認知科学 第 29 巻 第 3 号 (2022) pp. 371–389 https://doi.org/10.11225/cs.2022.036
Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, Vol. 29, No. 3, pp. 371–389

特集　合理性をめぐる認知科学　　招待論文

合理性と創造性の接点：多重目標と二重過程から意識的自己へ

服部 雅史 *

立命館大学

Rationality meets creativity: Multiple goals, dual processes, and the conscious self

Masasi Hattori*

Ritsumeikan University

This paper reviews and revisits the concept of rationality in the psychology of thinking. First, I consider the ambiguity of the concept of rationality. I point out that this ambiguity is due to (1) the indeterminacy of the normative system itself; (2) differences in the way the task, solver, and environment are perceived; (3) differences in viewpoints such as the theoretical and practical; and (4) the duality of cognitive processes. However, I show that rationality is a goal-dependent concept, and that such ambiguity can mostly be sorted out by the notion of conflict among multiple goals. Next, based on recent findings on reasoning and judgment in autism spectrum disorders, I point out that previous research required "clipped-out" thinking, which is assumed to be rational. Such thinking is non-creative as the goal is a predetermined one given from the outside of the target system. However, since such thinking can deviate greatly from what is rational in the ordinary sense, I point out that an aspect of creativity is essential to the concept of rationality in the event. Finally, I argue that a well-being perspective is indispensable for rational and creative thinking, and that the concepts of self and consciousness are indispensable for acquiring such a perspective.

Keywords: normativeness（規範性），adaptivity（適応性），autism spectrum disorder（自閉スペクトラム症），well-being（ウェルビーイング），consciousness（意識），

Received 4 April 2022

1. はじめに

「人間は合理的か」という問いは，おそらく人類の文化の始まりと同じくらい古くから，数え切れないほど多くの人々によって発されてきたにもかかわらず，今でも次々に新しい議論が積み重ねられている (e.g., Knauff & Spohn, 2021a; Viale, 2021)．この問いには，問うことをやめられない魅力がある．それは，この問いが，人間の認知や知性，また人間そのものを理解するために避けて通れないものであるからであろう．しかし，議論が尽きないのには，合理性という言葉があいまいであることも無関係ではない．合理性という概念は，個人間で，また個人内でも文脈によって，その解釈に大きな変動がある．本論文では，合理性という概念がなぜ，どのようにあいまいであるのか，また，その多義性が人間の思考や認知の研究にとってどのような意味があるのかを明らかにする．認知科学，特に思考心理学の見地から，近年の研究成果も踏まえながら，冒頭の問いに対する回答を模索する．人間の認知や知性の本性に少しでも近づくことを願いながら，合理性と意識性，また創造性との関係について検討し，認知科学研究の今後の方向性について考えたい．

2. 規範と解釈

人間が合理的であるかどうかに答えるためには，

本稿は，服部 (2017, 2018) で提示したアイデアの一部を，その後の複数の講演等を通して深めたものである．

* E-mail: hat@lt.ritsumei.ac.jp

合理性とはどういう特性であるかを明らかにする必要がある．そこで，まずは合理性にもっとも深い関係があると考えられる規範（norm）という概念からみていくこととする．

2.1 規範という物差し

思考心理学のもっとも初期の頃の実験研究は，三段論法（syllogism）を用いて行われた．Wilkins (1928) は，三段論法課題に日常的な材料と抽象的な材料を使った実験により，課題の内容によって成績が異なること，また「すべての X は Y である」という文は「すべての Y は X である」のように誤って解釈される傾向があることを明らかにした [1)]．論理学的正解と実験参加者の回答を比較するというやり方，すなわち，規範を物差しにして人間の行動を評価するという方法は，実験心理学において極めて頻繁に用いられる．たとえば，錯視研究において，対象物の客観的な測度（物差し＝規範）がなければ錯視という現象を同定することすら困難であるのと同様，思考についても，規範がなければ思考という現象の同定が難しくなる．規範とは，ここでは客観的に正しいもの，特に思考に関しては，主に数学などの形式的システムを指す．

