2D2-OS5b-2

人を騙すHAI

HAI for Deceiving People

寺田　和憲　　伊藤　昭
Kazunori Terada　　Akira Ito

岐阜大学
Gifu University

This paper addresses the relation among intention attribution, unexpectedness and deception. Unexpectedness indicates that a current strategy for understanding an observing behavior is no longer useful, and as a result, it trigges intention attribution to rationally understand the behavior. Deception is a class of behavior, which is realized by noticing an unexpectedness in the behavior, and cannot be understood without attributing intentions.

1. はじめに

本稿のタイトルは「人を騙す HAI」であるが，本稿の目的はロボット（エージェント）が人間を騙すためのアルゴリズムや倫理的問題を論じることがではない．本稿では，機械であるロボットに騙されたと感じることが，人間がロボットのことを人間的な（心を持った）存在としてみなしているという証拠になり，騙しのような意外性のある振舞の実装が HAI を考える上で重要であることを論じる．

ロボットは機械である．しかし，機械を超越した存在であってほしいという願望を人々は持つ．その願望にはロボットに心が宿って欲しいというものが含まれている．意識を持つことや記憶を持つことなど，心の定義は様々であるが，ここではひとまず心理学で心的状態として定義される願望，信念，意図を心と考え，その中で特に意図 (intention) を持っているように見えるロボットを作ることを考える．ロボットが本当に心を持っていることと心を持っているように見えることは別の問題である．ロボットが真に意図を持ち得るかどうかと意図そのものが何であるかについて議論することは本稿の範囲外と考え，意図を持っているように見えることにのみ焦点を当てる．なお本稿では，目的地や目標物のような物理的実体が心的に表象されたものを意図と呼ぶ．

対象が意図を持っているように見えるのは観察者がその対象に対して意図を帰属 (intention attribution) するからである．これは，意図が実際に観察対象の内部に存在するかどうかに関係なく，また，存在したとしても，真の意図が何であろうと観察者が恣意的に想定するものであることを意味する．心を持っているようなロボットを作るためには観察者としての人間が心を持っていると感じる要因をロボットに実装すればよい．外見は一つの要因であるが，本研究では振舞，特に人間とロボットのインタラクションにおける振舞の性質について考える．それでは，人間はどのような振舞の性質を持つ対象を意図的主体として捉え，機械として捉えるのだろうか．その答えの糸口は，機械-人間という概念的分類が振舞予測の観点から有効だということにある．

振舞予測戦略の違いから意図的主体と機械の違いを考える．機械に対しては我々は機能的観点から理解しようとする．Dennett はこの戦略を設計スタンスと呼んでいる [Dennett 87]．対象の機能的理解には目的論的 (teleological) 解釈と法則による解釈の二つの側面がある．目的論的解釈とは，その対象物が何かの目的のために設計されたもの，もしくは何かの目的のために使用されるというように，設計者や使用者を想定した上で目的を帰属するというものである．設計された人工物のもう一つの側面は，それがメカニズムやアルゴリズムなど，入出力関係の恒常性を保つ特定の法則の上に成り立っているということである．椅子の「座るためのもの」という解釈は目的論的解釈であり，「上に載せた物体の高さを常に一定に保つ」という解釈は物理法則によるものである．前者の解釈では，着座という行為の最終（目標）状態が注目され，後者の場合は，同じ入力に対して常に同じ出力を担保するという法則性が注目されている．この二つの視点に基づいて「機能」を再定義すると，「法則性の目的論的説明（解釈）」であり，設計スタンスは対象に法則と目的の両方を帰属するものだと言える．

機能的理解は自然に存在する対象についてもなされる．例えば，適当な高さの直方型の岩は椅子と同様の物理的恒常性を持っており，行為者を想定することで座るという機能を帰属することができる．また，多くの生物の振舞は設計スタンスによって理解し予測できる．本能として埋め込まれた反射的振舞は，入出力関係を規定したルールとその目的（観察者が恣意的に帰属するものではあるが）によって理解することができる．ただしこの場合は，創造主を想定する場合 [Kelemen 05] を除いて，目的は主体そのものに帰属される．

人間の振舞は目的論的解釈が可能である．しかし，人間の振舞を設計スタンスの法則的側面によって理解しようとすると破綻する．なぜなら，人間の振舞を決定している神経基盤に入出力関係の恒常性を担保する規則性が存在しない，もしくはそのような規則が存在し，人間の振舞が決定論的に解釈可能だとしても，入出力関係を規定する規則が多様すぎて知覚者の認知能力を上回るからである．そこで使われる戦略が心的状態を仮定するという方法である．

