3O2-OS-3b-8

定型性と逸脱性に基づくインタラクションのモデル化

An Interaction Model based on fixedness and deviation

寺田 和憲　　伊藤 昭
Kazunori Terada　　Akira Ito

岐阜大学工学部
Faculty of Engineering, Gifu University

This paper describes an interaction model to account for a difference between humans and machines in terms of fixedness and deviation of agent's behavior. Firstly, we argue inappropriateness of categorization of Dennett's three stances, and propose four new category of stances. Then we define one axis consists of two primary concept for classifying entities into human and machine: fixedness-deviation axis. Previous researches such as animacy recognition, interaction between child and robot, escaping behavior of pill bug, turing test with minimum modalities, and deception are explained by using fixedness-deviation axis.

1. はじめに

哲学者 Dennett は人が対象の振舞を理解し予測する際に物理スタンス，設計スタンス，意図スタンスという 3 つのスタンスを使い分けると考えた [Dennett 87]．物理スタンスとは主体の物理的組成，物理的性質，物理法則に基づいて振舞いを予測する戦略である．設計スタンスとは物理スタンスで想定される物理的組成などの細部を無視し，主体が設計意図に基づいて作られていることを前提として，様々な状況において設計どおりに振る舞うと予測する戦略である．意図スタンスとは主体の振舞いが意図，信念，願望などの心的状態、に基づいて合理的に生成されているという前提のもとに，振舞いの起源となる心的状態を帰属した上で，振舞いを予測する戦略である．ロボットは機械なので設計スタンスを採用するべきなのだろうか．もしそうであるならロボットは永遠に機械のままで人に近づけないことになる．本稿では，Dennett のスタンスに対する反論を出発点として人と機械を分離する特徴軸について論じ，ロボットが機械を超越した存在になるための必要条件として逸脱性を導く．また，定型性-逸脱性の軸を用いて主体の振舞いやインタラクション様式を説明することを試みる．

2. 人と機械を分離する軸

2.1 Dennett のスタンスに対する反論

Dennett は人（意図的主体），機械，物理現象をカテゴリーとして分類したが，それらを分類する軸が何であるかを明確に述べているわけではない．Dennett の 3 つのカテゴリー化は対象の振舞いを予測する戦略を決定する際に用いられる．戦略がどのように働くかを説明しているがそれを分離する軸については議論していない．本章では Dennett の 3 つの分類の問題点を指摘し，定型性と逸脱性による分類を導く．

Dennett の設計スタンスと物理スタンスの違いは法則の観点からみると粒度の違いでしかない．電子レンジの振舞いを理解するのに電磁波がいかにして水分子を振動させ熱となってあらわれるかについて理解する必要はなく，時間設定のダイヤルと温度の対応さえ理解しておけばよい．また，時計の振舞いを理解するためにはエネルギー源であるバネや振り子，電池の物理的挙動や歯車の摩擦について理解する必要はなく，3 本の針の動きを理解しておけばよい．時計の針の動きは定型性を保っている．その定型性が人為的に作られたものであったとしても，定型的であるという意味では，物体の落下や熱による気体の膨張といった物理的振舞いと同じである．時計の場合も物体の落下も特定の入力に対して特定の条件で同一の出力が保障されている．設計された人工物における定型性という観点は Dennett の論考には出てこない．しかし我々は，定型性こそがインタラクション対象を定義する軸の一つの極として重要だと考える．

Dennett の設計スタンスと物理スタンスの違いは法則の粒度の大小だけでなく，設計者の意図の有無という観点からも説明できる．Dennett の設計スタンスは設計者の意図に基づいて設計通りに振舞うことを予測するというものである．時計の長針が 12 周する間に短針が 1 周するという法則は，時計の設計者がバネや振り子，電池のエネルギー出力をメカニズムやアルゴリズムを利用してうまく制御した結果表出されるのである．設計者が存在しない限り，そのような法則が顕在化することはない．その意味で設計スタンスと物理スタンスは異なる．ただ，アンティキティラ島で発見された機械のように，機械であることは分かってもそれが何であるかが長年分からなかった物体の例や，古いアイロンをブックエンドとして活用することもあるため [Dennett 90]，設計者の意図が分からなくても第 3 者が帰属可能であるという意味で，その振舞いが生成している機能（定型性の目的論的解釈）は曖昧なものである．さらに，意図の判明しない人工物は多数ある．芸術作品もその一つである．芸術作品は制作者がなんらかの意図を持って作ったとしても，その意図は必ずしも正しく解釈されない．むしろ，芸術は鑑賞者が主観的に解釈するものである．また，ルーブ・ゴールドバーグ・マシン（ピタゴラ装置）は人為的に作られた物であり，入出力の恒常性を有しているが目的が存在しない．その意味で我々は，設計スタンスと物理スタンスを分離する観点として設計者や解釈者の存在を仮定することは不適切だと考える．

