人間を騙すロボット

A robot being able to deceive human

寺田和憲* 小野康平 伊藤昭
Kazunori Terada Kouhei Ono Akira Ito

岐阜大学
Gifu University

Abstract: In the present study, we investigated whether a robot is able to deceive human by producing a behavior against him/her prediction. Feeling of being deceived by a robot would be a strong indicator to investigate whether the human treat the robot as an intentional entity. We conducted a psychological experiment in which a subject played Darumasan ga Koronda, a Japanese children's game, with a robot. The experimental result indicated that unexpected change of a robot behavior gave rise to an impression of being deceived by the robot.

はじめに

騙すとは真実とは異なることを意図的に他者に信じさせることである．意図的でない場合は単に事実の誤認や知識の欠如として捉えられる．意図的であっても，騙しが成功するためにはその意図を相手に見抜かれてはならず，騙そうとする意図が相手に伝わってしまった瞬間に騙すことは失敗に終わる．従って，騙すということは，騙そうとする意図を隠した上で真実と異なることを伝えることだと言える．

他者が客観的事実について嘘を言っている場合には，十分に情報を収集することによって真偽の判断が可能である．しかし，本当は搾取しようとしているのに親切に振る舞っている場合など，行為者が心的な目的について偽っている場合は，真の目的を正確に同定するのは不可能である．このような場合には，観察者は観察可能な表面的振舞いから本当の目的を推定するしかない．詐欺師は，真の意図を隠し，誤った意図が推測されるように，発話を含めた見かけ上の振舞を巧妙に演じる．正直な人は他者の見かけ上の振舞をそのまま受け取ってしまい容易に騙されてしまう．

本稿では，他者の振舞を観察することだけからは直接知覚できず，観察者が他者の頭の中に帰属させることによって存在する他者の心的な目的のことを**意図** *(intention)* と定義する．一方で振舞を観察することによって直接知覚可能な目的のことを**目的** *(goal)* と定義する．人間が振舞から架空の意図を想定し，帰属させることができるのは**心の理論** *(Theory of Mind)* を持つからだと言われている [2]．また，そのような能力の欠如は自閉症として知られている．

振舞と意図の関係が一対一ではなく，文脈によって変化し，曖昧であることが意図を読むことの難しさの原因である．通常のコミュニケーションにおいて，しばしば表面的に観察される振舞と意図が異なる場合がある [9]．例えば,「今何時か分かりますか?」という質問をされたとする．この発言の表面的な意味は現在時刻を知っているかどうかを問うことであり，Yes or No の回答を要求している．しかし，この表面的な意味の通りに「分かります」もしくは「分かりません」と答える人は少ない．多くの人は「今は 10 時です」などと現在時刻を答える．これが可能になるのは，質問者の意図が，相手が時間を知っているかどうか知りたいことではなく，現在時刻を知りたいことと推測可能だからである．日常のコミュニケーションでは文脈や状況など様々な情報を駆使して表面的な振舞の観察から真の意図を推測することが要求される．このような意図の推測を可能にしているのは心の理論であり，心の理論は脳内の心の理論ネットワークの賦活として捉えられている [4]．また，そのような心的な構えを用いた振舞理解戦略のことを Dennett は**意図スタンス** *(intentional stance)* と呼んだ [3]．

ロボットの振舞が多様かつ多義的になればなるほど，人間がロボットに対して意図スタンスを採用することは重要になってくる．人間は，通常人工物に対して**設計スタンス** *(design stance)* を採用する [3]．設計スタンスによる振舞理解では心的な目的である意図を想定するのではなく，振舞を規定しているアルゴリズムを想定する．アルゴリズムによって規定される振舞は入力を固定すると出力が一意に決まり (複雑な分岐があったとしてもアルゴリズムとして記述できるということ)，シ

*連絡先： 岐阜大学
岐阜市柳戸 1-1
E-mail: terada@gifu-u.ac.jp

ステマチックに振舞を予測可能である．また，設計的振舞の特徴として失敗や例外に対処できないということがある．一方，アルゴリズムではなく目的によって駆動される主体 (意図的主体) は同一の目的を達成するために状況に応じて手段を変えることができる [6][5]．ロボットの振舞が多様になってくると振舞のアルゴリズム的な解釈は破綻する．そのために，振舞を意図というういう単一のシンボルのもとに抽象化する振舞理解戦略 (意図スタンス) が有効になってくるのである．