2. 意図スタンスと意外性

願望，信念，意図などの心的状態を帰属し (mental state attribution)，他者の振舞の理解と予測を行う戦略を意図スタンス (intentional stance) と呼ぶ．意図スタンスの採用を可能にしているのは，心の理論 (theory of mind) と呼ばれる認知的枠組みである．近年，その認知的枠組みを実現している神経基盤が明らかにされつつある [Gallagher 03]．意図を推測している際に上側頭溝や前部傍帯状皮質などの特定の部位が賦活す

連絡先: 寺田和憲，岐阜大学工学部，〒 501-1193 岐阜市柳戸 1-1，058-293-2792，terada@info.gifu-u.ac.jp

ることが知られており，それら心の理論関連部位の賦活が意図スタンスを採用していることの客観的指標となっている．

2.1 設計スタンスとの違い

設計スタンスと意図スタンスはどちらも目的を帰属するという意味では同じであるが，目的を帰属するのが観察対象そのものであるか，観察対象の設計者や使用者であるかという点で異なる．別の言い方をすると，その違いは「目的が振舞を駆動する」か「法則によって駆動された振舞が特定の目的のためになっている」かの違いである．

設計スタンスでは設計者や操作者を想定した上で目的を帰属する．行動主体の目的を想定しなければ設計スタンスを定義できないので，設計スタンスの基礎になっているのが意図スタンスだという考えがある．Kelemen は，幼児が「山は登るためのもの」,「雲は雨を降らせるためのもの」などのように，様々な主体に対して偏執的に目的論的解釈をしてしまうことと，創造主としての意図的主体を想定することの間に関連があることを主張し [Kelemen 05]，設計スタンスが意図スタンスから分化すると述べている [Kelemen 07]．これは発達心理学の知見からも妥当なものである．近年の研究によって 4 から 6 歳ぐらいの間に設計スタンス的理解をするようになると言われているが，これは意図スタンスの萌芽よりも随分遅い．そこで，Kelemen は次のような 3 段階を考えた．子供は，まず最初に人工物が意図的な行為の終端に関連していることを理解する．そして，人工物がその目的を体現するものとしての理解がなされる．そして最終的に，それほど明確に意図を想定しなくても目的そのものだけを直接理解可能な設計スタンスが確立するというものである．

このように意図スタンスと設計スタンスは関連があるが，二つのスタンスの決定的な違いは振舞の原因を帰属する際に法則性を想定するか意図を想定するかの違いである．設計スタンスでも意図は想定するが，それは振舞を規定するための法則性を導出するための意図にすぎない．別の言い方をすると，意図スタンスを採用すべき対象は意図に基づいて振舞を自由に変化させることができるのに対して，設計スタンスを採用すべき対象は振舞が法則性に縛られているということである．

2.2 意図帰属を引き起こすきっかけ

意図スタンスを用いることの利点は，一見すると脈絡のない複雑な振舞が，意図を帰属することによって単一の概念として抽象化されることである．そして，一旦意図を帰属してしまえば，その対象が将来や別の状況でどのような振舞をするかを予測可能なことである [Dennett 87][Gergely 95]．帰属する意図は観察者が恣意的に推測するものではあるが，観察対象が合理的な目的遂行者であることを仮定すれば，状況が異なっても振舞の予測が可能であるため，この戦略は頑健である．また，意図スタンスは，意図を恣意的に帰属することで，冗談や嘘，皮肉といった表面的な振舞と心的状態の乖離を解決し，それらを合理的に解釈することも可能にする．

ロボットは機械であるが，数多くの研究結果によって人間がロボットに対して意図スタンスを採用する可能性があることが示唆されている．コンピュータ科学の分野では，コンピュータの社会的反応に対して人間が対人的反応を示すことはよく知られている [Reeves 98]．また，コンピュータスクリーン上の単純な図形の運動を観察した人間が，図形に対して人間的属性を帰属することが知られている [Heider 44][Dittrich 94][Bassili 76]．

人間がロボットに対して意図帰属をするかどうかを調べた研究がある．Wang らはヒューマノイドロボットの頭部の動き方が意図帰属に影響を与えることを示した [Wang 06]．また，fMRI などを用いて脳内の心の理論関連部位の活動を観察することによって調べた研究がある [Krach 08]．この研究では，異なる外観を有するロボットと囚人のジレンマゲームを対戦中の被験者の脳活動を fMRI を用いて調べることによって人間らしい外観になるほど心の理論ネットワークの賦活が大きくなることが示されている [Krach 08]．また，発達心理学の分野において，6.5ヶ月児がヒューマノイドロボットに対して目的志向性を帰属することを示した実験 [Kamewari 05]，Meltzoff の [Meltzoff 95] 実験パラダイムをロボットに適用し，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た研究 [Itakura 08] がある．