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

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

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，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]．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定型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定型性を導出するための意図にすぎない．

2.2 スタンスの再定義

上述のように Dennett のスタンスは直感的には正しく思えるが，詳細に検討すると不適切な点があるため，我々は 3 つのスタンスを認知科学的手法を用いて再定義することを試みた [寺田 12]．実験では，3 つのスタンスを典型的に表現した 3 つのアニメーションを被験者に提示し，アニメーションに対する被験者の印象記述を分析することで，振舞い理解のための 4 つの言語的概念カテゴリーを明らかにした．次に，60 個の様々な対象の振舞いを参照基準として用いることで，4 つのうち 3 つの言語的概念カテゴリーが Dennett のスタンスと近いものであることが分かった．意図スタンスに相当する概念を構成する語句は，意識的，考えている，判断している，努力している，能動的，臨機応変などがある．設計スタンスに相当する概念を構成する語句は，予め決められている，正確，規則的，法則に従っている，アルゴリズムに従っているなどである．これらは，設計という意図的な行為によって形成される振舞いというより，法則などによって定型性を有する主体の振舞いの性質を表現しているため，このような振舞いを予測する戦略を**定型スタンス**と呼ぶ．物理スタンスに相当する概念を構成する語句は，自然な，ありのままの，身を任せているなどである．これらは，それ自身に振舞いを決定する原因や法則を持たず，外部要因によって決定する振舞いの性質であるため，このような振舞いを予測する戦略を**受動スタンス**と呼ぶ．

4 つ目の概念カテゴリーは Dennett のスタンスとは独立したものと考えられる．この概念を構成する語句は，複雑，非単調，変化がある，予測できないなどである．Dennett は振舞い予測という観点から論理を展開しているのにもかからわず，予測可能性について明示的に触れているわけではない．我々が発見した 4 つ目の概念は，予測可能性に直接関わる概念であり，3 つのスタンスに相当する概念とは明らかに独立している．我々はこのような振舞いを予測する戦略を**複雑スタンス**と呼ぶ．複雑スタンスは予測できない複雑な振舞いに対して用いる心的な構えで，振舞いを理解，予測できないものとして捉え，理解も予測も放棄する．予測困難さや複雑さを発生する原因としては，偶然や乱数，カオス力学，意図が存在するが，複雑スタンスではそれらの原因について言及するものではない．

2.3 人と機械を分ける軸

本稿の議論の目的は人と機械を分離する本質的な軸を導出すことである．前節では，認知科学的手法によるスタンスの再検討の結果，4 つの概念カテゴリーを導かれたことを述べた．4 つの概念カテゴリーは反意概念を想定することでそれぞれを一つの軸として考えることができる．本節ではそれらの軸によって人と機械を分離可能かどうかについて考える．

まず，受動性について考える．受動の反意語は能動なので，能動－受動という軸が人と機械の分離に適用できるかについて考える．自己推進性 [Premack 94] や重力法則違反 [Gelman 95] などのように，運動の起源が外部にあることが意図性知覚のトリガーだという説がある．そのため，物理現象のように，意図的でない主体の振舞いを決定づける性質として受動性は妥当なものに思える．しかし，機械には受動的なものも能動的なものも両方存在する．そのため，機械も人も能動的という同一のカテゴリーに存在することになる．さらに，能動性や受動性は振舞いの起源に言及するものであって振舞いそのものの性質ではない．従って，能動－受動という軸は人と機械の分離において用いることはできない．

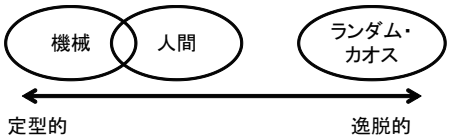

図 1: 振舞いの定型性と逸脱性による対象の分類

次に，意図性について考える．意図的の反意概念は非意図的であるが，それが具体的に何を示すのか定義が困難である．実際に，意図スタンスに関する概念は意識的，考えている，判断している，努力している，能動的，臨機応変などの語句によって形成されていた．これらのうち，能動的以外は抽象度が高く，それら自体を科学的に定義することが困難な語句である．従って，意図性を人-機械の分離軸として用いることは不適切である．

