ヒューマンエージェントインタラクションにおけるトリレンマ

Trilemma in Human Agent Interaction

松井哲也 [1*]　笹井一人 [1,2]
Tetsuya Matsui[1]　Kazuto Sasai[1,2]

[1] 大阪工業大学
[1] Osaka Institute of Technology
[2] 茨城大学
[2] Ibaraki University

Abstract: 本論文では，人間とロボットやバーチャルエージェントとのインタラクションを研究するヒューマンエージェントインタラクション (HAI) における 3 つの立場を，郡司の「人工知能・自然知能・天然知能」モデルを基に，既存の研究を位置づけていく方式で分類した．その際に着目したのは，各研究において「他者」をどのようにモデル化しているかということである．さらにそこに「自由意志・局所性・因果律」のトリレンマという概念を導入し，その観点から各立場がどのように記述できるかを検討した．その結果，現状の HAI 研究の見取り図を示すことができ，今後の HAI 研究の課題を示すことができた．

1　はじめに

この分野においては,「他者」を工学的にどのように扱うかという点が大きな課題となる．松井 [松井 22] では，郡司 [郡司 19] の「天然知能・人工知能・自然知能」という枠組みを援用して，既存の HAI 研究はその「他者」の扱い方によって,「天然知能的 HAI・人工知能的 HAI・自然知能的 HAI」の 3 つの立場に分けられることを示した．本論文においては，この松井の議論を拡張する形で，自由意志・局所性・因果律のトリレンマ [郡司 19][] の概念を取り入れ，HAI における 3 つの立場の差異をより明確にし，さらに松井 [松井 22] が触れていない HAI の先行研究を実際に挙げながら，それらが 3 つの立場のいずれに属するのかを検討することで，現時点での HAI 研究の見取り図を提示したい．

2　HAI 研究の 3 つの立場

HAI 研究は,「他者」をどのように捉えるかによって，3 つの立場に分類することが可能である [松井 2022][松井 22].

第一は，他者を，他者とインタラクションする「わたし」にとって完全にモデル化可能なものであると見なす立場である．このようなモデルを「他者モデル」と呼び [松井 23]，ロボットやバーチャルエージェントなどの AI システムにこのような「他者モデル」を実装することで，他者を理解できるエージェントを実現可能だとする立場である．この立場に立つ主要な研究者として，大澤正彦と坂本孝丈が挙げられる．大澤は複数の変数を導入した頑健な他者モデルを提案しており [大澤 20],「他者」を「わたし」の内部に投影することで，質の高いインタラクションが可能になるという志向を明確にしている．坂本は，ロボットと人の身体的な相互作用を観測する実験で得られたデータを基に，やはり多数の変数からなる頑健な他者モデルを提案して他者の振る舞いを予測可能なものとしようとしている [坂本 19][Sakamoto 21].

ここで，郡司 [郡司 19] の「天然知能・人工知能・自然知能」の概念を導入する．郡司は,「人工知能」を，世界を「私」を中心に置いたデータの集まりとして解釈し，それらのデータの中から自分にとって有用なものを，自分の基準で取捨選択する主体であると定義した．前述したような頑健な他者モデルを重視する立場は，他者を「私」にとって完全にモデル化可能なものであるとみなすという立場である．これは,「他者」という無限な存在 [Levinas 61] の構成要素の中から,「私」にとって有用で理解可能な要素のみを取捨選択するという立場であり，その意味でこのような HAI は「人工知能的 HAI」と呼べる [松井 22].

また郡司は，天然知能的思想には現象学の影響が強く見られることを指摘しているが [郡司 19]，哲学的な人工知能研究者とされる三宅 [三宅 16] はフッサール，メルロ・ポンティ，ユクスキュルといった現象学者・現象学的思想家を非常に重視しており，人工知能研究者と現象学の親和性の高さの傍証とすることができる．

*E-mail:tetsuya.matsui@oit.ac.jp

このような人工知能的 HAI は，他者を「私」の内部である，すなわち「私」の内部の論理で完全に記述可能なものであると見なす立場であると言える．第二の立場は，他者を「自分と同じルール・論理の中で動くもの」として定義する立場である．この立場では，他者を完全にモデル化する必要は無く，自分と同じルールに従うかどうかでその「他者性」を判定する．いわば，他者を「私」の延長と捉える立場である [松井 22]．