人間がコンピュータや図形の振舞に対して人間的属性を帰属できることは，人間らしさを感じるために必ずしも外見が人間に類似している必要はなく，振舞やインタラクションの様式がそれらしいものであればよいことを意味する．そのような観点から，認知科学の分野では，人間的属性，特に意図を帰属するために必要な振舞の属性を調べた多くの研究がある．それらの研究では，外見 [Krach 08]，自己推進性 [Baron-Cohen 95][Premack 94][Heider 44]，随伴性 [Bassili 76]，目的志向性 [Dittrich 94] などが意図帰属の原因として挙げられている．また，Heider[Heider 58] は人間的な振舞を特徴づける本質的な性質として，異なる状況においても同じ終極に到達するという等終極性の重要性を主張した．例えば，落石などの物理的な現象は，環境が異なると終極状態が異なるのに対し，目的を持った人間は状況が異なっても別の手段を用いて同一終極に到達することができるというものである．複数の行為系列が同一の終端に到達するのを観測することによって，観察対象の振舞を駆動している意図が同定できるために等終極性は意図の推定に寄与する．

Dennett や Gergely らは意図帰属の目的が振舞の予測であるという観点から，合理性を前提とすることが心的状態の帰属に重要であると主張している [Dennett 87][Gergely 95][Csibra 07][Gergely 02]．この主張は次の論理に基づいている．行為者は欲求に基づいて行動を開始し，信念 (何を知覚し記憶しているか) に基づいて行動を計画する．その際に，損失するエネルギーを最小にする，であるとか最短時間，最短経路で到達することいった合理性が考慮される．観察者が行為者の振舞を理解するためには行為者の欲求，信念，意図を知ることが必要不可欠であるが，観察者が直接知覚できるのは表面的な振舞だけなので，それらの心的状態は何らかの方法によって推測するしかない．しかし，振舞から意図を同定することは不良設定問題なので，様々な制約を想定した上でもっともらしいものを想定するしかない．このときに合理性を仮定すると，帰属するべき意図の探索範囲を限定できるために，推測が容易になる．さらに合理性は等終極的結果の観測と組み合わせることによって意図推定を容易にする．しかし，合理性がより重要な役割を果たすのは，将来の振舞を予測するときである．等終極性は意図の同定に貢献するが，異なる状況においてそのエージェントがどのように振舞うかを予測することはできない．しかし，状況認識を共有しており，合理性に基づいた最適な行動を選択するという前提があれば，行為者の行動と観察者の予測が一致する可能性が高くなる．合理的振舞が目的 (意図) と制約の関数として生成されるため，目的と制約が既知であれば生成される振舞がほぼ一意に定まるからである．

2.3 意外性

前説で挙げた属性はいずれも意図帰属を引き起こすきっかけや意図を推測するための根拠になり得るが，それらの属性が

観測されたからと言って必ずしも意図スタンスを採用しなければならないことはない．自律ロボットの多くは自己推進性や目的志向性など先に挙げた属性を備えている．特に Dennett と Gergely の主張する合理性に関して言えば，機械であるロボットは特定の目的を実現するための合理的手段を工学的に実現したものであるので振舞は合理的である．しかし，ロボットが合理性を備えているからと言ってロボットに意図を帰属する必要はない．なぜなら，ロボットの振舞は普通，設計スタンスによって，メカニズムやアルゴリズムなどの法則を想定することで理解がなされるからである．意図帰属の際に仮定される合理性は，帰属するべき意図の探索範囲限定のためのものであり，そもそも目的と振舞の関係が固定されている機械の振舞理解に，その原因となるような意図を帰属する必要はない．