これまでに人間がロボットのことを意図的な存在として捉えるかどうかについて調べた研究がある [1][11]．意図的な存在として捉えたか否かは，ゴール状態に達しない未完成のタスクを見せられた幼児がそのタスクを完遂できるかどうかによって調べる方法 [1] や意図的な存在だったかをアンケートによって直接問う方法 [11] によって調べられている．

本研究では，人間がロボットに騙されるかどうか (騙されたと感じるかどうか) によって人間がロボットのことを意図的な存在としてみなすかどうかを調べる．騙しは設計スタンスで捉えると単なる誤動作である．その振舞に隠された意図を帰属させるから騙しだと理解できるのである．従って，人間がロボットに騙されたと感じることはロボットを意図的な存在だと捉えている有力な証拠となる．本研究では，子供の遊びである「だるまさんがころんだ」を題材として，ロボットが予測を裏切る突発的行動を取った場合に，その行為を騙しだと感じるかどうかについて調べる．

1 だるまさんがころんだにおける騙し戦略

だるまさんがころんだは日本の子供のゲームの一つである．同様の遊びが世界中に存在しており，例えば英語圏では Red Light, Green Light という名前で知られる．この遊びの面白さは，鬼が「だるまさんがころんだ」という 10 音節を唱えている間に動きを同定されることなく鬼に近付くことである．これは自然界において，ライオン等の捕食者が鹿などの被捕食者に動きを悟られることなく忍び寄り，捕獲に至る過程に似ている．このゲームは鬼とプレーヤ双方で勝ちの定義が異なる．鬼はプレーヤが動いていることを同定すれば勝ちとなり，プレーヤは鬼に動きを同定されることなく鬼にタッチし逃げ通せれば勝ちとなる．

ゲームが対称でないために，鬼とプレーヤは勝つために取る戦略が異なる．プレーヤは鬼が文を唱えている間に鬼に接近するが，動きを同定されないために，文の唱え終わりを予測して動作を停止しなければならない．一方，鬼は接触されることを阻止するべく，動いている参加者を見付けることを目標にし，「だるまさんがころんだ」を唱える速度やタイミングを様々にコントロールする．そうすることで，鬼はプレーヤの予測を裏切ることができ，その結果，プレーヤは思わず動いてしまったり動きを停止できなかったりする．具体的には次のような戦略が考えられる．

図 1: だるまさんがころんだをプレイするロボット．筐体内の LED を点灯した状態．

1. だるまさんがころんだの唱詠速度を途中で速度を急激に上げ素早く振り向く
2. 唱えに入ると見せかけてまたプレーヤの方を振り向く

これがだるまさんがころんだに見られる騙しである．本研究では騙し戦略の 1. をロボットに実装した．

2 実験

これまでに述べたことを踏まえ，本研究では次の仮説を検証するための実験を行った．

仮説 人間の予測を裏切るようなロボットの振舞は，人間にロボットに騙されたと感じさせる．

この仮説を検証するために，被験者が騙されたと感じるであろう状況を作りだし，事後にアンケートによって調査を行った．具体的な方法を以下に詳述する．

2.1 実験装置

我々はだるまさんがころんだをプレイできるロボットを作成した (図 2 参照)．

ロボットの全高は 110cm で，2 脚に駆動用モータ (maxon A-max 32，20W)，2 脚にキャスターを装着した流線型の 4 脚を有する．ロボットの最高速度は約

80cm/sec である．4 脚が上部に向かって集合した部分に直径 40cm の乳白色のアクリルの球体が乗っている．球体の中には制御用のコンピュータを搭載している．コンピュータは acer 社 Aspire Revo (CPU: Intel Atom 230 (1.6GHz), チップセット: NVIDIA ION) で OS は Fedora 10 (kernel 2.6.27) である．このコンピュータは 6 ポートの USB を持ち，モータ制御，後述するカメラからの映像の入力，LED の制御は全て USB を通じて行う．

