AI時代の教育論文誌 第1巻 pp. 1-6（受理日：2019年5月5日）（AI時代の教育論文誌 第269巻 第9号 pp. 50-55）

チューリングテストによるAIと人の特徴分析の予備的研究

A Preliminary Study for Feature Analysis of AI and Humans Using Turing test

赤堀　侃司

（一社）日本教育情報化振興会・（一社）ICT CONNECT 21

人工知能（以下，AIと略す）と人の特性についてチューリングテストを用いて抽出した結果を基に，学習に適用することを目的として本研究を実施した。チューリングテストを用いて，問いかけに対してAIが答える回答と特定の人が答える回答を，60名の実験協力者に提示し，どちらの回答がAIかを判定してもらいその正答率を求めた。同時にAIの回答内容を吟味して，もし人間だと仮定したらどの年齢レベルかを推定してもらった。その結果，AIだと正答した率は17問全体の平均値が0.81と高い値であった。また人間だと仮定した時の推定年齢は，平均的にはおよそ中学生レベルと推定された。ただし，この結果は提示した問いの内容に強く依存することがわかった。さらに年齢推定において，その理由を自由記述で書いてもらい分析した結果，AIと人間の顕著な特徴が見出せた。この結果を元に，AI時代を生きる子どもたちの学習についての示唆を得た。

キーワード：AI，AI時代，チューリングテスト，ビッグデータ，学習方法

1．はじめに

これからの時代はSociety5.0に代表されるように，大きな社会変革と，未来の学びの姿が不連続的に進化すると言われている（Society 5.0 に向けた人材育成に係る大臣懇談会，2018）。その社会において，AIやビッグデータが大きな役割を果たすことは言うまでもない。その時代をAI時代と呼べば，その未来を生きる子どもたちには，どのような資質・能力が必要とされ，どのような学習が求められるだろうか（赤堀侃司，2019）。プログラミング教育（赤堀侃司，2018）やSTEM教育も，その学習内容・方法として有効であるかもしれないが，まだ模索中と言ってよいだろう。

そこでAI時代における学習の在り方を探求するためには，AIと人間はどこが違うのかを，始めに明らかにする必要がある。これまでにも，いくつかの優れた研究報告がある。例えば，新井紀子（2018）や奈良潤（2017），ゲルトギーゲレンツァー（2010）などが参考になる。そこで論じられてきたことは，AIの強みと弱みや読解力などの小中学生が身に付けたい学力や，人間の持つ暗黙知や優れた直感力とこれらの力を生かす方法など，これからの学習への示唆を含んでいる。

さらにロボット研究が進んでおり，理科授業においてロボットとの対話を通して理解を促進する研究（小松原剛志，塩見昌裕，他，2015）や，ロボットを介在した観察学習の研究（ジメネスフェリックス，加納政芳，他，2017）や，語学教育への適用も盛んである。

いづれにしてもAI時代には，新しい技術との協調が不可欠であるが（例えば，楠見孝，西川 一二，2018），そのためには，AIと人間との認知的な違い，最近ではロボットを始めとして感情研究が発展しているので（例えば，石川葉子，水上雅博，他，2018），感性の違いなども明らにする必要がある。

そこで本研究では，これまでの先行文献の知見を参考にしながら，チューリングテストによるAIと人の比較や特徴分析を行い，その結果を元にAI時代の学習について考察をして示唆を得ることを目的としている。

2．チューリングテストによる実験方法

2.1　チューリングテスト

チューリングテストは，アラン・チューリングの1950

受理日　2019年5月5日
Akahori Kanji: “A Preliminary Study for Feature Analysis of AI and Humans Using Turing test”
Japan Association for Promotion of Educational Technology，1-9-13 Akasaka Minato-ku Tokyo 107-0052 Japan
URL：http://www.gakujoken.or.jp/gakai/ronbun/akahori2019_07.pdf

年の論文「Computing Machinery and Intelligence」の中で提案された人間と機械を判別する思考実験として知られているが，その概念図を図１に示す。このチューリングテストを教材とする談話を，以下，簡便に談話と呼ぶ。