この事実が意味することは，法則が帰属できてしまう限り意図を帰属する必要がなく，ロボットが機械を超越した存在になり得ないということである．では，ロボットに意図帰属をさせるためにはどうすればよいのだろうか．我々は**振舞に心を感じるのは心を感じる必要があるから**だと考える．そこで，意図帰属が必要もしくは有益に働く状況を考える．そのような状況とは振舞の多様性が観測される状況である．意図的主体の振舞の多様性は 1) 意図遷移の結果としての振舞の変化，2) 特定の意図がバリエーション豊富な振舞を生成すること，3) 意図と振舞が一致しない場合（嘘，皮肉など）があること，に起因している．これらの原因によって多様な振舞が生成されるので，その振舞の起源になっている意図を推定することは，単一のシンボルで抽象化することによる認知負荷の軽減と振舞予測の簡単化に貢献する．このことは，振舞の多様性が意図帰属による振舞理解戦略が有効であることを示す信号になっていることを意味する．そこで，振舞の多様性を示唆することができる要因を考える．

振舞の多様性は方策の変化から生まれる．振舞の起源が意図であるならば，方策の変化は意図そのものが遷移したか，同一の意図を実現するためのインスタンスとして方策が変化したかのいずれかである．観察者の視点からすると，後者の場合には適切な意図を帰属していれば，方策の変化を吸収できるので，帰属した意図を変更する必要はない．しかし，前者の場合と後者の場合であっても誤った意図を帰属していた場合には，正しい意図を帰属しなおさなければならない．誤った意図帰属は誤った予測につながるからである．そのために，帰属していた意図が観察対象の現在の振舞を説明するために適切なものであるかどうかを判定することは重要である．これは，予測と観測が一致しているかどうかを判断することによって行われる．予測と観測の不一致は意外という主観的概念を生成する．すなわち意外性は意図帰属を再度行わなければならないことを知らせるための見逃してはならないは重要な信号だと言える．

脳活動の観測によって予測の不一致が意図帰属と関係があることが示唆されている [Pelphrey 03][Saxe 04]．Saxe ら [Saxe 04] は予想していた重さと違った重さの箱を持ち上げる人の映像が，予想と一致した重さを持ち上げる人の映像よりも STS の有意な賦活を引き起こすことを示した．また，Pelphrey ら [Pelphrey 03] は，映像中の歩行する人間が衝立の後ろを通過する際に，しばらく止まって現れない場合に，普通に通過する場合よりも STS の有意な賦活を引き起こすことを示した．STS は心の理論に関連した領域と言われているので，これらの研究は予測との不一致が意図帰属に何らかの形で関連している可能性を示唆している．しかし，意図帰属との関係を明示的に確認しているわけではない．

意外性は設計スタンスによっても解釈できる．設計スタンスでは対象の振舞を法則によって理解しているので，意外性を感じることは帰属していた法則が誤っていたかその法則が崩れたことを意味する．後者の場合は事故か故障である．しかし，いずれの原因によっても意外性を説明できない場合には意図帰属による理解を行うものと思われる．

3. 意外性と騙し

前章の議論から，「例えロボットであって意外な振舞をした場合には意図帰属が行われることがある」という仮説を導くことができる．冗談や皮肉，嘘，騙しなどは意外性が含まれ，意図帰属を行わない限り理解できない振舞である．これらの振舞では表面的な振舞と意図が乖離しており，表面的な振舞から直接推測できる意図とは異なる意図を帰属しない限り振舞の合理的解釈はできない．そして，それらの振舞の真の意図が認識されるきっかけは意外性である．

ここで，騙しに注目して意外性と意図帰属の関係について考える．騙しは特定の振舞のクラスを説明する概念で，その定義は意図的に相手の錯誤を引き起こすことである．騙しが成立するのは心的状態が外部から観測不可能だからである．行為者の視点から見ると，「騙そう」という意図のもとに生成された一連の振舞は一貫している．しかし，観察者は行為者の表面的な振舞に誘導されて誤った意図を帰属してしまう．騙されたことの認識は帰属している意図と振舞の乖離に気付くことで起こる．例えば，詐欺にあったことに気付くのは，代金を支払ったのに商品が届かないなど，通常の商取引の手順を逸脱した事象の発生である．このように，錯誤の認識が発生するのは相手の行為に不整合が発生し，それまでの解釈が意味をなさなくなるときである．この不整合は観察者に意外性を感じさせる．以上のことから，意外性が騙し認識のきっかけになると言える．

上述のような振舞の不整合は設計スタンスによっても解釈できる．意図スタンスでは騙しと解釈されるが，設計スタンスでは，事故や故障（相手の認知的故障，送金や物流システムの故障）であると解釈される．いずれの場合もこの不整合をうまく説明できるが，騙しであるという解釈は意図帰属を行わなければ発生しないため，ある振舞に対して騙されたと感じることはその振舞に対して意図帰属を行ったという証拠になる．