この立場の重要な研究者には寺田和憲がいる．寺田はロボットやエージェントの「社会性」というキーワードを好んで用い，ユーザがエージェントに社会性を感じることを重視する [寺田 14] 他，エージェントに「悪意」を感じるかどうかを探求した研究 [寺田 17] や「嘘をつく」という社会的な行動に焦点を当てた研究 [寺田] など，既存の社会的規範･ルールの内部での他者とのインタラクションという文脈を重視した研究を行っている．また，非言語情報などで感情や内部状態を伝達･受容する「社会的シグナル」に関する一連の研究 [植田 16][Feine 19][Cross 19] も，やはり他者と「私」が同じルールに従うことを前提とした立場であると言える．社会的シグナルとは,「ある特定の非言語的表出はある特定の情報を表している」という普遍的なルールに，社会の構成員の全てが従っているという考え方に基づいている．よって,「この表出はこの感情」といったリスト（「地図」）を所与のものとして前提しているという点で，自然知能的なものであると言える．熊崎博一は，ロボット・バーチャルエージェントを用いて発達障害児に対して面接の練習を行い，就職を支援するシステムを提唱している [Kumazaki 22]．これは，社会的ルールから逸脱しがちな人間を，社会的ルールの中に適合させることを目指すものであり，やはり他者を「私」の延長上に位置づけるものである．このような立場からの HAI 研究は，郡司 [郡司 19] の定義に合わせれば「自然知能的 HAI」と呼べるだろう．郡司は，自然知能とはあらかじめ世界の見取り図を持っており，それを基準として「私」にとって最善の選択ができる知能であると定義している．ここで，世界の見取り図に相当するものを「ルール・社会規範」だとすれば,「ルール・社会規範」があることをまずは前提として，それを基準として「私」と他者を定義・記述しようとする寺田らのアプローチは，自然知能的 HAI と呼ぶべきである．付言すれば，これは「他者」を実在物ではなく，社会的な構成概念とみなす立場であるとも言える．これは,「他者」を科学的な研究対象として分析・記述が可能であるとする人工知能的 HAI（ここでは他者を実在物と考えていることになる）と比較すれば，より言語論的回転の影響を強く受けていると言えるだろう．

最後に残った第三の立場が「天然知能的 HAI」である．これは，他者を，モデル化も規範の共有も不可能な「外部」[Levinas 61] であると定義する立場である [松井 22]．郡司 [郡司 19] は天然知能を，絶えず「外部」に開いた知能であると定義する．この外部とは，あらかじめ予測･知覚することはできないが，それが現に目の前に現れた場合にはそれとして対峙せざるを得ないような存在である．これを他者論の文脈で言えば，あらかじめ「やってくる」ことは予測できないし，どのような言動をするかも予測・理解できないが，否応なしに眼前に現れる他者である．思想的には現象学を重視し，他者モデルを作業仮説として重視する人工知能的 HAI や,「社会的な規範・社会性」という枠組みをあらかじめ設定する自然知能的 HAI と比較して，天然知能的 HAI の試みはまだ少ない．松井らは国際ワークショップにおいて「おばけ工学」を提唱し,「モデル化できない外部の他者」を HAI の中で扱うことを提唱している．

おばけ・天狗・神などを「外部の他者」性を持つエージェントとして捉える見方は松井 [松井 22] で詳しく述べられている．要約すると，これらを「外部の他者」と見なすのは，伝統的にこれらが人間世界（内部）の論理や因果関係から独立した存在であると見なされていたためである．小松 [小松 91] は，神隠し・天狗隠しを例として，神や天狗のようなエージェントの有用性を説いた．前近代における神隠し・天狗隠しは，失踪事件が起きた時にその責任を神や天狗に帰属させるものである．これは，人間界における失踪事件の責任追及を無効化すると同時に,「天狗の仕業であれば仕方がない」という理屈で，責任追及という行為そのものを無効化する機能を持っていた．これは，天狗や神が現実世界の論理や因果関係から隔絶した存在であるからこそ可能であったことである．現実世界（内部）の論理が遡及しないという意味で，彼らのいる世界は「外部」である．近代以降に因果論が世界の全てを覆うようになったため，このようなエージェントは駆逐された.「おばけ工学」とは，ロボットや AI を用いて，かつての天狗や神のような「外部の他者」性を持つエージェントを再現しようとする試みである．このような立場は，モデル化や社会性の実装といった人工知能的・自然知能的手段を採用せず，他者の不可解性・説明不可能性を積極的に擁護しようとする，天然知能的な立場である．