次に，定型性について考える．定型の反意語は非定型なので，定型－非定型という軸を考える．前述のように，機械の振舞いは定型性を有しているため，この軸の定型性の極は機械の振舞いを適切に表現している．では非定型性は人の振舞いを適切にあらわすだろうか．意図的主体の振舞いの性質として多様性がある．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．この多様性は非定型性とも言えるため，定型－非定型という軸は人と機械を切り分ける軸として適切だと思われる．

4 つ目の概念カテゴリーである複雑性について考える．複雑の反意語は単純である．単純－複雑は定型－非定型に対する主観的評価であると言える．複雑な機械は多数存在しているが，単純－複雑は機械と人間を切り分ける軸としても適切だと思われる．従って本稿では，{ 定型, 単純 }-{ 非定型，複雑 } のような軸を人と機械を切り分ける軸として導入することを考える．

3. 定型性と逸脱性によるモデル化

前章の議論に基づいて，我々は図 1 のような定型性と逸脱性に基づく対象の分類を考える．この図では概念的に定型性と逸脱性を一次元の軸上で表し，軸の一方の極に定型性，もう一方に逸脱性を配置している．機械は定型性が強い領域に位置し，人やエージェントは定型性と逸脱性の両方を備えた領域に位置する．物理現象は定型的側面と逸脱的側面の両方を備えている．

右の極を非定型や複雑ではなく逸脱という言葉に置き換えた理由は次の通りである．振舞いを非定型性にする原因は偶発性，ノイズ，意図など多様であるが，定型性を全く持たない振舞いはランダムである．しかし，定型でかつ複雑なものは存在する．複雑さは一般的には多数の要素多様が絡み合い様々な様

態が出現することと定義される．計算機科学において，有限長のデータ列の複雑さを表す指標のひとつにコルモゴロフ複雑性がある．コルモゴロフ複雑性は，出力結果がそのデータに一致するプログラムの長さの最小値として定義される．ここでの複雑性は複雑性が高くてもその系列がアルゴリズムによって生成されるため，定型性を持っていると考えられる．また，我々の考える逸脱性はカオスにおける複雑性とは少し異なる．前述したように，逸脱とは定型性が前提となっている．しかし，カオスの複雑な振舞いは予測できないだけで決定論的法則に従っている．

逸脱は一般的に平均的な基準からの偏向という統計的性質と，本筋や規則から逸れることの両方の意味で用いられる．本稿では，主に後者の意味として用いる．逸脱は規則から逸れることを意味するため，規則の存在を前提としている．規則には，行為や手続きなどを行う際の標準となるように定められた事柄，きまりという意味と法則，秩序という意味がある．これらには，守るべきという主観的な意味が含意されている．我々は，周期運動や単なる入出力関係のような意味のない振舞いの性質も軸の左の極に含まれると考えるため，左の極を表現する言葉として，規則性や法則性ではなく，より抽象度の高い定型性という言葉を用いる．

心理学において，心を帰属できるような対象の属性として，自己推進性，目的志向性，随伴性，合理性，物理法則違反など多くの要因が提案されているが，それらを備えてもなおロボットは心を持っているようには見えない．我々は，この原因を**法則的理解による人らしさの消滅**と考える．例えば，随伴的な振舞いをするロボットはインタラクションの初期では人らしさを感じるかもしれない．しかし，随伴性を特定のセンサ入力に対する動作出力というアルゴリズムで法則化してしまうと，とたんに人らしさは感じられなくなってしまう．この現象は，振舞い予測の観点から，心を感じることの必然性を考えると説明できる．意図帰属の目的は心的状態の帰属による振舞いの予測にある．従って，より簡単な法則的理解が可能な場合には，そもそも意図帰属の必要がない．一方で，定型性がないランダムやカオス的な振舞いに対しても意図性を感じられない．これは，予測することが不可能であるため，振舞い予測のための心的状態の帰属が意味をなさないからである．

4. 定型-逸脱モデルによる現象の説明

本章では，これまでの研究報告に対して定型性と逸脱性に基づくモデル化による説明を試みることで，定型性-逸脱性の軸が機械と人間を区分する適切な軸であることを示す．

4.1 アニマシー知覚

Fukuda らは，人とインタラクションするロボットの反応様式に 1/f ゆらぎを加えると，よりアニマシーが知覚されることを示した [Fukuda 10]．1/f ゆらぎに従う振る舞いでは適度な規則性と規則からの逸脱性が混在している．逸脱性から感じられる意外性が意図を持った主体としてのアニマシー知覚に寄与したとも考えられる．