意図的な騙しは生物の中でも人間と高等霊長類にしかできない特殊な振舞である [Byrne 95][Hare 06]．騙しを実現，認識できるのは心の理論を持つからであり．心の理論を持たないとされる自閉症者は騙しの認識ができないと言われている．例えば，自閉症者は他者が達成しようとしていることを物理的な妨害によって阻止することはできるが，騙すことによって阻止することはできない [Sodian 92]．また，自閉症者は，騙しなどの心的状態への反応が含まれたアニメーションの説明することを求められたとき，健常者よりも正確に記述することはできなかった [Abell 00]．Castelli らは fMRI を用いて Heider らと同様のアニメーションを使って実験を行った [Castelli 00]．その中で，アニメーション中の対象が騙し，説得，嘲り，驚かすなどの振舞を行った場合に目的指向的振舞やランダムな振舞よりも心の理論関連部位が賦活することを示した．

意外性がロボットに対する意図帰属を引き起こすことを明示的に調べた研究は我々の知る限り存在しない．しかし，意外性に明示的に言及していないものの，騙されたことの認識が意図帰属の証拠となるという動機のもとに，人間がロボットに騙されたと感じるかどうかを調べた研究がある [Short 10]．Short らは上半身ヒューマノイドロボットによるじゃんけんタスクを取り上げた．ただ，彼女らの研究における騙しはどちらかと言

うと，不正やごまかし (cheat) に属するものである．彼女らは不正なし（統制），口頭不正，動作不正の 3 条件によって実験を行った．口頭不正条件では，ロボットはじゃんけんの手を出した後，負けていたとしても「勝った」と宣言する．動作不正条件では負けていたとしても，物理的に勝ちの手に変形させ，「勝った」と宣言する．彼女らの仮説は口頭不正は故障（認識エラー）として認識され，動作不正は騙しとして認識されるというものであり，実験によってそれらは支持された．また，口頭，動作の両不正条件で不正なし条件よりも心的状態の帰属が有意に高かったことが確認された．

4. まとめ

本稿では認知科学の知見を引用しながら，次のような論理を展開した．人間は対象の振舞の理解と予測のために設計スタンスと意図スタンスを使い分けている．設計スタンスでは，設計者の意図とともに機能を実現している法則を推測する．意図スタンスでは，振舞を駆動している意図を推測する．ロボットの振舞は通常設計スタンスによって理解されるが，ロボットが心を持っているように見えるためにはロボットに対して意図帰属ができる必要がある．これまでに様々な意図帰属の原因が提案されているが自律ロボットはそれらを満たしている．そこで意図帰属の必要性と有用性を再考し，意図帰属を行わなければならない状況を考え，振舞の多様性が観察されたときに意図帰属の必要性が発生することを導いた．振舞の多様性の認識は予測と観測の不一致から発生する意外性が知覚されたときなので，意外な振舞は意図帰属を誘発する原因になり得る．そしてこれがロボットにも適用されると考え，「ロボットが意外な振舞をしたときに意図帰属が行われる」という仮説を導いた．また，意外性が含まれ，意図帰属を行わなければ理解できない振舞の一つとして騙しに注目し，騙しの認識においても意外性がきっかけとなることを説明した．

参考文献

[Abell 00] Abell, F., Happe, F., and Frith, U.: Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development, *Cognitive Development*, Vol. 15, pp. 1–16 (2000)

[Baron-Cohen 95] Baron-Cohen, S.: *Mindblindness: An Essay on Autism and Theory of Mind*, The MIT Press (1995)

[Baron-cohen 03] Baron-cohen, S.: *The Essential Difference: The Truth about the Male and Female Brain*, Basic Books (2003)

[Bassili 76] Bassili, J. N.: Temporal and Spatial Contingencies in the Perception of Social Events, *Journal of Personality and Social Psychology*, Vol. 33, No. 6, pp. 680–685 (1976)

[Byrne 95] Byrne, R. W.: *The Thinking Ape: Evolutionary Origins of Intelligence*, Oxford University Press (1995)

[Castelli 00] Castelli, F., Happé, F., Frith, U., and Frith, C.: Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns, *Neuroimage*, Vol. 12, pp. 314–325 (2000)

[Csibra 07] Csibra, G. and Gergely, G.: 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans, *Acta Psychologica*, Vol. 124, pp. 60–78 (2007)

[Dennett 87] Dennett, D. C.: *The Intentional Stance*, Cambridge, Mass, Bradford Books/MIT Press (1987)