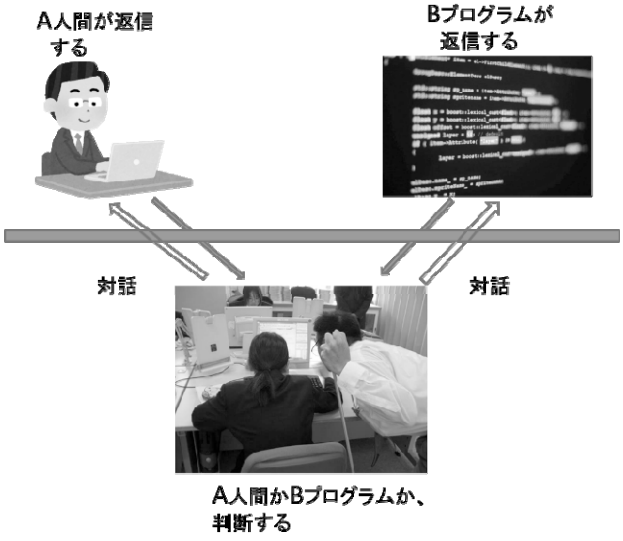

図1　チューリングテスト

図１のように，壁で仕切られた部屋を想定する。壁の手前の人間が，A人間とBプログラムの両方に問いを発信する。問いを受けたAとBはどちらかが返信するが，その問いと回答を繰り返す。問いを発信した手前の人間は，図１のように返信を受けるが，その返信はAの人間からなのかBのプログラムからなのかを判定する思考実験である。１回の試行では，Aの人間かBのプログラムのどちらかが返信するが，この試行を何回も繰り返して，その判定結果である正答率が同じ程度であれば，プログラムは人間と同じ知能を持つと言ってもよいではないか，という考え方である。本研究では，BのプログラムにはAI技術を採用し，AIがどの程度人間に近いかを実験することにした。

また，近年のメディア環境の発展は著しく，テキストだけでなく，音声，写真，動画などの多様な情報媒体が一般的になったことから，本研究では，人間と機械との識別をこれらすべての情報媒体を対象にしてチューリングテストを行うことにした。その理由は，学習教材は上記のようなすべての情報媒体を含むからである。その概念図を図２に示す。

ただし本研究では，テキスト，写真，イラストの３種類の情報媒体だけを対象にして実験することにした。音声は，音声の質によるバイアスがかかる危惧があるために除外して，テキストにして提示した。AIとして，市販のGoogle Homeを用いた。問いに対して音声で回答が返ってくるが，これをテキストにして実験を行うことにした。

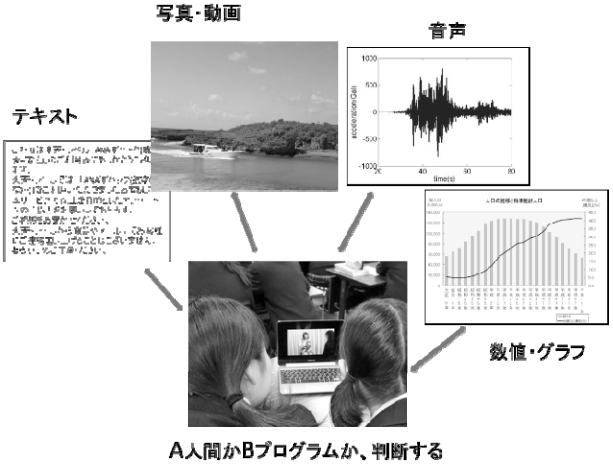

図２　他の情報媒体を含むチューリングテスト

2.2. 実験方法

本研究では，図１や図２のようなリアルタイムの判定をする実験ではなく，実験協力者に提示する談話を予め準備して，その談話に対して判定してもらう方法を用いた。その談話は，筆者が問いを発信して，AIの回答はGoogle Homeの応答とし，人間の回答は筆者の応答として，テキストにして提示する談話である。具体的な談話例を，図３に示す。

１　あいさつ

- おはようございます。
  - A　おはようございます。
  - B　おはようございます。
- 今日も暑いですね。
  - A　暑いときこそ、ホラー映画を見るのはどうでしょう。
  - B　今年は、特に暑いですね。
- 行ってきます。
  - A　行ってらっしゃい。お気をつけて。
  - B　行ってらっしゃい。

図３　提示する談話例

図３では，３つの問いに対してAかBのどちらかがAIでどちらかが筆者であるが，それを判定するという単純な方法である。この談話例では，Aがもし人間だとしたら，３つの問いのすべてに対してAの回答は人間からの

回答である。実験協力者には，AIはAかBのどちらですかという質問形式で判定をしてもらった。この談話例では，正解はAがAIであるので，Aと判定すれば正解になる。本研究では，図３を談話または問いと呼ぶことにする。問いは，実際には音声であるが，本実験では先に述べたようにテキストにして表示した。実験協力者には，AIと判定した場合，そのAIは人間だとしたらどのレベルかを判定してもらった。さらにその理由や根拠も自由記述で回答してもらった。なお，そのレベルを本研究では推定年齢と呼び，以下の４段階に分けた。

