してやり取りが行われる他の自然人やソフトウェアとのトラストは、これらの間のインタラクションが円滑に行われるために必要である。本報告では、このトラストを法的位置づけの観点、技術的実現性の観点から考察する。トラストの研究は多岐にわたるので、概要および本論文の内容に係る研究を取り上げて紹介する。次に CA の法的位置づけについての研究を紹介する。そのうえでトラストの基礎になる ID 認証について概念を説明する。次に、(1)自然人の本人と、(2)その代理者となる AI Agent や CA、さらに、(3)それらにインターネットを介してインタラクトする自然人ないし AI Agent や CA を想定し、(1)(2)(3)の間で成り立つべき ID 認証、トラストの枠組み、および個々の ID 認証、トラストに関わって生じる問題点について述べる。とくに、マルウェアや BOT による乗っ取り、なりすましとその対応策について言及する。次に、1 人が複数の AI Agent や CA を操る場合、および複数人が 1 個の AI Agent や CA を操る場合の問題点について説明する。

キーワード：AI エージェント、トラスト、サイバネティック・アバター、AI

１．はじめに

人間の代理として働く AI をここでは AI エージェント(以下では AI Agent と略記する)と呼ぶ。AI Agent が代理している人間から独立し、かつ自律的に活動できる度合いにはかなり幅がある。AI エージェントには種々の具体的実現形態があるが、例えば、サイバネティック・アバター(Cybernetic Avatar、以下では CA と略記する)(石黒, 2021)もその一つと考えられる。CA はインターネット上、あるいはメタバース上に本人とは異なる外見をもつ存在として使われることが現在では多い。しかし、将来は自律化が進み、本人とはまったく別の場所で自律して本人の代理として行動する存在となる可能性もありえる。

小塚（2022）は CA がメタバース内だけで行動する場合にはメタバース固有のルールに従えばよいが、現実世界と関わりがある場合は現実世界のルールが優先すべきだという主張を展開している。本稿では AI Agent や CA はインターネットを介して現実世界とインタラクトし、現実世界のモノを購入したり、現実世界の人間や組織と契約することを想定して

[1] 理化学研究所・革新知能統合研究センター・チームリーダー

いる。したがって、小塚の主張するように、以下では現実世界にルールや法制度を前提として議論を進める。哲学的に考えれば、AI Agent や CA は、抽象的にはラトゥール(Latour,2019)が提案するアクターネットワーク理論における actant に対応する概念と考えることもできる。

本報告のもう一つのテーマであるトラストは多数の意味で使われる。トラストを主題にした研究は文系、理系の双方において多数存在し、すべてを網羅することはできない。諸分野におけるトラストを捉え直し、AI とトラストの関係を含めて総括する研究が行われた(木村, 2021)。この成果を踏まえて、２章では人間と AI agent や CA の間のトラストの概念を概観する。３章では AI agent や CA の法的位置づけについて俯瞰する。４章では人間と AI agent や CA の間にある認証およびトラストについて述べ、そこで問題となるマルウェアや BOT による乗っ取りやなりすましの問題に触れる。５章では１人が複数の AI Agent や CA を操る場合、および複数人が 1 個の AI Agent や CA を操る場合の問題点について説明する。6 章はまとめである。

２．トラスト研究の背景

２．１．トラスト研究の流れ

ホッブスに始まり、ルソー、カントを経て現代にいたる社会学におけるトラストの研究の流れがある。これは主として囚人のジレンマあるいは社会的ジレンマを題材にしたトラストの研究である。囚人のジレンマとは、例えば、複数の囚人が協力して脱獄する穴を掘って成功すれば全員脱獄できるのに、一人の囚人がこの計画を看守に密告するとその囚人だけが釈放される場合は、結局は互いに協力して脱獄する穴を掘らなくなってしまうという現象である。囚人同士が、相手をトラストしていればこのジレンマを解決できる。似たような社会状況は多く存在するので、トラストは社会を円滑に動かす役割がある。これを社会秩序、社会契約、社会心理学、紛争解決などへと発展させていった流れを小山(2018)がまとめている。

Botsman(2018)は具体的なトラストの在り様を、企業が個人のトラストを得るには、企業の行動が個人にとって十分に予測可能であることが重要だと種々の例をあげて主張した。例えば、ブラブラカーは、運転車が現在地から、これから向かう到着場所を売り出して乗客を募る。このサービスは当初キャンセルが多くて成功しなかったが、オンライン決済の前払いを導入するとキャンセルが激減してビジネスとして成功した。つまり、前払い決済により、未知の顧客をトラストできるようになったことで成功した。新しいサービスは自分が良く知っているサービスに類似していることも予測可能性を高めるのでトラストを生み出す。たとえば、エアビーアンドビーは、ロンドン在住の人がニューヨークに旅行するときの宿泊先を紹介するために、ニューヨークの宿泊先がロンドンのどのような宿泊先に似ているかを提示して、成功した。これは既知の情報を与えて、未知のサービスを予測可能にしたことになる。いずれも、トラストがビジネスの牽引力になっている。AI がこれらのような信頼性や予測可能性という意味でのトラストを持つことができれば、人々が AI に対して抱く感情が改善するだろう。

山田(2021)はエージェントと人間のインタラクションを研究している。インタラクションは人間と擬人化エージェント間、人間とロボット間、エージェントを介した人間と人間の間

の 3 種に分類し、この 3 種のインタラクションがうまく行くためには、(1)エージェントのデザイン、(2)エージェント間で授受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 つ要素が重要であるとしている。そのための評価方法として、AI の性能に対する人間による主観的期待値と、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を AI の信頼性と定義している。このように AI の信頼性には人間が持つ主観が入っているので、トラストと呼んでもよいであろう。

エージェントの一種である AI Agent に対する人間からみたトラストを、AI とのインタラクションを通じて醸成していく方法もまた山田(2021)によって提案されている。(山田, 2021)は、山岸の言説による他者への信頼(山岸, 1999)を AI への信頼と置き換えた構造であると主張している。Anderson(2021)は、さらに人間とマシンが共存する世界、例えば、自動運転車と人間が運転する車が共存する世界においては、自動運転車が乗っている人間からトラストされるために自動運転車を悪意の攻撃から守る設計が重要であることを主張している。