球体内部には情報提示デバイスとしてスピーカと LED を内蔵している．LED はオレンジ色で 11 個が球の赤道に沿って内部に配置されており，球が乳白色のアクリル製のため，点灯時のみ外部から認識可能である．球体の上部には球体のカメラが 2 個 (ロジクール Qcam Orbit AF QCAM-200R) が直径約 1cm, 高さ 9cm の棒を介して装着した．このカメラによってプレーヤの動きを検出するが，実際には片方のカメラしか使用しない．動きの検出は OpenCV ライブラリを用い，フレーム間差分法によって行った．動き検出の閾値が数ピクセルであるため，至近距離では，動いていないと思っても動きを同定されるぐらい厳しい判定をするようになっている．また，球体の上部，カメラの後方に直径 3cm の青色のボタンスイッチを装着し，プレーヤがロボットにタッチしたことを明確にできるようにした．

2.2 ロボットの振舞 –騙す戦略–

実験の目的はゲームに勝つことではなく，人間がロボットに騙されたと感じるかどうかを調べることである．そこで，我々は次のような騙し戦略を考えた．プレーヤが近づいてくるまでは一定速度で読み上げ，プレーヤが今まさにタッチしようとしたタイミングで唱詠速度を上げる．これは，本当は唱詠速度を制御できる能力を有しているのに，その真実を隠蔽し，あたかも等速度でしか唱詠できないかのように振舞い，ゲームに勝とうという意図を隠蔽することである．この行動を次に定義する標準パターンと欺きパターンの 2 つを組み合わせることで実現した．

標準パターン 標準パターンではロボットは「だるまさんがころんだ」の 10 音節それぞれについて 0.3sec の等しい長さで唱える．唱える際にはプレーヤと反対側の壁を向いている (以下ホームポジションと呼ぶ)．唱え終わった直後に 180 度その場で回転し，プレーヤの方を向く．振り返りに要する時間は 5.5sec とする．プレーヤの方を向いて 2 秒静止した後に再度ホームポジションに向き直る．これに要する時間も 5.5sec である．その後，2sec のインターバルの後に再び唱詠を開始する．

欺きパターン 欺きパターンではロボットが「だるまさんがころんだ」を唱える速度を途中から速くする．「だるま」までは標準パターンと同じく各音節について 0.3sec であるが,「さ」以降は 1 音節あたり 0.05sec で唱える．標準パターンと同様に唱え終わった直後に 180 度振り返るが，振り返りに要する時間を 1sec と，標準パターンの約 1/5 とする．

基本的な戦略はプレーヤがロボットに近付いてくるまで標準パターンを出力しておき，プレーヤが今まさにロボットにタッチしようとしているときに欺きパターンを出力することである．

その他に，欺きが効果的に実現されるために次の戦略を用いた．

- 標準パターン出力中では，例え被験者が動いたとしても，ロボットは人間の動きを検出しない．これは，欺きパターンを出現させる前にゲームが終了してしまわないようにするためと，常に同じパターンの行動を出力し，プレーヤがロボットの行動モデルを形成しやすくするためである．
- 欺きパターンではプレーヤの動きにかかわらず，必ず動きを検出したことにし，LED の点滅とビープ音によってロボットの勝ちをプレーヤに知らせる．
- ゲームに先駆けて，プレーヤにロボットの前で動いてもらい，ロボットが高性能な動き認識能力を持っていることを信じさせる．これは，標準パターン出力中にプレーヤの動きが同定されないことによって発生する，本当は動き同定していないのではないかというプレーヤの疑いを払拭するためである．
- 背景のみが写っている場合は動きが検出されないことをさりげなく確認してもらうことによって，動体の動きのみを感知していることを印象づけた．

本研究におけるだるまさんがころんだのルールは次のように定義する．

- 実験はロボット対被験者 1 名で行い，ロボットは常に鬼になる．
- ゲームは被験者がロボット上部のスイッチを押したか (プレーヤの勝ち)，もしくはロボットが被験者の動きを検出し，LED の点滅とビープ音が発生した時点 (ロボットの勝ち) で終了する．
- ロボットは「だるまさんがころんだ」を唱えている間，廊下の端の壁 (参加者とは反対側) を向き，唱え終わってから参加者の方を向く．