規範という物差しを前提とした思考研究のアプローチは，人間の非合理性を顕著化させた．一連の「ヒューリスティックスとバイアス」（heuristics and bias; HB）研究 (e.g., Manktelow, 2012 服部・山監訳 2015) によって，比較的簡単な問題に対する人々の回答が，規範から大きく，しかも一定の方向にずれることを示す経験的事実が蓄積されるにつれて，合理性のパラドックス (rationality paradox; Evans & Over, 1996 山訳 2000) が認識されるようになった．すなわち，人間は優れた言語システムをもち，論理学を構築するような高度な知能をもつのに，実験室ではなぜこんなにも愚かな誤りをするのかという疑問である．人間の非合理性を強調する心理学に対して，哲学者 Cohen (1981) は，「心理学実験で人間の非合理性を示すことができるか？」という疑問形の挑発的タイトルの論文の中で，「否」と断じて人間の合理性を擁護した．Cohen のように人間は本質的には合理的であるする立場を，Stanovich (1999) はパングロス主義（Panglossian）と呼んだ．パングロス主義者の論点を Evans (1993) は，規範系の問題，解釈の問題，外的妥当性の問題の三つに分類した．このうち，3 番目の外的妥当性の問題 [2)] は，重要ではあるが本論との直接的な関係はわかりにくいので省略し，以下では，合理性の多義性の観点から，第 1, 2 番目の論点，およびそれと密接に関係する問題について順に論じる．

2.2 規範は一つではない

Cohen (1981) は，素朴な直感に一致しない規範理論は人々の合理性を判断するための枠組みとしては不適切であるとした (p. 317)．この考えに基づけば，人間は常に合理的であり，その行動が規範に一致しない場合は，規範理論の方を考え直すべきということになる．この主張が正しいかどうかは別として，与えられた課題の規範理論が必ずしも唯一に決まるわけではないことには留意が必要である．このことが，合理性の多義性の一因であると考えられる．以下に二つの例を示す．

推論の心理学でよく使われるウェイソン選択課題 (Wason, 1966) では，片面にアルファベット，もう片面に数字が書かれた 4 枚のカード（A, K, 4, 7）が机の上に置かれ，実験参加者は「カードの片面が母音ならば，その裏面は偶数である」という規則が正しいかどうかを確かめるために，どれを裏返すべきかを答える．正解は A（母音）と 7（奇数）であるが，多くの参加者は 7 の代わりに 4（偶数）を選ぶ．この課題で想定されている規範はいわゆる古典論理であり，K. Popper の反証主義哲学である [3)]．しかし，この課題は，Bayes 的な仮説更新課題として定式化して参加者が期待獲得情報量を最大化すると仮定することが可能であり (Oaksford & Chater, 1994)，その場合の正解は，多くの参加者の回答に一致する

1) その後，前者は主題材料効果 (thematic materials effect; e.g., Wason & Shapiro, 1971) や信念バイアス (belief bias; e.g., Evans et al., 1983) などの内容効果に関する研究に，後者は不正換位 (illicit conversion; Chapman & Chapman, 1959) から確率表象モデル (probabilistic representation model; Hattori, 2016) への対称性推論（symmetry inference）に関する研究に展開し，現在もこの研究の流れは継続している．

2) これは，たとえば，人間の仮説検証過程を調べるための 2-4-6 問題 (Wason, 1960) といった課題は人工的すぎるため，そこでの誤りは人間の非合理性を示すものではないといった論点である．ただ，人工的課題に対する不適切にみえる反応は，ある（不適切な）解釈がなされた結果と捉えることも可能であり，そうするとこの問題は解釈の問題に統合されると考えられる．

3) 母音を V，偶数を E とするとき，$\forall x[V(x) \rightarrow E(x)]$ という仮説の反証テストとしてカード a が $V(a) \wedge \neg E(a)$ であることを確認する必要があるため，母音かつ奇数の事例を求めて 4 と 7 を裏返す．

A と 4 である．

もう一つの例は，タクシー問題 (Tversky & Kahneman, 1982) と呼ばれる有名な確率判断課題である．この課題は，街に緑と青の 2 色のタクシーが走っており，ひき逃げ事故の目撃者がタクシーは青色だったと言ったが，証人は 8 割の確率で正しい色を答える（2 割の確率で逆の色を答える）とき，青色タクシーが事故を起こした確率を答えるというものである．この課題では，「街中のタクシーの 85% が緑色で残りの 15% が青色」という基準率（base rate）情報が与えられるため，それを考慮して Bayes の定理を使って計算すると正解は 41% となるが，多くの参加者は 80% 前後と回答する．ここで想定されている規範は Bayes の定理であるが，この課題は，信号検出理論を用いて定式化して参加者が正しい同定を最大化すると仮定することが可能であり (Birnbaum, 1983)，その場合は，多くの参加者の回答に近い 82% が正解となる．

Einhorn & Hogarth (1981) が正しく指摘したように，規範理論が適切かどうかは仮説次第である (p. 55)．つまり，課題の捉え方次第で規範は変わりうる．さらに 7 節で論ずるように，課題の捉え方は，課題自体だけでなく，課題に取り組む主体，また課題と主体を取り巻く環境をどう考えるかによっても変わる．