トラストの強さの測定方法についての研究もされている(Bauer, 2018)。トラストを個人的な感覚の表現で測定することだけでは、大規模な場合への適用には客観性を欠くと考えられる。トラストする/しないの 2 項対立だけではなく連続値、具体的には A さんが「B さんが X を Y という状況で行う」と予測する確率とし、これを種々の状況 Y で評価し、場合によっては平均化する方法が提案されている。

ロボットや AI を見たとき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い。このネガティブなバイアスは、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、主観的な評価であるトラストは過小に評価されてしまうような状況である。これには、個々人のパーソナリティーが影響している。Stein(2020)は、複雑かつ高い能力、人間に近い表示をもっている AI に対する気持ち悪さとトラストの関係を西欧において調査している。人間に近い表示形態の場合、不気味の谷という気持ち悪さが強い。一方、表示形態は精巧でなくても、高い能力を示す AI は、人間がその能力を知り始めると、その存在を見直す傾向が表れる。ただし、人間と AI の友だち関係までは進展しない。この点は、AI との友だち関係や親近感を持ちやすい日本人のメンタリティとの差があるとしている。

２．２．トラストと認証

セキュリティの文脈でのトラストを(松本, 2021)に沿って紹介する。従来のトラストの実現方法は外界の攻撃からシステムを守る防火壁を強化する方向であった。しかし、最近の攻撃は巧妙化し、防火壁では守り切れないことが分かってきた。そこで、現れたのが防火壁を信頼しないこと、すなわちゼロトラストという考え方である。具体的には、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常に検証する作業を実行する。このときには外部の信頼できる認証局を使った認証プロセスを用いる。ネットワーク経由でつながっている相手が期待通りの検証の仕組みを持っているかをチェックするのがリモート・アテステーションという技術であり、電子署名が使われる。このようなトラストでは人間は関与せず、マシン対マシンのトラストである。

ここでトラスト対象が人間、マシンのいずれでも可能であるような拡張を行うために用語の定義をする。

エンティティ：対象になる人間とソフトやAIのことを合わせてエンティティと呼ぶ。

本人と本体：あるエンティティXが呼称Yで参照されたとき、Xが人間の場合、呼称Yの本人、ソフトやAIの場合はYの本体と呼ぶ。

アイデンティ:エンティティX本人あるいは本体を同定できる情報をアイデンティ(ID)と呼ぶ。

ID認証：呼称YからエンティティXのIDを確認することをID認証と呼ぶ。

IdP:あるエンティティからの別のエンティティのID認証を要求されたときID認証を行うシステムをIdP(アイデンティティプロバイダー)と呼ぶ。IdPはエンティティXに関するID認証ができればXの認証情報を発行する。

エンティティAが呼称bのエンティティBをトラストするためにはセキュリティの観点からはIdPによるID認証によってbに対応するBである本人ないし本体を認証することである。

セキュリティ以外の観点から見ると、トラストはID認証にとどまらなくなる可能性がある。この枠組は(Olivier, 2008, 2009)で詳しく取り上げられている。(Olivier, 2008)では、仮名や匿名の相手をトラストすることが可能であるとしている。仮名とはBにつけた仮の名前であり、呼称bの一種と考えられる。ID認証には仮名とBの対応表あるいは対応付けの方法が必要になる。

匿名のエンティティに対するトラストは、実現がより複雑になる。すなわち呼称がない、あるいは呼称があっても対応表がない状態である。よって、当該エンティティの行動履歴、所属組織、そのエンティティが他のエンティティとつながるネットワーク構造など、その他の情報を収集して匿名のエンティティのIDを認証する必要がある。技術的にはゼロ知識証明や自己主権型アイデンティ(Self Sovereign Identity：SSI)、などを使うことも考えられる。ID認証を要求したエンティティは通常1回限りの認証情報を見るだけである。よって要求元のエンティティはトラストしたい相手のプライバシーを攻撃する手段を持たない。しかし、認証システムがあるエンティティに関する認証を行う場合、多くのエンティティからいろいろな強さの認証を要求されることがある。もし、仮名や匿名のエンティティに対する認証を行うとすれば、認証のたびにそのエンティティに関する多様な情報を収集することがあるだろう。したがって、認証システム自体がトラストされることは必須である。

３．AI Agent

３．１．法的位置づけ

本報告はAI AgentやCAと自然人の間のトラストを検討することが目的なので、まずAI Agentの法的位置づけを歴史的に俯瞰する。

(Olivier, 2008)によれば、エンティティが自然人でないという事実それ自体は、法的権利の帰属を妨げるものではない。しかし、どのような条件下で法的主体性を帰属させることが意味を持つのか、また、どの程度まで法的主体性を認めるべきかが問題であるとしている。

Solum(1992, 2019)は、AI Agentが法人格を持つ条件を、複雑な行為を行う能力および意図的に（自己）意識を持って行動する能力という観点から定義している。これは、人格とは

意図的に行動する能力を意味するという伝統的な考え方を AI に沿うように法人格という形で拡大したものである。AI は自然人ではないので、自然人にのみ付与される人格は持てないが、法的な格を持てることはトラストの議論において重要な点である。Pagallo(2013)は単純ツール AI、完全な自律的 AI[2]の中間に限定的自律 AI というカテゴリーを導入した。ここで対象とする AI Agent は限定的自律 AI に対応する。Pagallo は AI Agent に目的に応じた法人格などを与えることを否定はしていないが、それが問題を全て解決するとは見ていない。むしろ、AI やロボットに人間に対して、ある行動を促進あるいは抑制するようなデザインを組み込むことを提言している。しかし、このような人間の行動を予測し制御するようなデザインは個別性が高く、一般化は困難である。むしろ、一般人に理解可能な明確さと透明性を持つという条件の下で AI Agent の利用目的に対応して部分的な法人格を与えることによって、人間と AI が共存する見通しのよい社会が作れるのではないかと筆者は考える。

石井(2021)によれば EU は「AI または AI の搭載されたロボットに法人格を与えるべきか」をいう問題設定に対して、人格付与を否定している。これに対して、新保(2021)は AI Agent の法的位置づけを検討している。新保によれば、米国の統一電子取引法（UETA）14 条は、「個人が電子エージェントの行為、または、結果としてもたらされた条項および合意に気付かなかった、または、確認しなかったとしても、当事者の電子エージェントの相互作用によって契約は成立し得る」という例を挙げて契約における AI Agent の動作結果を人間の場合と同様に認める方向を示唆している。