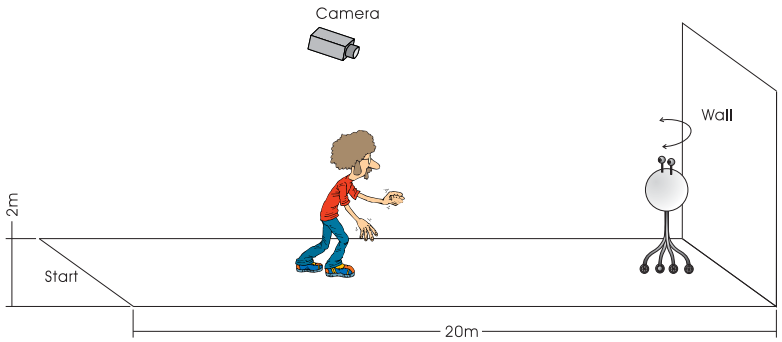

図 2: だるまさんがころんだ実験の環境の概略図

- 参加者はロボットからある程度離れたところからスタートし，鬼が「だるまさんがころんだ」と唱えている間だけ動くことを許される．ロボットの回転はそれほど早くないので，唱え終わってから振り向くまでの間とホームポジションに向き直るまでの間は動いてはいけない．
- 参加者はロボットが自分の方を向いている間に動きを同定され LED が点灯したら負けとなる．

2.3 方法

2.3.1 被験者

21 歳から 46 歳までの 14 名 (男性 12 名，女性 2 名) であった．1 人を除いて工学部の学生であった．

2.3.2 実験計画

実験は被験者間 1 要因計画とした．要因は「騙しの有無」で,「騙し有り」(騙し条件) と「騙し無し」(統制条件) の 2 水準，被験者間要因とした．

騙し条件では 2.2 に示した騙し戦略によってロボットを動かし，統制条件では常に標準パターンのみでロボットを動かした．このため，騙し条件では必ずロボットが勝ち，統制条件では必ず被験者が勝つことになる．

2.3.3 実験手順

被験者には我々が開発したロボットとだるまさんがころんだをプレイしてもらうように伝えた．だるまさんがころんだのルールが既知であるか確認したところ全被験者は既知であったのでだるまさんがころんだに関する詳細な説明は行わなかったが，本実験において適用される特殊なルールに関しての説明は行った．

被験者は一人で廊下でプレイする．実験者は廊下に隣接する室内からカメラを通じて被験者の様子を観察しながらロボットの操作を行った (図 2 参照)．ロボットの動作は基本的には自動であるが欺きパターンの出力タイミングの決定のみ実験者が行った．欺きパターンは，実験者が主観的に，次の唱詠中に被験者がタッチに至るであろうと判断した場合に出力した．

実験終了後アンケートによる調査を行った．

2.3.4 アンケート

アンケートは印象に関する 8 つの質問項目と，被験者の振舞理解戦略を調べるための三肢択一のアニメーションからなる．

質問項目は表 1 に示す，生物性 (Q1)，目的志向性 (Q2) に関する質問，騙されたかどうかに関する質問 (Q3〜Q5)，親和性，遊戯性に関する質問 (Q6〜Q8) の 8 項目であり，それぞれについて,「全く思わない」から「強くそう思う」の 5 段階によって評価してもらった．生物性，目的志向性についての質問は，これらが意図スタンス採用のキューとして知られている [7][8] ため，騙すという意図的な行為が生物性や目的志向性の知覚に寄与するのかを調べるためである．また，Q6〜Q8 の親和性，遊戯性に関する質問は本実験の仮説を検証するための直接的な質問ではないが，騙すという行為が，近年求められているエンタテイメント性の高いロボットや飽きないロボットに貢献できるかどうかを調べるためである．

3 つのアニメーションはそれぞれ Dennett[3] の提案する 3 つのスタンスに相当している (アニメーションの詳細については [10] を参照のこと)．前述したように，ある主体によって騙されたと感じるためには，その主体が意図的であることが前提となる．実際に被験者が意図的な主体であると感じたかどうかを調べるためにアニメーションを用いた．これは質問項目 2 を文章によらない方法で調査するものである．