4.2 子供とロボットのインタラクション

高橋らはロボットと子供の関係性を新奇性と親近性という 2 つの軸でモデル化した [高橋 11]．新奇性のある対象とは，子どもにとって珍しい普段あまり接したことが無い対象を指し，親近性のある対象とは，子どもが慣れ親しんで愛着や安心を感じる対象を指す．新奇性に基づく興味は時間とともに減衰し，親近性に基づく興味は時間とともに増加するため，高橋らは親近性を高めることがロボットとのインタラクションの継続に繋がると考えている．新奇性に基づく興味は Dennett の設計スタンスに相当するものと言うことも可能である．設計された対象の定型性はその対象と初めて対峙したときには分からない．そして，人は定型性を理解しようとして，入出力関係を確かめる行動を起こす [寺田 07]．定型性に対する探索的行動は，Baron-Cohen がシステム脳と呼ぶ男性に典型的な特質 [Baron-Cohen 04] によって引き起こされると考えられる．定型性が理解され既知のものになれば，その対象と関わる理由が消失し，興味が失われる．高橋らの提案する親近性は愛着や安心という感情的反応に主眼が置かれているが，Dennett の意図スタンスに近いものだと考えられる．前述のように我々は，意図スタンスの本質が逸脱性にあると考えている．親近性を感じる対象に対して興味が持続し，インタラクションが継続可能な理由は，意図的な主体の振舞いが適度に定型性を逸脱するため，逸脱の都度新規性が高まるからだと言うことができる．愛着や安心という感情的反応は行動主体が観察対象に接近した状態を保つという機能がある．この機能を発現させているトリガーとして観察対象の振舞いから知覚される逸脱性を仮定することはそれほど不自然ではない．

4.3 ダンゴムシの脱出行動

逸脱的振舞いを生成できる能力は意図的主体や知性を定義づけるための本質的能力のひとつではないかと考える．振舞を遺伝子によって規定された生物はオンラインで方策を変化させることができない．例えばバクテリアの捕食や鮭の遡上は特定の刺激に対する固定的反応である．しかし，固定した方策しか持たないことは変化する環境や競合状態においては弱点となる．特に，絶えず競争にさらされ，相手を凌駕しなければ生き延びられない状況では，予測を上回るもしくは予測できない振舞を生成できること，すなわち固定的振舞を逸脱できることが有効である．そのための戦略の一つとして生物が採用しているのが突然変異である．突然変異は振舞を遺伝子によって規定された生物が方策を変化させられる機会であるが，その機会は自己複製時などに限られている．そのために，突然変異をしてもなお個体内で方策は固定され，依然として競合と適応の観点から脆弱である．この弱点は個体内で方策を変化させることによって克服される．この能力は試行錯誤を可能にし，単一の目的に対して異なる解を見つける可能性を高める．

ヒトなど高い知性を持つとされている種はこの能力を有しているが，昆虫においても固体内での方策変化が発現することが知られている．森山は未知の環境における予想外の行動の発現を心の存在の証と考え，ダンゴムシを用いた様々な実験を行った [森山 11]．森山によると，ダンゴムシの行動の定型性を利用し，どんな行動をとっても行き止まりに遭遇するような環境を人工的に作り閉塞状態に陥らせると，壁を登って環境を脱出するという逸脱的行動が観察されるという．森山はこのような現象を発現させている仕組みを心と考えている．我々は振舞い認知の観点から逸脱性を意図帰属の原因として考えているが，動物行動学における心の定義と認知科学における心の定義が一致することは興味深い．

4.4 ミニマムチューリングテスト

利用可能なコミュニケーションモダリティを最小限に限定した環境を用いたチューリングテストにおいて，相手が人であることを確かめるために，定型性と逸脱性が用いられるという報告がある [竹内 05, 飯塚 12]．竹内らの実験では実験参加者は電子ドラムを通じて別室にいる他者とインタラクションすることが求められた．インタラクションの初期では，一方が特定のリズムパターンを作り出し，他方がそれに追従する模倣が見ら