３．２．AI Agent と CA の関係

次に AI Agent と CA の関係について述べる。筆者は CA と本論文で定義している AI Agent は概念的には多くの部分で一致していると考えている。新保(2021)は CA を本人とは独立した本体としている。新保は特にロボットのような有体物の CA と AI ソフトのような無体物の CA を区別し、主に有体物 CA の法的位置づけについて議論している(新保, 2021)。石井(2021)は、CA を「身代わりとしてのロボットや 3D 映像等を示すアバターに加えて、人の身体的能力、認知能力及び知覚能力を拡張する ICT 技術やロボット技術を含む概念」という(内閣府, 2021)の定義を引用しており、本人との一体感が強い場合を考察している。さらに CA は、①操作者本人の情報、②その能力、③その外見や性格などを総合した主体であり、本人に代わって、リアル・バーチャルの世界で活動する存在と定義している。この場合は本稿で想定している AI Agent に近い概念となる。

４．AI Agent と CA のトラスト

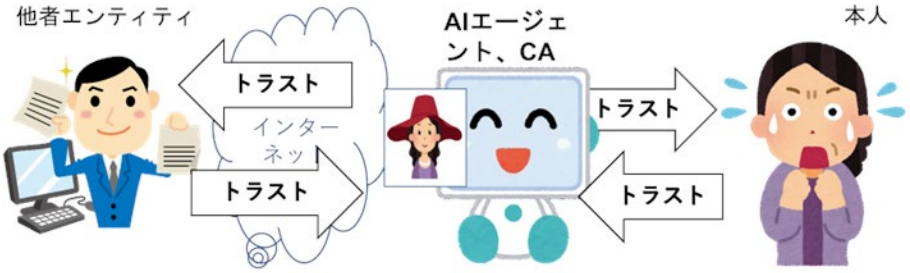
従来は、インターネットを介して接触するエンティティ同士がトラストし合うという 2 方向のトラストを考えておけばよかった。一方のエンティティが自然人である場合は、自然人はサービスの申し込み、商品の購入契約などで多大な労力をかけないと失敗することがありえる。したがって、なんらかの支援手段が有力だが、AI Agent や自律性を持つ CA はこの支援手段となる。また、エンティティ同士の会話やコミュニティ活動では CA が活躍する

[2] これはいわゆる汎用 AI あるいは Artificial General Intelligence と呼ばれる概念に対応する。

が、AI Agent の場合と同じく CA 本体と対応する自然人本人の間でも情報のやりとりをすることになる。

このような状況に鑑み、以下では、(1)自然人である本人と AI Agent や CA 本体の間の両方向のトラスト、および (2) AI Agent や CA とインターネットなどの情報環境を介してやり取りするエンティティの間の両方向のトラストの計４方向のトラストについて考える必要がある。ここで (2) のインターネットを介して対峙する相手側のエンティティとしては、プラットフォーム、サービスや商材を提供する事業者、自然人そして、自然人の代理をする AI Agent や CA などが想定される。以下ではこのエンティティを他者エンティティあるいは短く「他者」と呼ぶことにする。この４方向のトラストの関係、および CA の場合に考察すべき項目を図１に示す。

図１　他者エンティティ・本体・本人の間のトラスト



この図では本人と AI Agent ないし CA が１対１に対応する場合を示している。なお、本論文のこれ以降の図では、AI Agent、CA の箱の左側にある顔は CA の外見を表すとする。

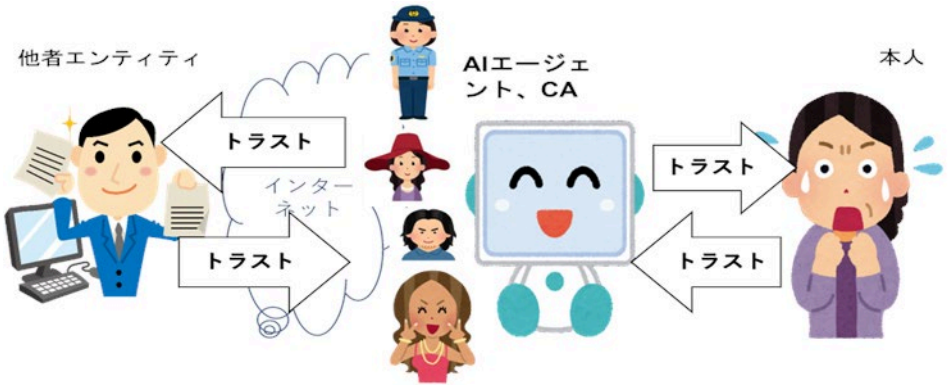
以下では、この４方向のトラストに関する問題点を説明する。

４．１．CA の外見

CA は本人の外見を本人以外のものにするという役割が重視されているので、ここでは外見の問題を扱う。

石井によれば、本人が外見を自分の意思で変えることは自己イメージコントロール権とされる(石井, 2021)。自分自身とは異なる外見の CA を使うことは、例えて言えば、お化粧をしたり仮面を被ったり、あるいはサングラスをかけることと類似しており、それ自体なんらの問題もない。つまり、図２の中央に示された他者から見える CA の４種類の外見のうち、下３つの一般人の外見なら、たとえ本人と大きく異なるものであっても、少なくとも法的には問題にならない。

図 2　CA の外見



しかし、一番上の警察官の外見にすると、他者エンティティは一般人に対する場合と異なる対応をしなければならないと考える、あるいはやや無理な要求にも答えようとするかもしれない。これは公職の地位を騙った詐欺行為になりうる。自己イメージコントロール権があると言っても、本人はその権利の利用には限界があることを意識する必要はある。

法的な問題はないかもしれないが、一番下の人の外見が他者エンティティの好みだったとすしよう。メタバースの中だけに他者エンティティと CA の接触が限定されていれば、問題はない。実際、高齢者の男性が若い女性の外見の CA を使う例はありえるし、彼が若い女性になりきるという非日常を楽しむという CA の利用法もある。ただし、他者エンティティが CA の外見を気に入って、なんらかの便宜を与えた後に本人の真の外見を知るようなことがあれば、騙された気分になるかもしれない。法的な問題ではないにしても、CA の外見は本人の外見とはまるで違うことは当初から認識しておくべきという論点はある。