１：幼児レベル，２：小学生レベル

３：中学・高校生レベル，４：大学生かそれ以上

実験協力者は，都内の大学生で男女それぞれ30名からなる合計60名であり，2018年10月20日に都内の大学で実施した。合計17の談話を用意した。

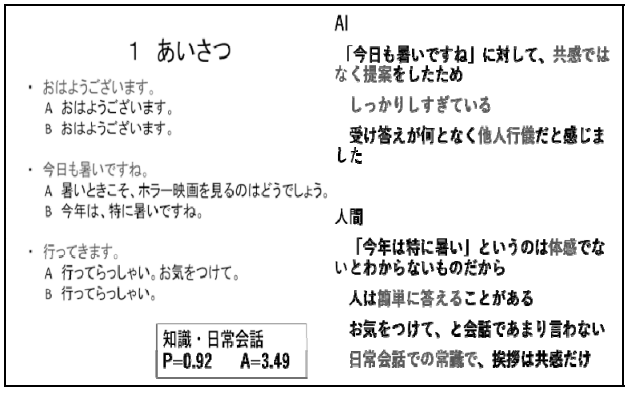

図４　談話と推定年齢と理由の例

詳細はこの後に示すが，回答のイメージをしやすくするために，談話と回答の結果をまとめた例を図４に示す。

図４において，□で囲まれた数字は，Pが正答率で，Aが推定年齢の平均値を示す。この場合AがAIなのでAと答えれば正解で，60名の実験協力者の正解率は0.92ときわめて高い。またその時のAIの回答のレベルを人間に例えれば，60名の実験協力者の推定年齢の平均値が3.49であり，高校生から大学生レベルである。自由記述による主な判断理由を，図４の右に示す。

３．分析と結果

3.1　正答率と推定年齢の結果

表１に，17問の正答率と推定年齢および分類の一覧表を示す。正答率は正解した割合のことであるから，この値が50％に近ければ，AIの回答か人間の回答かの区別が難しいことになるので，AIは人間との差がほとんどないという結果になる。

また推定年齢は，人間に例えればどのレベルかの推定なので，この値が大きければAIの回答のレベルは高いことになる。ただし推定年齢は，表１の脚注に明記しているように，正答者数を母数としている。つまり正しくAIだと判定した数を母数として，平均値を算出している。

さらに「１：幼児，２：小学生，３：中学高校生，４：大学生以上」のように，各レベルに対して値を割り当てている。これらの値は名義尺度であるので平均値を出すことは統計的には意味はないが，分析結果を解釈しやすいように数値として集計処理し，平均値を算出した。

表１　17問の正答率・推定年齢・分類の一覧表

No.	題 名	正答率	推定年齢[*1]	分 類
1	あいさつ	0.92	3.49	知識・日常会話
2	夏至	0.97	3.72	知識・教科書
3	首都	0.98	3.59	知識・教科書
4	足し算	0.83	2.16	思考・文章理解
5	大きさ	0.65	1.72	思考・比較
6	味噌汁	0.82	3.18	知識・手順
7	交通	0.80	3.44	思考・手順
8	海水	0.72	2.56	思考・理由
9	道路	0.85	2.14	思考・判断
10	桜	0.40	2.38	認識・写真
11	台風	0.92	3.38	認識・写真
12	スキー	0.73	2.98	認識・写真
13	兄弟	0.82	2.22	認識・イラスト
14	ビール	0.85	2.37	思考・理由
15	掛け算	0.83	2.36	思考・意味理解
16	プログラミング	0.85	3.69	知識・専門用語
17	眠れない	0.87	2.46	思考・判断

推定年齢[*1]：正答者を母数とする
１：幼児，２：小学生，３：中学高校生，４：大学生以上

また分類の項目は，問いの内容を検討して表１のように示した。その理由は，正答率も推定年齢も問いの内容に強く依存するからである。本研究では，知識，思考，認識の３つのカテゴリーに分類し，さらにそれらをサブカテゴリーに分類した。但し，３つのカテゴリーを中心に分析した。例えば，１の「あいさつ」の談話は図３に示した通りで，これは知識・日常会話と

して分類した。表１のように，知識の問いが５問，思考の問いが８問，認識の問いが４問である。認識については，正確には写真やイラストを見せて何を連想するかという問いであり，パターン認識や連想の問いなどの表現がわかりやすいが，本研究では認識というカテゴリー名にした。