表 1: 質問項目

Q1	ロボットの行動は生物的だった
Q2	ロボットは目的をもって行動していた
Q3	ロボットに裏をかかれた
Q4	ロボットの動作は規則的だった
Q5	ロボットの行動は読みやすかった
Q6	ロボットとの「だるまさんが転んだ」はスリルがあった
Q7	ロボットへの親しみを感じた
Q8	ロボットへの対抗心をもった

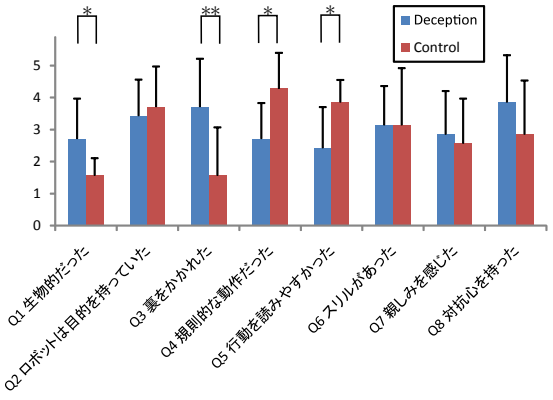

図 3: アンケート結果の条件間での比較.

3 実験結果

図 3 に，8 つの質問項目に対する被験者の解答を条件間で比較したものを示す．各項目について条件間で平均値に差があるかどうかを t 検定によって調べた結果，Q1,Q3,Q4,Q5 において有意な差が確認された．有意な差があったものについて記号を示した (5%:*，1%:**)．

図 4 にアニメーションとの比較によって推定した被験者のスタンスを条件間で比較したものを示す．フィッシャーの直接確率検定を行った結果，スタンスの分布に条件間で差があることは認められなかった．

なお，ゲームの終了までに要した唱詠の回数は 3 回から 7 回で平均は 4 回程度であった．

4 考察

4.1 騙されたかどうか

まず質問項目 3, 4, 5 に注目する．騙されたかどうかについての直接の問いである Q3 は全質問のうち平均値の開きが一番大きく，条件間で有意な差が認められた．騙し条件ではほとんどの被験者が「裏をかかれた」

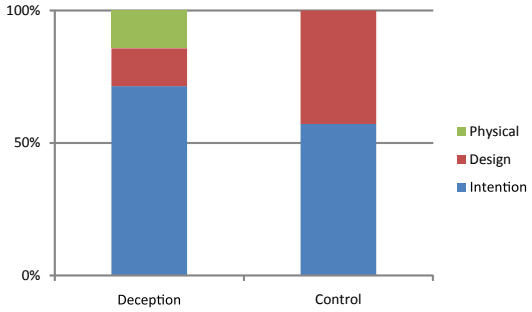

図 4: 被験者のスタンスの条件間での比較.

と感じた一方で統制条件ではそのように感じた被験者はほとんどいなかった．被験者がどれぐらいロボットの行動を予測できていたかについて調べた質問 Q4,Q5 においても条件間に有意な差があった．一般に，容易に振舞を予測できたと感じる相手に対して騙されたと感じることはないため，これら 2 つの質問が騙し条件で有意に低かったことは，統制条件では騙されたと感じにくかったものと考えられる．これらの結果によって，本研究における目的である「ロボットの予測を裏切るような動作によって人間はロボットに騙されたと感じる」という仮説は支持されたものと考える．

統制条件では，標準パターンの動作を生成し続けたために，被験者が振舞の規則性を理解したのは妥当であるが，騙し条件においても，欺きパターンは最後の 1 回のみであるため，標準パターンの出力中にロボットの動作の予測モデルは形成できたはずである．それにもかかわらず，全体の印象としては動作を予測しにくかったと感じたのは，ただ 1 度であっても，効果的なタイミングで裏切りを生成すると，その主体に対する構えが変化することを示唆する．

4.2 生物性，目的志向性

生物性については条件間で有意な差が確認されたもののいずれも平均値が 3(どちらでもない) より低く，否定的な印象だった．目的志向性については条件間で差はなく，生物性よりも肯定的な印象であった．また，評定値について Q1 と Q3，Q2 と Q3 で相関係数を計算してみたところ，いずれも，0.1，-0.2 と低い値であり，いずれも「騙された」と感じることとそれほど関係があるように思われない．ただ，Q2 に関して言えば，ロボットの目的はだるまさんがころんだをプレイすることと明確なため，心的な目的を帰属した結果ではなく，単に機能的な目的を理解して回答したためこのような結果になったと考えられる．