また，Matsui et al.[Matsui 21b] は論理的に不完全な発話を行うバーチャルエージェントが，むしろユーザの対話継続意欲を誘発することを示して，モデル化不可能性が重要な役割を果たしうることを示した．これはエージェントに不可解性・説明不可能性を実装することの工学的利点を示したものであり，モデル化可能・理解可能なエージェントこそが有用であるとする，人工知能的・自然知能的 HAI の立場に問題提起を行うものである．さらに Matsui[Matsui 21a] は，幽霊や異星人などの「異類」を信奉する度合が高い人ほど，失敗をしたロボットに対する信頼が下がりにくいことを

示した．前述のように，幽霊などの異類は論理や因果関係，説明可能性から自由な「外部の他者」であると考えるなら，この結果はやはり説明不可能なエージェントこそが,「信頼」という工学的課題を解く際にも有効に働くことを示すものである．これらは天然知能的 HAI の実践例である．

3　HAI 研究におけるトリレンマ

郡司 [郡司 19, 郡司 20] は，天然知能を論じる上で,「自由意志·局所性·因果律」のトリレンマについて述べ，ある系の中ではこの 3 つは同時には両立せず，2 つしか満たすことはできないことを示した．このトリレンマは，元来は量子力学の領域で「アインシュタイン＝ポドルスキー＝ローゼンのパラドックス (EPR パラドクス)」と呼ばれるパラドクスを基にしている [Torrens 18]．ここでは，観測者の自由意志と，量子の局所性 (自分の状態を知らずに相手の状態を知ること) と，因果律は両立しないことを示したものである．ダメットは「酋長の踊り」という思考実験で，我々が自明なものであるとしている因果律が実際には相対的なものに過ぎないことを示している [Dummett 78]．郡司はさらにこれを拡張して，微視的な量子の世界に限らず，我々の経験的な日常世界でもこのトリレンマが成立することを示した．ここでは郡司によりアレンジされた「酋長の踊り」を参照する．

これは以下のような思考実験である．とある未開地の村には，若者はある年齢になると成人と認められるために一人でライオン狩りに出かけて，ライオン狩りを成功させて帰ってこないといけないという掟がある．対象となる若者は，2 日かけてライオンの生息地まで行き，2 日かけて狩りを行い，2 日かけて帰ってくる．その村の酋長は，若者が出発した後は，若者が帰還する 6 日後まで若者の無事を祈って踊りを踊る．さて，ある日この村に西洋人の科学者（因果論者）がやってきて，この風習のことを聞いて当然の疑問を口にした.「4 日目には，もう若者のライオン狩りが成功したかどうかは決まっているのだから，5 日目と 6 日目の踊りは不要でしょう」と．つまり，仮に酋長の踊りと若者のライオン狩りとの間に因果関係が存在することを認めたとしても，それは時間的に「酋長の踊り」が「若者のライオン狩りの成功」に先行している場合に限られる．しかし，それを指摘された酋長は答える.「いや，私は今まで若者のライオン狩りの儀式がある度に 6 日間休まず踊り続けてきた．その結果，若者はみんな無事に帰ってきたのだ」．ダメットは，この酋長を説得して，酋長の行為では因果関係が成立していないことを示せるか，と問う [Dummett 78]．結論は，言葉だけでは酋長の経験則を否定することは不可能である．そして，酋長の主張は，我々の直感にも反しない．森田 [森田 11] が言うように，試験が終わってすでに自分の点数が確定してしまった後でも,「自分が受かっていますように」とお祈りをすることは不自然な行為ではない．ダメットは，このような「逆因果」が我々の認知と何ら矛盾しないことを示し，因果律とは相対的なものに過ぎないことを示した．さて，この西洋人の科学者の世界観の中では，因果関係の逆転は認められない．一方で，酋長が自身の自由意志で踊っていること，また酋長と若者がお互いの状態知ることができないことは前提とされている．すなわち，自由意志・局所性・因果律のトリレンマの中で，自由意志と因果律を擁護し，その代償として局所性を切り捨てている．