3.2　正答率の分析

図５に，知識，思考，認識のカテゴリー毎に，正答率の高い順に並べ替えたグラフを示す。

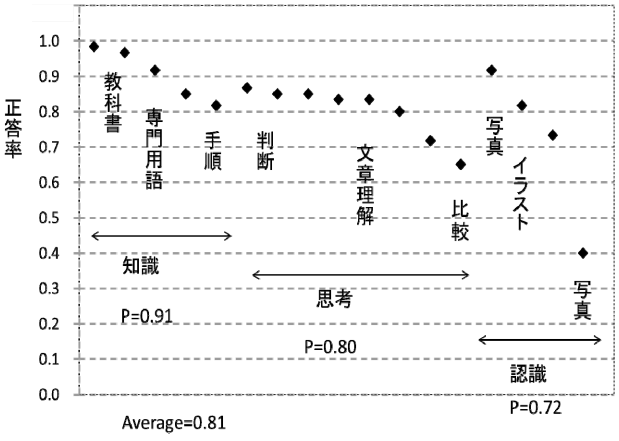

図５　カテゴリー毎の正答率のグラフ

図５に示す通り，17問の平均正答率は0.81であったので，この値は予想より高かった。この理由として，実験協力者である大学生は，日常的にAI技術が組み込まれたデジタル機器に触れていて，その特徴を無意識ながら理解していることが挙げられる。スマホを日常的に使っているので，スマホに組み込まれている音声認識や画像認識のレベルの高さについては熟知していると思われるからである。その意味では，実験協力者としての大学生は，チューリングテストについては適任者かもしれない。

この正答率は，当然ながらカテゴリーによって異なる。知識の平均正答率は0.91であり，思考の平均正答率は0.80で，認識の平均正答率は0.72であった。このことから，AIが返信した知識カテゴリーについては，AIらしい特徴があって容易に判別しやすかったと言える。思考カテゴリーについては，図５に示すように判断と文章理解と比較のサブカテゴリー間で差が大きい。すなわち同じ思考カテゴリーであっても，問いの内容に強く依存する。さらに認識カテゴリーでは，同じ写真であっても正答率に大きな差がある。

3.3　推定年齢の分析

図６に，知識，思考，認識のカテゴリー毎に推定年齢の高い順に並べ替えたグラフを示す。

図６に示すように，17問の平均推定年齢は2.81であった。表１に示すように，２が小学生，３が中学高校生レベルなので，小学高学年から中学高校生の間なので，およそ中学生レベルと推定された。しかし，これはカテゴリーに強く依存することは言うまでもない。

図６に示すように，知識カテゴリーの平均推定年齢は3.54と高く，しかもばらつきが小さい。AIは高校生や大学生レベルの高い回答をしたことがわかる。思考カテゴリーは，平均推定年齢が2.40と小学生から中学高校生の間であり，ばらつきが極めて大きい。これからAIの思考レベルは進化途上であって，理由や文章理解などのサブカテゴリーに大きく依存すると言える。認識カテゴリーの平均推定年齢は2.74であるが，これもばらつきがきわめて大きい。

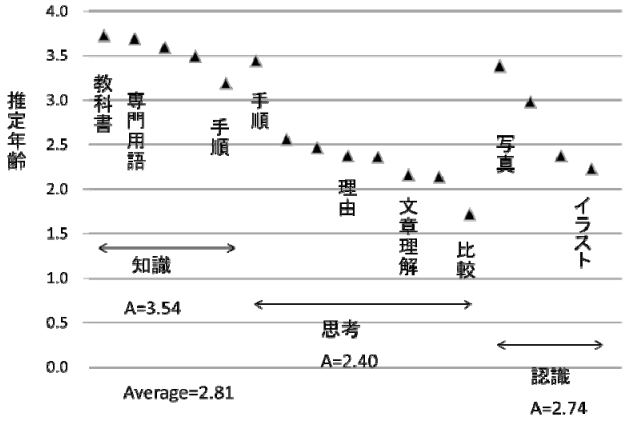

図６　カテゴリー毎の推定年齢のグラフ

3.4　正答率と推定年齢の相関分析

図７に，正答率と推定年齢の相関分析のグラフを示す。相関係数は0.51で，この値自身にあまり意味はないが，図７の回帰直線の上部に知識カテゴリー（■で示す）が，下部に思考カテゴリー（●で示す）が，全体に認識カテゴリー（▲で示す）が分布していることが興味深い。サブカテゴリーの内容に依存することはあるが，カテゴリー毎の特徴が読み取れる。その詳細は後で示すが，実験協力者に何故AIだと判定したのかという問いに対する自由記述の回答を分析した結果の特徴を，図７に書き加えた。