2.3 解釈の任意性

パングロス主義者の 2 番目の論点，解釈の問題とは，端的にいえば，実験課題を解く参加者が実験者が意図するように課題を解釈するとは限らないということである．たとえば，Tversky & Kahneman (1983) が示したある問題では，参加者は，ある人物（リンダ）がフェミニスト（A）のステレオタイプであるかのように記述された文章を読むと，その人物が，銀行員である（B）よりフェミニストの銀行員（$A \wedge B$）である可能性が高いと判断してしまう．連言の確率 $P(A \wedge B)$ が単一事象の確率 $P(B)$ を超えることは論理的に不可能なので，これは連言錯誤（conjunction fallacy）と呼ばれる．しかし，もし文脈から「銀行員」が「フェミニストでない銀行員（$\neg A \wedge B$）」をさすと解釈可能なのであれば，これはエラーとは言えない (Smedslund, 1990)．課題文が正しく解釈されれば，参加者は規範的に解答できた可能性がある (ただし，Moro, 2009 も参照)．

ところが，解釈の問題は見かけほど容易に規範系の問題と区別できるわけではない．まず，規範が複数あるということは，参加者側と実験者側で規範が異なりうることにもなるので，解釈の問題との区別が難しくなる（3 節）．また，解釈の問題は，もう少し微妙な目標多重性（4 節）という別の問題とも関わるため，やはり規範との区別が困難になる（7 節）．それぞれ改めて議論したい．

以上のように，規範性には不定性が存在するが，規範性は間違いなく合理性の一部であるため，必然的に合理性にも不定性が持ち込まれることになる．これが合理性の多義性の一因である．しかし，合理性には別の種類の多義性・曖昧性も存在する．以下ではそれをみていきたい．

3. 実践，資源，進化

3.1 実践的合理性

合理性の概念が一つでないことは，古くから認識されてきた．理論的合理性（theoretical rationatliy）と実践的合理性（practical rationality）の区別はギリシャ時代に遡る (Manktelow, 2004)．この区別について，哲学者 Harman (1995) は次のような例を挙げている (p. 175)．たとえば，「ジェーンは，重要なテストの前日，テストでよい結果を出すために今夜勉強しなければならないことがわかっていて，友人のパーティに行けば本当に後悔することもわかっていながら，パーティに行ってしまう」という場合，本当にするべきことを正しく理解しているという意味で，ジェーンの信念は（理論的に）合理的であるが，行為の方が（実践的に）非合理的である．別の例として，「ボブは，テストの採点は匿名化して実施されていることを知りながら，自分の点が悪かったのは採点者の人種差別のせいだと決めつける」という場合，ボブは信念の方が（理論的に）非合理的である．

両者の概念的な区別はこれでよいとして，心理学の問題はここからである．ジェーンの信念は本当に合理的であろうか，また行為は本当に非合理的であろうか．ジェーンは，（本人が意識しているかどうかはわからないが）明らかに二つの目標を持っている．一つはテストでの好成績，もう一つはパーティでの歓楽（または友情を深めること）である．しかも，これらは同時に達成できない．第 1 の目標を達成するためには勉強する（パーティに行かない）と

いう行為が合理的であるが，第 2 の目標を達成するためにはパーティに行く（勉強をしない）という行為が合理的である．それなのに第 1 の目標の達成が合理的で，第 2 の目標の達成が非合理的であるとされるのはなぜなのか．それは，第 1 の目標の方が本人に強く意識されているからであろうか．しかし，意識されている目標を達成することの方が，意識されていない目標を達成することよりも，それが意識されているという理由だけで合理的であることの根拠とするべきであろうか．この問題は，本稿の最後（9 節）で議論するが，ここではまず，意識的自己と無意識的自己をあわせて考えると，理論的合理性と実践的合理性の区別が見かけほど明快ではなくなる点を指摘しておきたい．

もう一つの論点となるのは，パーティに出かけてしまったジェーンは，本当に後悔するであろうかという点である．するかもしれない．しかし，後悔したからといって，その行動が合理的ではなかったといえるであろうか．つまり，後悔（の予期）は合理性（あるいは幸せ）の決定要因として適切であろうか．この問題も 9 節で議論する．

3.2 限定合理性・適応的合理性・生態学的合理性

実践的合理性の「実践」の概念の中に認知資源の観点を明確に持ち込んだのが H. A. Simon である．新古典派経済学の「（合理的）経済人」（homo economicus）は，決定理論の最適解に基づく決定をするが，実際の人間はこの姿からは程遠い．