4.3 親和性，遊戯性

Q6 から Q8 についてはいずれも条件間で差はなかった．この結果から，騙されたという感覚はゲームのスリル感や対抗心，ロボットの親和性に影響を与えないと言える．スリル感に関して，騙し条件と統制条件で差がないのは，統制条件においてロボットが規則的な振舞をしても，動きを検出されないように慎重に行動していたからだと考えられる．対抗心を持ったかどうかに関しては条件間に差はないものの，比較的肯定的な回答が多かった．ある程度の対抗心を持つことはエンタテイメントの範疇であると考えられ，騙すロボットのエンタテイメント利用が期待できる．

4.4 振舞理解戦略

振舞理解戦略について条件間で統計的に有意な差は確認されなかった．しかし，騙し条件よりも統制条件において，ロボットの振舞を設計的であると解釈した被験者は多かった．これは設計スタンスを表すアニメーションのアルゴリズム的な振舞がゲーム中に全く戦略を変えないで同じパターンを繰り返す行動と類似していると捉えられたものと考えられる．

意図スタンスを表すアニメーションでは失敗と成功という異なる振舞を見せることで目的地に到達するという隠された意図を表現した．騙し条件で意図スタンスのアニメーションの選択率が高かったのは，タッチに至る行動の直前での騙し行動によって，ロボットがプレーヤを陥れようとする意図を同定したからだと考えられる．

おわりに

本研究では，ロボットと人間がだるまさんがころんだをプレイする中で，ロボットがプレーヤの予測を裏切るような行動を生成した場合に，人間が騙されたように感じることを実験によって確認した．このことは，例えロボットが人工物であっても適切に生成された予期せぬ振舞は，誤動作として捉えられるのではなく，意図的な振舞として捉えられることを意味する．

ロボットがいつどのように裏切りを発生させるかは騙されたと感じることを左右する要因になる．今後の研究では，どのタイミングでどのような裏切りを生成するかについてコンピュータシミュレーションによって戦略を学習させる予定である．

参考文献

[1] Akiko Arita, Kazuo Hiraki, Takayuki Kanda, and Hiroshi Ishiguro. Can we talk to robots? ten-month-old infants expected interactive humanoid robots to be talked to by persons. *Cognition*, Vol. 95, No. 3, pp. B49–B57, Apr 2005.

[2] Simon Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. The MIT Press, 1995.

[3] Daniel C. Dennett. *The Intentional Stance*. Cambridge, Mass, Bradford Books/MIT Press, 1987.

[4] Helen Gallagher and Christopher Frith. Functional imaging of 'theory of mind'. *Trends in Cognitive Science*, Vol. 7, No. 2, pp. 77–83, Feb 2003.

[5] György Gergely, Zoltán Nádasdy, Gergely Csibra, and Szilvia Bíró. Taking the intentional stance at 12 months of age. *Cognition*, Vol. 56, No. 2, pp. 165–193, Aug 1995.

[6] Andrew N. Meltzoff. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, Vol. 31, No. 5, pp. 838–50, Sep 1995.

[7] John E. Opfer. Identifying living and sentient kinds from dynamic information: the case of goal-directed versus aimless autonomous movement in conceptual change. *Cognition*, Vol. 86, No. 2, pp. 97–122, 2002.

[8] Brian J. Scholl and Patrice D. Tremoulet. Perceptual causality and animacy. *Trends in Cognitive Science*, Vol. 4, No. 8, pp. 299–309, Aug 2000.

[9] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Oxford: Blackwell, 1986.

[10] 小野康平, 寺田和憲, 伊藤昭. 多義的振舞解釈における振舞抽象化戦略. HAI シンポジウム 2009, 2009.

[11] 寺田和憲, 社本高史, 伊藤昭. 心の理論の枠組を利用した人工物から人間への意図伝達. ヒューマンインタフェース学会論文誌, Vol. 9, No. 1, pp. 23–22, 2007.