一方，酋長の世界観の中では，自分が自由意志で踊っていること，因果関係が逆転しうることは自明のこととされているだろう．一方で，自分が踊っている限り，若者は無事である，すなわち自分が若者の状態を把握できることは確信している．すなわち，彼は自由意志と局所性を擁護し，その代償として因果律を切り捨てている．因果律を切り捨てているという点で，この立場は天然知能的であると言える．ここで第三の立場を導入してみよう．この村に今度は運命論者がやってきて，酋長が踊りを踊って，若者が無事に帰ってくることは，最初から決まっているのだと主張したとしよう．この運命論者の世界観では，酋長には「踊るか，踊らないか」を自分で決める自由意志が無い．すなわち，この運命論者は局所性と因果律は擁護するが，その代償として自由意志を否定せざるを得ない [郡司 19].

このトリレンマに対する立場は，それぞれ自然知能・天然知能・人工知能に対応するものだと言える．よって，自然知能的 HAI，天然知能 HAI，人工知能 HAI それぞれの立場に立つ研究者らのアプローチ方法をよりわかりやすく比較するには，それぞれの立場で研究者らがどの立場を取っているかを考察するのが有用だろう．以下ではそれを試みたい．なお，以下では説明のため，先に提示した順ではなく，自然知能・人工知能・天然知能の順で考察を行う．

まず自然知能的 HAI について述べる．ここでは，自分と他者が同じ規範に従って動くことを前提としている．その規範の中では他者の自由な意思は認めるが，その自由意志によって他者が規範を逸脱することには極めて不寛容である．さらに,「私」の出す社会的なシグナルには,「他者」は必ず返答してくれないといけない（逆も然り）という前提がある．すなわち，この立場は，他者の自由意志を認めない．自由意志を認めない代わりに，決定論と因果律を維持しようとするのが自然知能的 HAI の立場である．

他者の自由意志を認めず，未知のものはあくまでもデータの付けたしとして認識する自然知能的 HAI の「わたし」は，自閉症者に例えることができる 2. では，

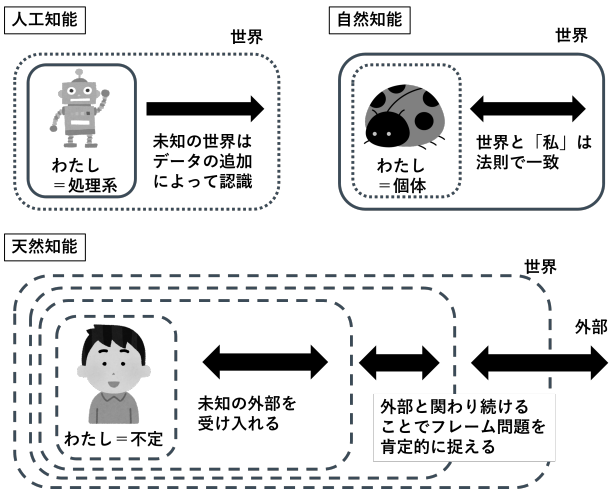

図 1: HAI の 3 つの立場の概念図

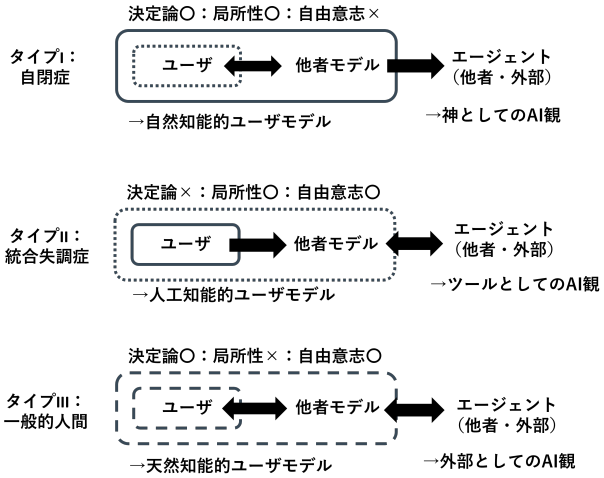

図 2: HAI の 3 つの立場における他者のモデルとトリレンマ

このような他者のモデルを実装した人工物 (以降 AI と呼ぶ) は，どのようなものであるとみなせるだろうか．他者の自由意志を認めず，世界の全てをあらかじめ認識しているという前提を持つこの AI は，神に例えることができるだろう．まとめると，自然知能的 HAI は，自閉症的な「わたし」のモデルを基に，他者を中にあらかじめ存在するものとして扱う (つまり，この AI は，世界の中に他者がいることをあらかじめ「知っている」)，神のような AI を作ろうとしている立場である．