４．２．他者と AI Agent、CA 間のトラスト

他者は現実には自然人をインターネットを介してトラストしていたので、それが AI Agent や CA になっても同じインターフェースでやり取りするなら、同様のトラストをすればよい。インターネットを介してトラストを構築するためには、通常は相手のエンティティを ID 認証する(崎村, 2021)。

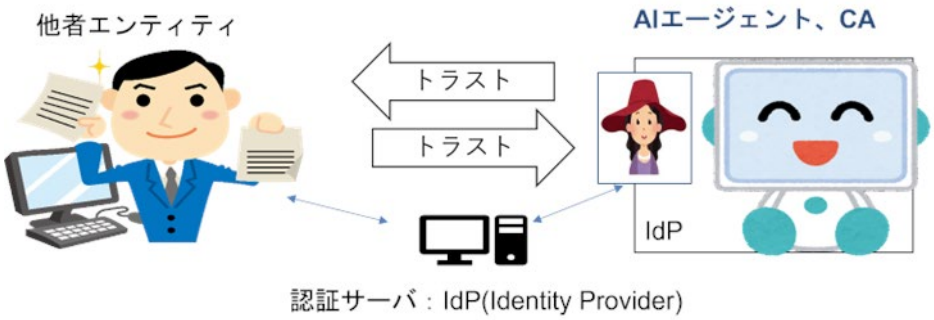
エンティティの認証は WebAuthn と FIDO2.0 という標準を組み合わせる方法が有名であり、FIDO2.0 は生体認証が使える標準になっている。エンティティが人間でないと生体認証は使えないので、AI Agent や CA の場合は電子署名などの手段で代用しなければならないので、そのような代用を認めるような標準にしなければならない[3]。この認証ではエンティティの認証の鍵となる認証情報を外部の IdP が管理しているので、AI Agent や CA も IdP に認証ための自分の認証情報を登録しておかなければならない。認証要求を受けたエンティティが IdP との間で認証に成功すると IdP は認証済 ID を生成し、認証要求元に OpenID Connect という標準化されたプロトコルを用いて認証済 ID を送る。このような流れでエンティティ同士の ID 認証ができる。ひとたび ID 認証が確立すると、その後の個別のアクセスは OAuth2.0 というプロトコルで行われる。なお、IdP を本人やその AI Agent

[3] 筆者の理解では、WebAuthn は生体認証であることを強制していないように見える。

の内側にもつ自己主権型アイデンティ(Self Sovereign Identity)という方法もあるがあまり普及していない。

上記の ID 認証は機械的なものであり、図３のように片方のエンティティが AI Agent ないし CA であったとしてもこの認証作業ができるソフトを組み込めばよい。ところが、トラストは ID 認証に基礎をおくとはいえ、それだけで実現するわけではない。

図３　他者エンティティと AI エージェント間のトラスト



被代理者情報：　AI Agent や CA は自然人である本人の代理であるなら、そのことも証明できなければならない。この証明は AI Agent や CA 本体の ID 認証情報に、AI Agent や CA が代理ないしは代弁している自然人本人に関する ID 認証情報を付加する必要がある。自然人本人が背後にいて、AI Agent や CA はその本人の代理であるという認識を他者が持っていれば、法的には民法 109 条による表見代理と考えられ、AI Agent や CA の行為の最終的責任は本人に帰属する。よって、AI Agent や CA が本人の代理という立場になると、本人の ID 認証情報は他者が AI Agent や CA をトラストするためには重要な情報である。

４．３．AI Agent、CA の法的自律性

AI AgentやCAに自律的な動作を認めると、民法99条1項が定める代理行為ではなく、民法100条が定めるようにその行為はそれら自身のための行為となるため、AI AgentやCAに何らかの法的な人格を想定しなければならない。

AI Agentへの人格や法人格付与は3.1節で述べたようにEUでは否定されているが、米国では契約の有効性は認められる。この状況とAI AgentやCAの予想される開発、利用の速度からみて、人格、法人格のような根本的な法的位置づけを目指すよりは、米国のような契約の有効性に持ち込んでAI AgentやCAの限定的自律性を契約行為とみなして実現を図るほうが現実的であろう。

そうなると、AI AgentやCAのトラストは契約を確実に履行する実績によって構築されることになる。これはID認証とは全く異なる技術であり、以下の要件を持たなければならない。

要件１：契約履歴を記憶しておき、要求に応じて正しく開示する。

要件２：開示要求側のエンティティが、開示された契約履行履歴を機械学習などのAI技術によってトラストできるかどうかを判定する。

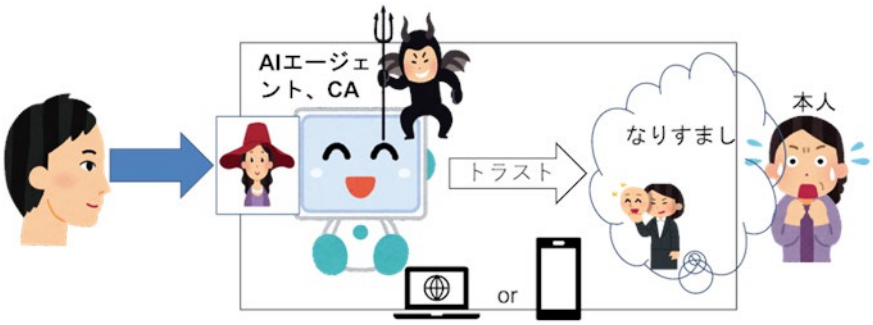
このように、自律性を持つAI AgentやCAのトラスト構築には高い知識処理技術が必要

になる。また、機械学習を使うなら、当然ながら100%確実ではなく、得られるのは契約が履行される予測確率なので、トラストする側のエンティティはその予測確率に基づく判断をしなくてはならない。

４．４．AI Agent、CA のなりすましと乗っ取りによる他者への危害

AI Agent や CA 本体に他のエンティティが騙されるケースとして、図４に示すような本体が対応する自然人の本人以外の人に<u>なりすまし</u>をされて悪用されるケースと、本体がマルウェアや BOT に<u>乗っ取られる</u>ケースがある。

図４　本人のなりすましと乗っ取り



本人以外の悪意のある人間が、本人になりすまして AI Agent や CA を操る場合に限って言えば、AI Agent や CA が本人を ID 認証する標準としてすでに述べた FIDO2.0 を使えば生体認証や多段階認証の利用によって高いレベルの安全性を持つ ID 認証が可能であるが、万全とは言い切れない。また、マルウェアや BOT は本人にそれなりの知識があって用心をしている場合でも、破られて、図４に示すように乗っ取られる場合もありえる。