れる．しかし，それは長く続かず，主導者と追従者が入れ替わる現象が見られた．この現象は原初的コミュニケーションプロトコルの生成プロセスとしても興味深いが，人的（意図的）な振舞いとされる模倣が” 模倣するアルゴリズム” として捉えられると，もはや人的な振舞いとして感じられなくなり，ターンテイキングという別の振舞いパターンによる検査が行われたと説明することができる．飯塚らも，遠隔状態で指先のみがインタラクションできる装置を用いて類似の実験を行い，ターンの形成からターンテイキングへの発展という定型パターンの生成，逸脱，さらに別の定型パターンの生成という過程が発生することを確かめた．

4.5 騙し

寺田らは人間とロボットの間の騙しが定型性の逸脱によって発生すると考えた [寺田 11]．狼少年が信用を失うように，常に嘘ばかり言っていると信用されなくなる．すなわち，騙しが成立するためには，通常は本当のことを言うなどして相手を信用させなければならない（正直者というモデル化を誘導する）．一方で，常に正直に振舞っていると，競合状態においては，容易に搾取される．従って，騙しが成立する条件は，正直な（定型的）行動と裏切り（逸脱的）行動の割合という統計的なモデル化が可能である．そのような観点から，寺田らはだるまさんがころんだを題材として，ロボットが競合ゲームにおいて定型的に振舞うよりも逸脱的な振舞いをした場合に，実験参加者がロボットのことを意図的な存在であると認識したという実験を報告している．

同様に，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．ロボットが実際にはじゃんけんに負けているにも関わらず,「勝った」と宣言するごまかしをした場合に，参加者が心的状態を帰属してロボットの振舞いを説明するという結果が得られている．

5. まとめ

本稿では，人と機械を分離する特徴軸として定型性と逸脱性を両極とする軸を提案した．導出のために，まず，Dennett の提案する 3 つのスタンスが対象を適切に分類しているかどうかについて検討し，我々が過去に行った認知科学的実験の結果に基づいてスタンスを再定義した．新たなスタンスは意図スタンス，定型スタンス，受動スタンス，複雑スタンスである．これらのスタンスを構成する概念が人と機械を分離する軸になり得るかについて議論し，定型性と逸脱性という軸が妥当であることを導いた．さらに，この軸を用いてアニマシー知覚，子どもとロボットのインタラクション，ダンゴムシの脱出行動，ミニマムチューリングテスト，騙しという現象の説明を試みた．

参考文献

[Baron-Cohen 04] Baron-Cohen, S.: *The Essential Difference*, Penguin Books (2004)

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

[Dennett 90] Dennett, D. C.: The Interpretation of Texts, People and Other Artifacts, *Philosophy and phenomenological research*, Vol. 50, pp. 177–194 (1990)

[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)

[Gelman 95] Gelman, R., Durgin, F., and Kaufman, L.: Distinguishing between animates and inanimates: not by motion alone, in Sperber, D., Premack, D., and Premack, A. J. eds., *Causal cognition : a multidisciplinary debate*, chapter 6, pp. 150–184, Oxford University Press (1995)

[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*, pp. 415–449, Oxford University Press (2007)

[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)

[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

[高橋 11] 高橋 英之, 宮崎 岡田 浩之, 大森 隆司：「新奇性」と「親近性」の軸から子どもとロボットの関係性を捉える, HAI シンポジウム 2011, pp. I–2B–2 (2011)

[寺田 07] 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)

[寺田 11] 寺田 和憲, 伊藤 昭：人間はロボットに騙されるか?–ロボットの意外な振舞は意図帰属の原因となる–, 日本ロボット学会誌, Vol. 29, No. 5, pp. 43–52 (2011)

[寺田 12] 寺田 和憲, 岩瀬 寛, 伊藤 昭：Dennett の論考による 3 つのスタンスの検証, 電子情報通信学会論文誌 (A), Vol. J95-A, No. 1, pp. 117–127 (2012)

[森山 11] 森山 徹：ダンゴムシに心はあるのか, PHP 研究所 (2011)

[竹内 05] 竹内 勇剛, 杉江 舞子：踊るエージェントを通した相互作用による対人認知過程, 情報処理学会研究報告 (知能と複雑系), No. 109, pp. 9–14 (2005)

[飯塚 12] 飯塚 博幸, 安藤 英由樹, 前田 太郎：身体的相互作用におけるコミュニケーションとターンテイキングの創発, 電子情報通信学会論文誌 (A), Vol. J95-A, pp. 165–174 (2012)