知識カテゴリーは，「教科書的・定型的・辞書的・専門的・順序正しい回答」という特徴があった。人間はあい

まいで，その時その場で対応する柔軟性を持っているが，AIの回答した専門的な内容を読んで，その推定年齢は高いと判断したと思われる。思考カテゴリーは，逆に推定年齢は低く，その理由として「理由・推測・意味付け・あいまいな問いが難しい」が挙げられた。AIは，まだ思考レベルでは人間に劣っていると大学生たちは判断した。特に，理由を述べる，文章を理解する，意味付けをするなどでは，人間の能力が優れていると感じたと言える。さらに認識カテゴリーでは，写真やイラストを見てどのような連想をするのか，どのように認識したのかという問いで，連想の違いによってばらつきが大きい結果になっている。

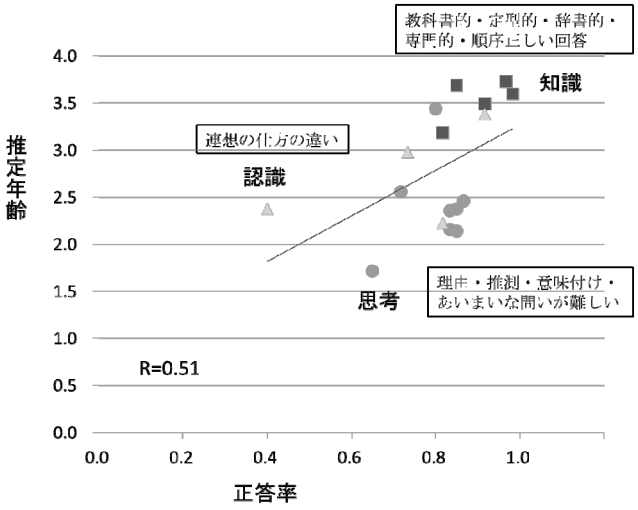

図7　正答率と推定年齢の相関グラフ

3.5. 自由記述の内容分析

先に述べたように，実験協力者に何故AIだと判定したのかその理由を自由記述してもらったが，その内容を主観的に集計し特徴を抽出した。17問X60名の回答を，カテゴリー毎にそれぞれの特徴をKJ法にしたがって分類し，AIと人間の特徴を抽出した。その結果を表２に示す。

表２　自由記述の内容分析のまとめ

	AI	人間
知識	辞書的・専門的・順序正しい定型的・正確・数字など	感覚的・常識的・時と場合による・あいまいさ・大まかなど
思考	最適解を出す・文章理解が困難・一般的な方法・関連付けが困難・前例による	相手の立場・推測して理解・個人的な経験・暗黙的な知識・意味付け・文章を補う
認識	表面からの情報・細分化した情報・表面的な連想	全体的な情報・内容的な情報・総合的な連想

表２に基づいて簡単に述べる。知識カテゴリーでは，AIと人間の対比では，辞書的－感覚的，専門的―常識的，順序正しい－時と場合による，定型的－あいまいさなどが抽出された。AIが辞書や検索などを介在して回答することを考えれば，データに基づいているのでこのような特徴は納得できる。これに対して人間は，常識を持っている。この常識をデータとしてAIに実装することが極めて困難であることは，1960年代から1990年代の第２次AIブームの研究で得られた知見であるが，改めてこの常識的な知識は本質的な意味を持っていることがわかる。

思考カテゴリーでは，AIと人間の対比では，最適解を探す－相手の立場を考える，文章理解が困難－推測して理解，一般的な方法－個人的な経験，関連付けが困難－意味づけをする，などの特徴が見出せた。この場合はどうすればいいですか，という問いに対して，AIはデータに基づいて前例を探して最適解を求めるが，人間は最適というよりも相手の立場も考慮して回答する。それは，時と場合によるという柔軟性を持っているからと考えられる。問いに対して，人は状況に応じて，時と場合によって，個人的な経験を含んで答える。また，あいまいな文章を提示した問いでは，AIは文章理解は難しいが，人間は文章を補完して読むので，つまり推測しながら読むので理解できる。このことが，AIはまだ思考という点では進化していないと判断して推定年齢が低くなったと言える。