石井(2021)は、事前の措置として市場に CA を上市する条件として、①なりすましや乗っ取りを防ぐ方法を充実して他者の権利利益を侵害しないようにすること、②自身の CA が他のエンティティの CA によるなりすましや乗っ取りによって起きる攻撃から防御できる堅牢性をもつことを保障する仕組みを提言している。これはリスク管理の考え方である。一方、事後の措置として、法的規制として③なりすまし行為に対する処罰規定、④なりすましを働いた者に対する CA の利用禁止、⑤不正利用のおそれのある CA の遠隔自動停止などを提言している。

以上の提言のうち、③、④は法的な決め事の問題であり、なりすましが発覚したという前提なら法的決め事として実現できる。ただし、なりすましが自然人によって引き起こされたなら、法的措置として有効であるだろうが、マルウェアや BOT による乗っ取りの場合は、その行為を働いたものを突き止めること自体が困難である。

①は主に法的決め事というものの、技術的観点からすれば、事前に全ての権利侵害のケースを洗い出せるか不明である。AI Agent や CA の既存の権利侵害事例を集積してデータベース化しておき、新たに上市しようとする AI Agent や CA がこのデータベースの事例に当てはまる動作をしないことをいちいち確認しなければならない。厳密に全ての場合を確認することは AI Agent や CA の開発コストを上昇させる。

②、⑤には技術的解決策が必要になる。なりすましたAI AgentやCAが別のエンティティに対して行ってきた行為が、なりすましているエンティティの行為なのか、正当な本人の行為なのかを判断しなければならない。もちろん、なりすましたAI AgentやCA自体はID認証をされているので、通常ならトラストすべきところである。

相手側のAI AgentやCAがなりすましかどうかを判定するには、以下のようなAI技術が必要になる。すなわち、AI AgentやCAがなりすまされていないとき行動と、なりすまされたときの行動のログを使って、なりすましかどうかを判定するAIを学習する。ただし、なりすまされた場合の行動は少数しか集められないので、うまく学習できるかどうか疑問が残る。

一般的には学習が難しいとなると、当該AI AgentやCAが過去に行ったやりとりの履歴があれば、AI技術によって、そのAI AgentやCAの行為を予測するAIシステムを構築することも考えらえる。予測される行為の範囲から逸脱していれば、なりすましたAI AgentやCAであることを疑う。ただし、AI AgentやCAが初見の場合には過去のやりとりデータがない。

これらの状況を改善する方法としては、AI AgentやCAの開発メーカあるいは管理企業が個別のAI AgentやCAの行為を常に収集し、個別のAI AgentやCAごとに行為の予測モデルを作って提供する方法、あるいは個別のAI AgentやCAの行動を監視して予測の範囲を超えるかどうかをチェックし続ける方法が考えられる。ただし、この方法には４つの問題がある。

(1) 全てのAI AgentやCAの行為を開発企業や管理企業が常時監視することは、機械的に行うにしても、大きな労力、通信コストが必要になるので、経済性に疑問がある。
(2) 個別のAI AgentやCAが代理をしている本人の個人情報保護に抵触する可能性がある。
(3) なりすましでない場合でも、予測される範囲の行為ではないからといって、本人でないとも言い切れない。なぜなら、自然人の本人はモノゴトの考え方や判断方針を変える可能性は常に存在する。
(4) 本人がこれまで行ったことがないタイプの作業だと予測が難しい。たとえば、はじめて不動産を買う場合、はじめて海外旅行に行く場合などがある。この場合は、多数の人々がこれらの「はじめて」の行為の場合にどう行動したかのログを集めて利用する方法がある。ただし、個人情報の保護には留意しなければならない。

エンティティとやり取りするに際して、守る側のエンティティが常にこのようなりすましかどうかの判断を行うことは、重い処理であり現実的かどうかは疑問が残る。

このような考察により、AI AgentやCAのなりすましに対しては③、④のような法的手段が経済性を考えれば有望であろう。ただし、既に述べたように、マルウェアなどによる乗っ取りへの有効性には疑問が残る。したがって、なりすましや乗っ取りの被害を受けた場合の法的ないし経済的救済策として保険制度を整備しておくことが重要である。具体的には、被害者への金銭的補償、本人への免責範囲を保険のポリシーとして設定することになるだろう。

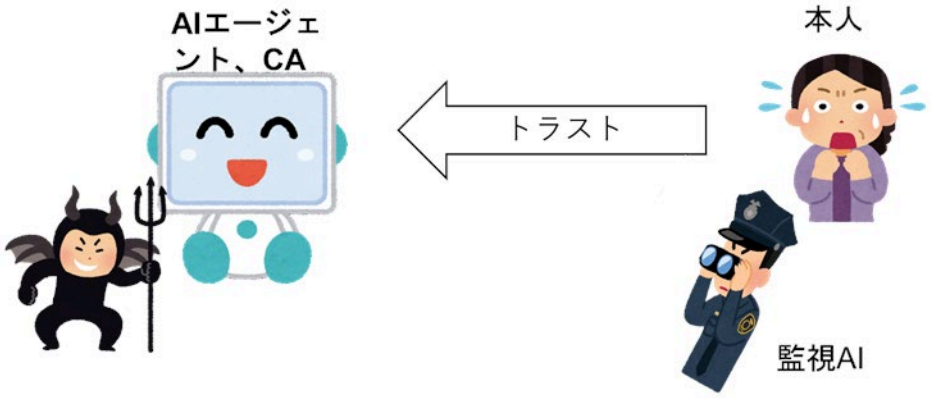
４．５．乗っ取られた AI Agent や CA の本人への危害

乗っ取られた AI Agent や CA が本人になりすまして他のエンティティを欺く以外に、本人を欺こうとする場合もある。

たとえば、本人が商品購入をするために AI Agent を使ったとき、悪徳な商品販売会社の商品を推薦したり、法外な金額で購入契約[4]をしてしまうことが考えられる。あるいは他のエンティティからのメッセージを改竄することもあるだろう。

AI Agent や CA がマルウェアや BOT に乗っ取られる可能性はコンピュータウイルスに感染するような確率で常につきまとう。ここで問題になるのは、自然人である生身の本人は AI Agent や CA をトラストする電子的な手段を持たないことである。そこで登場するのが図５に示す監視 AI として示した本人とは別の電子的エンティティがセキュリティ技術、AI 技術などを用いて AI Agent や CA を監視する方法である。