次に人工知能的 HAI は，他者を「わたし」の内部とする立場である．この立場に立つ研究者らにとって,「他者」とは完全に「わたし」の中のモデルに従って行動する存在である．彼らが認識できる「他者」とは，あくまで「わたし」の中の変数からなるモデルに投影された存在であり，その行動は完全に記述可能でなければならない．もし他者が自分の予想を裏切る，つまり自由意志を見せるような振る舞いをした場合は，その振る舞いを無視するか,「こんなものは他者ではない」とみなす．例えば，ロボットであれば「故障したのだ」とみなすのだ．この立場では，他者を認識する「わたし」は処理系であり，データとしての他者を認識して処理する装置である．この立場では,「わたし」は決定論を放棄している．他者が因果律で記述できることよりも，自分の持っているモデルにおさまることのほうを重要視する．言い換えれば，決定論を放棄することで，他者の自由意志と局所性を維持しようとする立場である．因果律を認めないということは，一見すると経験科学として不合理なように思われる．しかし，前述のように，ダメットは因果律は相対的なものに過ぎないことを示しており [Dummett 78]，量子力学においても因果律の逆転の可能性は議論されている [森田 11]．

決定論よりも,「わたし」の持っている他者モデルのほうを擁護するこの立場は，統合失調症に例えることができる 2．では，このようなモデルを実装した AI とは何だろうか．それは，ユーザにとって求める出力を返してくれる，便利なツールに過ぎないものだろう．人工物知能的 HAI 研究者は，AI をあくまでツールとしてみなし，レヴィナスの意味での他者性を認めない立場である．

最後に天然知能的 HAI である．この立場では，レヴィナスを引用すれば「他者は決して全体性に回収されることのない無限の存在」として定義する [Levinas 61]．この立場では，他者は「わたし」にとって全く予測できない存在であるため，他者の自由意志は自明である．また，この「わたし」は外部と接することによっていかようにも変化していく可能性を持っている存在であり，その意味で決定論的性質を持っている．一方で，このモデルでは,「わたし」と他者は常に双方向的なインタラクションによって繋がる．つまり局所性は放棄されている.「わたし」だけが他者のことを認識している，もしくはその逆ということはありえない．どちらかがどちらかを認識する時には，必ず相手も認識していることになる．このような他者観は，一般的な人間の持っている他者観に最も近いものだろう．よって，天然知能的 HAI とは，実は一般的な人間の感覚に最も近い研究姿勢であると言える．では，このモデルを実装した AI はどのようなものか．それは,「わたし」にとって全く予期できない外部，レヴィナスの意味での真の他者である．

何より，局所性を否定することによって，HAI が対象とするロボットやバーチャルエージェントを，人間世界のしがらみから自由な存在としてデザインすることが可能になる [松井 22]．小松和彦は，前述のように，前近代の日本の農村における「神隠し」は，人間社会の外部にいる神や天狗に責任を負わせることで，人間社会内の責任追及を無効化するシステムであったと指

摘した [小松 91]．これは社会の外部に，その社会の内部のルールが全く通用しない他者を想定しているという点で，単なる社会的なルールを規定するシステムという枠に収まらないものである．局所性を排した天然知能的 HAI は，神や天狗のような人間社会の外部に立つロボットやバーチャルエージェントをデザインできるという発展性を含んでいる

4 結論

本論文では，郡司 [郡司 19] を参照しつつ松井 [松井 22] を補足するという形で，ヒューマンロボットインタラクション (HAI) におけるトリレンマについて論じた．その結果，自由意志，局所性，因果律のうち，人工知能的 HAI は決定論を，自然知能的 HAI は自由意志を，天然知能的 HAI は局所性を切り捨てることで成り立っていることを論じた．そして，局所性を切り捨てている天然知能的 HAI が最も発展性を有しているということを述べた．今後課題とすべきなのは，なぜ HAI の中で天然知能的 HAI が小勢力に留まっているのかであろう．その理由としては,「局所性を否定する」という思考法が工学者にとって馴染みにくいこと，および機械学習やデータサイエンスの影響下にある HAI 研究者の多くが「理解不可能な外部の他者」という概念を理解できないことに求められると思われる．この問題の解消には，そのような研究者に「外部」と接する体験をさせることが有用ではないかと思われる．