認識カテゴリーでは，AIと人間の対比では，表面的な情報－全体的な情報，表面的な連想－総合的な連想などの違いが抽出された。AIの画像認識は表面の情報から抽出するが，人間はその他にも脳に蓄えられた情報や経験などを追加するので，このような違いになったと言える。但し，画像認識技術は急速に進化しているで，かなりのレベルで感情までも判定することができる。実験協力者はそこまでの知識がなく，写真やイラストを含んだ感性的な問いでは判定を間違えることがあり，ばらつきが大きくなったと考えられる。

4．結論と考察

以上の結果と分析を，結論としてまとめる。

（1）大学生 60 名を対象にしたチューリングテストでは，AI と人間の判定は 0.81 と高い正答率を示した。これは，現代の若者が日常的にスマホなどに

触れていて，AI を身近に感じているからと考えられる。

（2）AI の正答率や推定年齢は，問いの内容である知識・思考・認識のカテゴリーに大きく依存する。全体的には，知識カテゴリーの推定年齢が高く，思考カテゴリーが低く推定された。全体の平均は，ほぼ中学生レベルと推定された。

（3）AI と人間の特徴の対比では，自由記述の内容を分析した結果，専門的―常識的，順序正しい―時と場合による，一般的な方法―個人的な経験，関連付けが困難―意味付けをする，表面的な連想―総合的な連想などの特徴が見出せた。

以上であるが，学習という観点で考察すれば，AIの優れた専門的内容や論理的な思考については，あいまいで感覚的で経験的に表現する人間が学ぶことは多い。例えば，文章を書く，引用する，論文として完成するなどは，AIと協調する必要があろう。思考カテゴリーについては，人間の優れた特性が抽出され，情報が不足している場合には，個人的な経験に基づいたり，他の情報から補足したり，常識という膨大な知識で補完するという優れた特性があることがわかった。一言で言えば，総合的に処理しているわけで，総合的な処理や情報の関連付けは，AI時代においても，さらに重要で必須の資質・能力になると思われる。

認識カテゴリーについては，AIは進化途上であるがロボットの感情や感性も研究されており，AI時代にはロボットとの共存は当たり前の光景になるであろう。広くとらえれば，それはコミュニケーションの在り方と言ってもよいが，人とのコミュニケーションの他にAIとのコミュニケーションが問われる。人と人の間でも難しい関係が，感情を読み取ったり表現したりするAIの出現で，学校生活も含めて大きく変わる社会になるかもしれない。

本研究結果は，サブカテゴリーの分類や実験協力者の特性などに依存することが大きいので，例えば，ブルームの教育目標分類に基づく学習カテゴリーや，大規模な実験協力者数による調査などを今後の課題として，研究を継続する予定である。

［謝辞］

最後に本論文は，（NPO）教育テスト研究センターの支援と，科学研究費助成金・基盤研究C（代表，赤堀侃司，課題番号15K01034）の支援を受けたことを明記して，厚くお礼申しあげる。

【参考文献】

赤堀侃司（著）（2018）プログラミング教育の考え方とすぐに使える教材集、ジャムハウス

赤堀侃司（著）（2019）AI 時代を生きる子どもたちの資質・能力、ジャムハウス

新井紀子（著）（2018）AI vs. 教科書が読めない子どもたち、東洋経済新報社

石川葉子，水上雅博，他（2018）感情表現を用いた説得対話システム、人工知能学会論文誌 / 33 巻 1 号 p. DSH-B_1-9

楠見孝，西川一二（2018）人と AI 協働社会の認知に及ぼすコンピュータ経験と不安の影響、日本認知心理学会第 16 回大会セッション ID: p06-004

ゲルトギーゲレンツァー（著），小松淳子（翻訳）（2010）なぜ直感のほうが上手くいくのか？,インターシフト

小松原剛志、塩見昌裕、他（2015）理科室で授業の理解を支援するロボットシステム、日本ロボット学会誌 33（10），789-799

ジメネスフェリックス，加納政芳，他（2017）学び方が変化するロボットとの共同学習がもたらす Learning by Observing の実現可能性、人工知能学会誌 32 巻 2 号 p.D-G51_1-12

Society 5.0 に向けた人材育成に係る大臣懇談会（2018）、Society 5.0 に向けた人材育成 ～ 社会が変わる、学びが変わる ～、http://www.mext.go.jp/component/a_menu/other/detail/_icsFiles/afieldfile/2018/06/06/1405844_002.pdf（2019 年）

奈良潤（著）（2017）人工知能を超える人間の強みとは、技術評論社