図５　乗っ取られた AI エージェント、CA の監視



AI Agent や CA を監視するソフトウェアが乗っ取られた PC やスマホに常駐するのでは、その監視ソフトが乗っ取られていることもありえるから、危険性は減らせない。

そこで考えられるのは、前節で述べたように AI Agent や CA の開発業者などが CA を監視する AI ソフトを動かし、個々の AI Agent や CA の挙動を遠隔から監視する方法である。BOT 対策の場合は、通常はアクセスされない外部の特殊なサイトとの通信を監視する手法が知られている。BOT が乗っ取った場合には、この対策は有効と考えられる。

このような仕組みを使う場合、自然人の本人がトラストする相手は本体の AI Agent や CA というよりは、それらを開発して、監視サービスまで提供している企業ということになる。これは、個人としてはトラストしやすい相手だろうし、企業も社会的評価を受けるので、悪事は働かないというインセンティブがあると考えられる。

ただし、既に述べたように、開発業者ないし監視を専業とする業者が自社が受け持つ全ての AI Agent や CA を監視するためには、業者と AI Agent や CA の間の通信コストは大きなものになり、かつ業者のサーバにおける監視ソフトの稼働も大きな負荷となることはやむを得ない。

---

[4] 霊感商法として問題になっている。

５．本人と本体が１対１以外の場合

ここまで自然人の本人１人に対して AI Agent や CA の１体が対応する場合を考察してきた。この節では、それ以外の場合について検討する。

５．１．１人がＮ体を使う場合

石黒(2021)によれば、そもそも CA は自然人１人の能力を拡張するためにあり、当然ながら１人が使う CA は複数すなわち N 体（N>1）になり、しかも N がかなり大きな数もあり得る。

１人 N 体で N が大きい場合は、１人の自然人は多数の AI Agent や CA を同時に意識し、即時に対応することは困難である。その結果、AI Agent や CA は自身の判断で活動する自律性を持たざるをえなくなる。自律性を持った場合についての法的問題点はすでに 3.2 節で述べた。N 体の AI Agent や CA がお互いに無関係な行為や契約をしている場合は、米国の限定自律性を持つ電子エージェントの契約の正統性に依拠する方法が有力である。だが、N 体が互いに関係する場合は、追加的な取り決めが必要である。

一連の行為シーケンスが N 体に分かれて行われる場合、その一連の行為が矛盾するものではないが、本体 Y が行為 y を行う前に本体 X が行為 x を行う必要がある場合は、X と Y はその順序性を互いに認識しておかなければならない。この順番を乱して行為しようすると、間違った結果になるかもしれない、ないし行為そのものができないということがありえる。N 体を管理監督する本人が、このような処理順序を指定できればよいが、自然人である本人には過大なタスクかもしれない。その場合は高い自律性をもつ本体が必須になる。

N 体の行為に矛盾が生じた場合はさらに厄介である。外部のエンティティ D に対して、1 体 A が契約 C を受諾し、別の 1 体 B が契約 C の受諾を拒否したとしよう。D が A、B ともに同一の本人の代理だと認識していたとすれば、表見代理とみなして A、B の契約に関する行為の時間順序に沿って本人が行動したと D は見なせる。その場合は契約に関する法令にしたがうことになるだろう。したがって、契約の受諾ないし拒否の行為の時間的順序を正確に記録しておくことが A,B,C にとって必須となる。

上記の問題は、N 体が自律しており、個別の法人格が与えられた場合は、本人は N 体に対する依頼者であり、責任は N 体、依頼者の本人、および相手側の契約当事者の間での法律に基づく紛争処理に委ねられると考えられる。

５．２．本体の集合が階層構造を持つ場合

ここまでは主に１本人 N 本体の場合で、N 本体が全て本人に直属するというフラットな構造を持つ場合を検討した。しかし、AI Agent や CA が自律性を持つ場合は、5.1 節で述べたように、自然人の本人には手に余る場合もある。

この問題を解決するためには、N 体の中に全体を統率する司令塔になる 1 体を置き、その 1 体の指令によって他の N-1 体が動くような階層的な命令系統を構成しなければならない。この問題は、並列分散処理のアルゴリズムとしてソフトウェアの世界では多くの蓄積がある。ここで扱う状況では、どの 1 体を司令塔にし、そして司令塔とその他の N-1 体との間の責任分担を決める必要がある。これは、契約の履行においてマルチステークホルダーの関係を明確化するという契約法務の知識が使える可能性がある。これらの技術的、法的な知

見を動員すれば原理的には解決可能な問題と考えられる。

1 司令塔本体、N-1 本体の構造は図 6 の左側に示す 2 層からなる階層構造だが、図 6 の右側に示すように、より複雑な階層構造も考えられる。

図 6　階層構造を持つ本体の集合

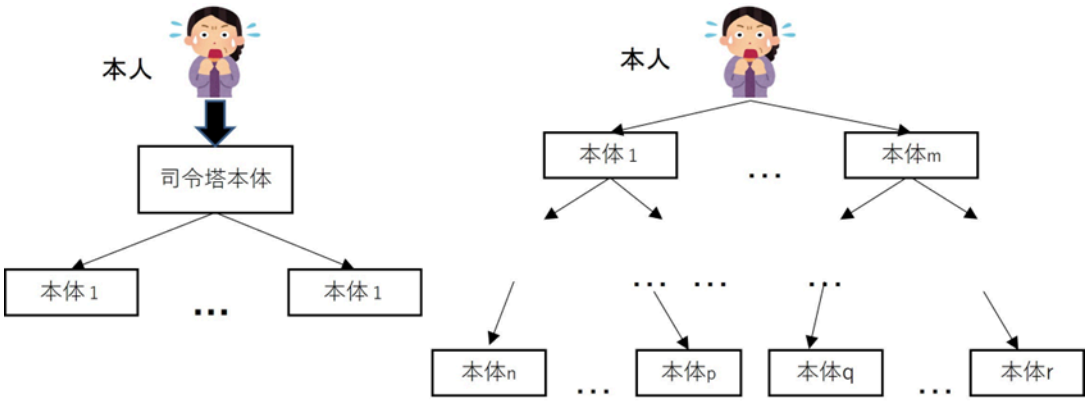
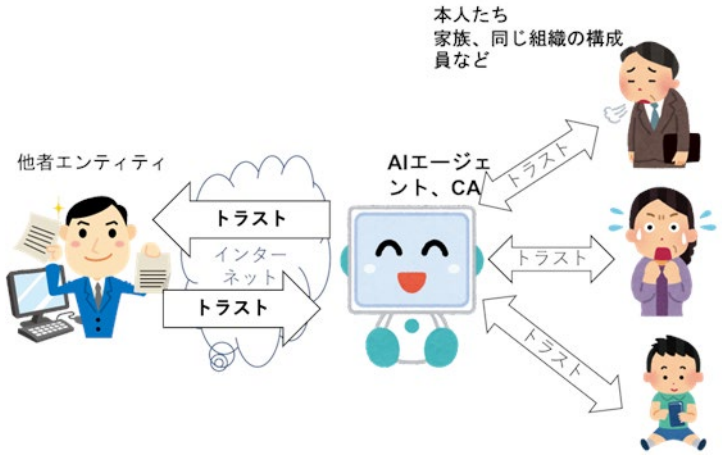