参考文献

[Cross 19] Cross, E. S., Hortensius, R., and Wykowska, A.: From social brains to social robots: applying neurocognitive insights to human–robot interaction (2019)

[Dummett 78] Dummett, M.: *Truth and other enigmas*, Harvard University Press (1978)

[Feine 19] Feine, J., Gnewuch, U., Morana, S., and Maedche, A.: A taxonomy of social cues for conversational agents, *International Journal of Human-Computer Studies*, Vol. 132, pp. 138–161 (2019)

[Kumazaki 22] Kumazaki, H., Yoshikawa, Y., Muramatsu, T., Haraguchi, H., Fujisato, H., Sakai, K., Matsumoto, Y., Ishiguro, H., Sumiyoshi, T., and Mimura, M.: Group-based online job interview training program using virtual robot for individuals with autism spectrum disorders, *Frontiers in Psychiatry*, Vol. 12, p. 704564 (2022)

[Levinas 61] Levinas, E.: Totalité et infini: essai sur l'extériorité (1961)

[Matsui 21a] Matsui, T.: Relationship between users' trust in robots and belief in paranormal entities, in *Proceedings of the 9th International Conference on Human-Agent Interaction*, pp. 252–256 (2021)

[Matsui 21b] Matsui, T., Tani, I., Sasai, K., and Gunji, Y.-P.: Effect of Hidden Vector on the Speech of PRVA, *Frontiers in Psychology*, Vol. 12, p. 627148 (2021)

[Sakamoto 21] Sakamoto, T., Sudo, A., and Takeuchi, Y.: Investigation of model for initial phase of communication: analysis of humans interaction by robot, *ACM Transactions on Human-Robot Interaction (THRI)*, Vol. 10, No. 2, pp. 1–27 (2021)

[Torrens 18] Torrens, F. and Castellano, G.: EPR paradox, quantum decoherence, qubits, goals and opportunities in quantum simulation, *Theoretical Models and Experimental Approaches in Physical Chemistry: Research Methodology and Practical Methods*, Vol. 5, pp. 317–334 (2018)

[郡司 19] 郡司ペギオ幸夫：天然知能, 講談社 (2019)

[郡司 20] 郡司ペギオ幸夫:やってくる, 医学書院 (2020)

[坂本 19] 坂本孝丈, 吉岡源太, 竹内勇剛 : 話しかけ場面における相手の受容度に応じた接近行動のモデルに基づく分析, 知能と情報, Vol. 31, No. 5, pp. 842–851 (2019)

[三宅 16] 三宅陽一郎 : 人工知能のための哲学塾, ビー・エヌ・エヌ新社 (2016)

[寺田]

[寺田 14] 寺田和憲, 勅使宏武, 伊藤昭 他 : ロボットが表出する感情の社会機能的評価, 研究報告ヒューマンコンピュータインタラクション (HCI), Vol. 2014, No. 8, pp. 1–8 (2014)

[寺田 17] 寺田和憲：機械に対する悪意の帰属 (2017)

[小松 91] 小松和彦：神隠し: 異界からのいざない, 弘文堂 (1991)

[松井 22] 松井哲也 : ロボット工学者が考える「嫌なロボット」の作り方: ヒューマンエージェントインタラクションの思想, 青土社 (2022)

[松井 23] 松井哲也 : HAI における 「他者モデル」 の限界と アップデートへの見通し, 認知科学, Vol. 30, No. 4, pp. 536–541 (2023)

[植田 16] 植田一博, 小野哲雄, 今井倫太, 長井隆行, 竹内勇剛, 鮫島和行, 大本義正 : 意思疎通のモデル論的理解と人工物設計への応用 (¡ 特集¿ 認知的インタラクションデザイン学), 人工知能, Vol. 31, No. 1, pp. 3–10 (2016)

[森田 11] 森田邦久 : 量子力学の哲学: 非実在性・非局所性・粒子と波の二重性, 講談社 (2011)

[大澤 20] 大澤正彦, 奥岡耕平, 坂本孝丈, 市川淳, 今井倫太 : 認知的インタラクションフレームワークに基づいた他者モデルの提案 HAI シンポジウム 2020 (2020)