[Dittrich 94] Dittrich, W. H. and Lea, S. E. G.: Visual perception of intentional motion, *Perception*, Vol. 23, No. 3, pp. 253–268 (1994)

[Fukuda 10] Fukuda, H. and Ueda, K.: Interaction with a Moving Object Affects One's Perception of Its Animacy, *International Journal of Social Robotics*, Vol. 2, No. 2, pp. 187–193 (2010)

[Gallagher 03] Gallagher, H. and Frith, C.: Functional imaging of 'theory of mind', *Trends in Cognitive Sciences*, Vol. 7, No. 2, pp. 77–83 (2003)

[Gergely 95] Gergely, G., Nádasdy, Z., Csibra, G., and Bíró, S.: Taking the intentional stance at 12 months of age, *Cognition*, Vol. 56, No. 2, pp. 165–193 (1995)

[Gergely 02] Gergely, G., Bekkering, H., and Király, I.: Rational imitation in preverbal infants., *Nature*, Vol. 415, No. 6873, p. 755 (2002)

[Hare 06] Hare, B., Call, J., and Tomasello, M.: Chimpanzees deceive a human competitor by hiding, *Cognition*, Vol. 101, No. 3, pp. 495–514 (2006)

[Heider 44] Heider, F. and Simmel, M.: AN EXPERIMENTAL STUDY OF APPARENT BEHAVIOR, *The American Journal of Psychology*, Vol. 57, No. 2, pp. 243–259 (1944)

[Heider 58] Heider, F.: *The Psychology of Interpersonal Relations*, Lawrence Erlbaum Associates (1958)

[Itakura 08] Itakura, S., Ishida, H., Kanda, T., Shimada, Y., Ishiguro, H., and Lee, K.: How to Build an Intentional Android: Infants' Imitation of a Robot's Goal-Directed Actions, *Infancy*, Vol. 13, No. 5, pp. 519 – 532 (2008)

[Kamewari 05] Kamewari, K., Kato, M., Kanda, T., Ishiguro, H., and Hiraki, K.: Six-and-a-Half-Month-Old Children Positively Attribute Goals to Human Action and to Humanoid-Robot Motion, *Cognitive Development*, Vol. 20, No. 2, pp. 303–320 (2005)

[Kelemen 05] Kelemen, D. and DiYanni, C.: Intuitions About Origins: Purpose and Intelligent Design in Children's Reasoning About Nature, *Journal of Cognition and Development*, Vol. 6, No. 1, pp. 3 –31 (2005)

[Kelemen 07] Kelemen, D. and Carey, S.: The Essence of Artifacts: Developing the Design Stance, in *Creations of the mind: Theories of artifacts and their representation*, Oxford University Press (2007)

[Krach 08] Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., and Kircher, T.: Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI, *PLoS ONE*, Vol. 3, p. e2597 (2008)

[Meltzoff 95] Meltzoff, A. N.: Understanding the Intentions of Others: Re-enactment of Intended Acts by 18-Month-Old Children, *Developmental Psychology*, Vol. 31, No. 5, pp. 838–50 (1995)

[Pelphrey 03] Pelphrey, K. A., Singerman, J. D., Allison, T., and McCarthy, G.: Brain activation evoked by perception of gaze shifts: the influence of context, *Neuropsychologia*, Vol. 41, pp. 156–170 (2003)

[Premack 94] Premack, D. and Premack, A. J.: Moral belief: Form versus content, in *Mapping the mind: Domain specificity in cognition and culture*, pp. 149–168, Cambridge: Cambridge University Press (1994)

[Reeves 98] Reeves, B. and Nass, C.: *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, CSLI Publications (1998)

[Saxe 04] Saxe, R., Xiao, D.-K., Kovacs, G., Perrett, D. I., and Kanwisher, N.: A region of right posterior superior temporal sulcus responds to observed intentional actions., *Neuropsychologia*, Vol. 42, No. 11, pp. 1435–1446 (2004)

[Short 10] Short, E., Hart, J., Vu, M., and Scassellati, B.: No fair!!: an interaction with a cheating robot, in *HRI '10: Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pp. 219–226, New York, NY, USA (2010), ACM

[Sodian 92] Sodian, B. and Frith, U.: Deception and sabotage in autistic, retarded and normal children, *Journal of child psychology and psychiatry, and allied disciplines*, Vol. 33, No. 3, pp. 591–605 (1992)

[Wang 06] Wang, E., Lignos, C., Vatsal, A., and Scassellati, B.: Effects of Head Movement on Perceptions of Humanoid Robot Behavior, in *HRI '06: Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 180–185 (2006)