図 6 の右側の階層的な場合、本人は第 1 層の本体に働きかけ、ある層の本体は一つ上の階層の本体から指令を受けて、一つ下の層に本体に指令をだす。最下層の本体たちは、他のエンティティとやり取りする。また、最下層の本体とやり取りする他のエンティティは、本人が誰かを知りたいかもしれない。その場合には、本人から最下層の本体に至るルートの情報が必要である。なぜなら、このルート上の各本体が自律的に行為をできる範囲や行為の責任の情報が記載されている。他のエンティティとの間で責任問題がおきたとき、その責任を負う本体を特定しなければならない。場合によっては責任が本人にまで及ぶ可能性もある。

５．３．N 人が 1 体を使う場合

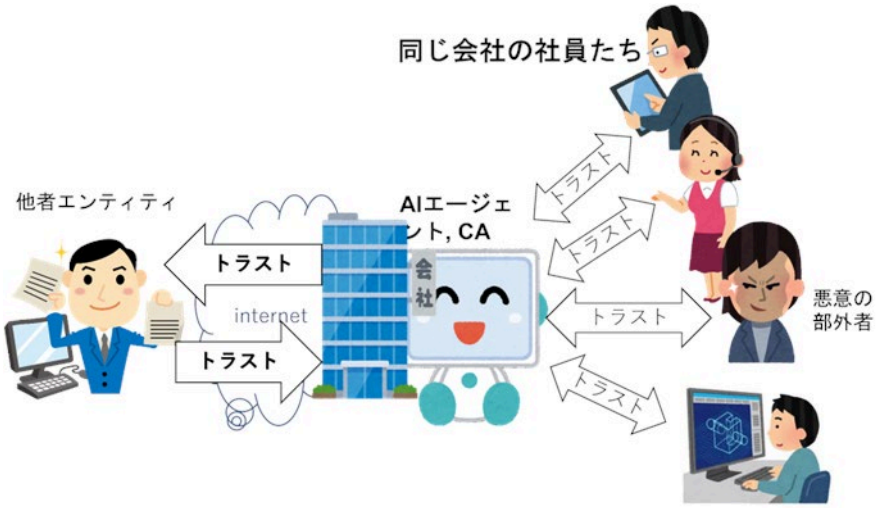
N 人が 1 体を使う典型的な場合は図 7 に示す家族のメンバーが家族を代表する 1 体の AI Agent や CA を使う場合であり、家族としての購買や契約などにおいては便利であろう。

図７　３人家族と１本体



会社などの組織は法人としては１個であっても、背後に多数の社員がいて、１個の法人を支える活動をしている。そこで図８に示すように会社を代表する１体の AI Agent や CA を複数の社員が使うことが考えられる。

図８　会社を代表する AI Agent や CA



例えば、ある製造業のセールスを行う AI Agent や CA が、売り込みや契約をタスクとする人、製品の技術的説明をする開発系技術者、法的問題に対応する法務部門の人が場合に応じて本人になって活動することはあり得るだろう。また、相手側からすれば、1 体の AI Agent や CA と交渉するだけで間に合うので便利である。

このような AI Agent や CA の使い方は相手側にはシンプルなインターフェースで便利かもしれないが、使い手の本人たちの間で十分な意思疎通や責任分担ができている必要がある。

AI Agent や CA に自律性がある場合は、相手側とのやり取りから、どの人を本人として指定するかを選択して決めなければならない。相手側とのやり取りの内容を分析してこの

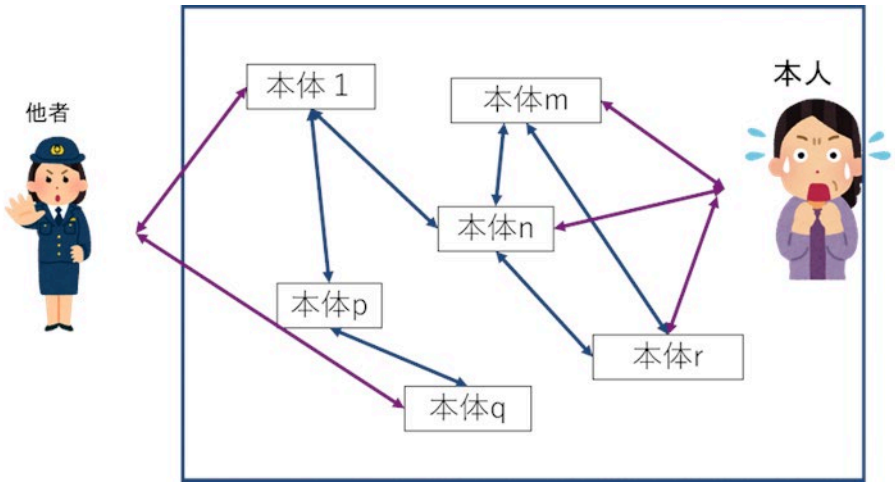
選択を行うには、高度な自然言語処理技術および推論、判断などの AI 技術が必要であるが、技術的には今後の課題であろう。

さて、図８の右側の上から 3 番目に悪意を持つ外部者が社員を装って侵入することもあり得る。多くの社員が 1 体の AI Agent や CA を使う場合は管理が甘くなって、このような侵入を許しやすくなるかもしれない。技術的な対応策は、4.4 節で述べたことの応用になる。つまり、AI Agent や CA は個々の社員の行動を予測し、その予測と違う行動をする社員がいたら、悪意の外部からの侵入者とみなして、認証を取り消すという処理を行う。このような機能を装備した AI Agent や CA は高額になるかもしれないが、会社として予測される損害を考慮して導入を考えるべきだろう。

５．４．アクターネットワーク

図６のような複雑な階層構造をもつ CA の集合体の全容は本人にはもう分からないと考えたほうがよいだろう。そこでは、本人の代理としての CA という関係ではなく、複数の CA 本体たちと自然人本人がネットワークでつながる図９のような関係のほうが現実的であろうと考えられる。これは、アクターネットワーク理論において、自然人がアクターであり、CA がアクタントになるという概念に近い。人間のアクターが CA のアクタントに働きかけ、CA はそれに応えるというインタラクションが行われる。このような形態が社会で一般化したとき、自律性を備えた CA あるいは AI Agent がどのような法的位置づけ、たとえば法人格、限定的人格、を持つべきかという Solum(1992, 2019)の問いかけが現実的に議論されるようになると予想される。

図９　自律した多数の CA と自然人が形成するネットワーク



６．まとめ

自然人本人と AI agent や CA との間のトラスト、AI agent や CA とインターネットを介してやり取りが行われる他の自然人やソフトウェアとのトラストについて、本体や本人などの諸概念の提案と定義を行った。次にトラストを上記の 3 者間で成立させる前段階としての ID 認証、次いでトラストの形成について法的位置づけの観点、技術的実現性の観点から考察した。本人のなりすましや、乗っ取られた本体が本人の意図とは違う行動をする場合

の対策などを検討した。本報告は全体像の概観と、問題点の指摘が中心であり、これを出発点として、より具体的な方策ないし技術的実現方法を考えることが今後の課題であろう。

謝辞

本研究は JST RISTEX「人と情報のエコシステム」研究開発領域:研究開発プロジェクト「PATH－AI:人間-AI エコシステムにおけるプライバシー、エージェンシー、トラストの文化を超えた実現方法」の補助を受けて行っている。

参考文献

Anderson R., Shumailov I. (2021) Situational Awareness and Adversarial Machine Learning – Robots, Manners, and Stress, https://www.cl.cam.ac.uk/~rja14/Papers/situational-awareness2021.pdf (last access 2022/3/7)

Bauer, P. C., Freitag M. (2018) Measuring trust". In *The Oxford Handbook of Social and Political Trust*, edited by Eric M. Uslaner, 15-36. Oxford University Press.

Bootsma R. (2018): TRUST 世界最先端の企業はいかに〈信頼〉を攻略したか. (関 美和訳).日経 BP 社.

Latour B. (2019) ブリュノ ラトゥール(伊藤嘉高訳). 社会的なものを組み直す: アクターネットワーク理論入門. 法政大学出版局, 2019

Olivier D., Chiffelle J. (2008) WP2:D 2.13 Virtual Persons and Identities. Future of Identity in the Information Society.

Olivier D., Chiffelle J., Buitelaar H. (2009) WP17: D17.4: Trust and Identification in the Light of Virtual Persons. Future of Identity in the Information Society

Pagallo U. (2013) The Laws of Robots-Crimes, Contracts, and Torts, Springer

Solum L.B. (1992) Legal Personhood for Artificial Intelligences. 70 North Carolina Law Review. 1231.

Solum L.B. (2019) Artificially Intelligent Law. available at SSRN: http://dx.doi.org/10.2139/ssrn.3337696

Stein, J.-P., Appel, M., Jost, A., & Ohler, P. (2020) Matter over mind? How the acceptance of digital entities depends on their appearance, mental prowess, and the interaction between both. *International Journal of Human-Computer Studies*. Advance publication online. https://doi.org/10.1016/j.ijhcs.2020.102463

石井 夏生利 (2021) サイバネティック・アバターとプライバシー保護を巡る法的課題.人工知能 36(5), p.578-584.

石黒 浩 (2021) アバターによる仮想化実世界の倫理問題. 人工知能 36(5), p. 558-563.

小山虎（編著）(2018) 信頼を考える. 勁草書房.

小山虎 (2022) 人文・社会系のトラスト研究の系譜. 俯瞰セミナー＆ワークショップ報告書：トラスト研究の潮流 ～人文・社会科学から人工知能、医療まで～. CRDS-FY2021-WR-05. p.5-11. (https://www.jst.go.jp/crds/pdf/2021/WR/CRDS-FY2021-WR-05.pdf)

小塚 荘一郎 (2022) 仮想空間の法律問題に対する基本的な視点 ―現実世界との「抵触法」的アプローチ, 『情報通信政策研究』 第 6 巻第 1 号, p.IB-1 - IB-13.

木村 康則 他 (2021) 俯瞰セミナー＆ワークショップ報告書：トラスト研究の潮流 ～人文・社会科学から人工知能、医療まで～. CRDS-FY2021-WR-05. p.5-11. (https://www.jst.go.jp/crds/pdf/2021/WR/CRDS-FY2021-WR-05.pdf)

崎村夏彦 (2021) デジタルアイデンティティー.日経 BP

新保 史生 (2021) サイバネティック・アバターの存在証明 ―ロボット・ AI・サイバーフィジカル社会に向けたアバター法の幕開け ―. 人工知能 36(5), p. 570-577.

内閣府 (2021) 政策統括官（科学技術・イノベーション担当）付未来革新研究推進担当：ムーンショット型研究開発制度の概要

山田 誠二 (2021) ヒューマンエージェントインタラクションと信頼工学. 俯瞰セミナー＆ワークショップ報告書：トラスト研究の潮流 ～人文・社会科学から人工知能、医療まで～. CRDS-FY2021-WR-05. p.82-90. (https://www.jst.go.jp/crds/pdf/2021/WR/CRDS-FY2021-WR-05.pdf)

松本泰 (2021) 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～. 俯瞰セミナー＆ワークショップ報告書：トラスト研究の潮流 ～人文・社会科学から人工知能、医療まで～. CRDS-FY2021-WR-05. p.65-72. (https://www.jst.go.jp/crds/pdf/2021/WR/CRDS-FY2021-WR-05.pdf)

山岸俊男 (1999) 安心社会から信頼社会へ. 中央公論社.