> 【限定合理性（bounded rationality）の原理】
> 複雑な問題を定式化・解決するための人間の心の能力は，現実世界において客観的に合理的な行動（あるいはその適切な近似）をするために解くべき問題の大きさに比べて非常に小さい．(Simon, 1957, p. 198, 筆者抄訳)

このように，Simon (1957) は合理性の限界を人間の能力の問題としたが，人間だけの話ではない．現実世界において問題解決をしようとすれば，計算の各ステップに，どんなに短くてもゼロより大きい時間がかかる限りは，ステップが膨大ならば，（記憶容量の問題がなくても）実用的時間内に計算が終了しない．限定合理性の概念は，理論的な整合性ではなく，環境の中で限られた資源しか持たない行為者が，より適切に行動することを問題にする．

Anderson (1990) の適応的合理性（adaptive rationality）の概念もこの観点を共有している．適応的合理性の一般原理は，「認知システムは常に生体の行動の適応を最適化するように作動する」(p. 28) こととされている．この最適化は，生物の進化によって実現される．また，この原理に基づいて，認知システムの目標と環境の形式モデルとから導かれる最適行動として生体の行動が予測できるというのが合理分析（rational analysis）の基本的な考えである．2.2 節で触れた期待獲得情報量最大化の規範 (Oaksford & Chater, 1994) も，この考えから導かれたものである．限定合理性に比べると強調点が環境側に移動しているものの，基本的な考えは同じである (Anderson, 1990, p. 31; ただし，Simon, 1991 の異論も参照)．また，Gigerenzer らが唱える生態学的合理性（ecological rationality）の概念も「現実との適合性」(Gigerenzer et al., 1999, p. 5) を定義としている．生態学的合理性の概念は，単純なヒューリスティックの使用（特にその単純さ）に強調点があるが，環境の中で知的に生きることを問題にしている点は同じである．

ここで取り上げた合理性（以降は総称として適応的合理性と呼び，Anderson の概念を狭義のものとする）は，環境に注目し，生体との相互作用の中での合理性を考える点が共通している．それは，生体内外の資源を考慮に入れ，生物の進化の観点も内包した実用的観点からの合理性の概念である．その点において，2 節で論じた規範的合理性や理論的合理性とは大きく異なる．つまり，合理性には，基準の異なる複数の概念がある．

3.3 合理性は目標依存

合理性に複数の概念が存在する理由について，さらに考えを進めるための素材として，ビュリダンのロバ（Buridan’s ass）の寓話を取り上げたい．道の分岐点に立っている空腹のロバが，道の先の等距離のところに同量の干草が置かれているときに，どちらにも行けずに餓死してしまうという話である．

この話は，どちらを選択しても同じなのであれば迷って選べないのは愚かであるという意味合いで引用されることが多い．しかし，文脈次第では別の解釈を考えることもできる．「同じ」というのが「ほぼ同じに見える」というだけで実は少しは違うかもしれない場合はどうであろうか．しかも，少しの違いが極めて重要であり，さらには，ロバが実はロ

ボットで餓死しないとしたらどうであろうか．たとえば，1000 年後の人類の移住先の候補として二つの地球外惑星を検討しているとき，どちらもほぼ同じに見えるという理由だけでいずれか一方に即決するのは，明らかに愚かな行為である．つまり，何が合理的な判断であるかという評価は，目標次第で変わる．この評価の不定性は，3.1 節で提示した隠れた目標を考えたときのジェーンの合理性の問題とも重なる．

規範的合理性と適応的合理性の違いも，目標の違いによって説明できる．もし干草に早くありつくことが主要な目標ならば，積極的理由がなくても悩まずに一方を選択するのが合理的であるが，もし干草の量を正解に見積もることが主要な目標ならば，積極的理由がないのに一方を選択するのは合理的ではない．規範的合理性（や理論的合理性）の基準は，考え（信念）の理論的・形式的整合性を保持することが目標であるときに採用される．認知資源や計算コストのことは考慮しない．一方，適応的合理性の基準は，時間や資源に制限がある環境の中で，できるだけ望ましい結果を得ることが目標であるときに採用される．その評価は結果次第であり，理論的な整合性は（あってもよいが）なくてもよい．

同様に，適応的合理性自体についても，目標が変われば，合理的であった行動がそうではなくなることもある．たとえば，求められること（目標）が，二つの都市のどちらの人口が多いかを直感で答えるというような（些末な）課題である場合，Goldstein & Gigerenzer (2002) が鮮やかに示したように，単に一方の都市名を見たことがあるというだけの理由でその都市を選ぶという単純な方法（再認ヒューリスティック recognition heuristic）は，環境に適合した合理的なものである．しかし，転職や結婚のような重大な意思決定の場合，この方法は使えない (Sternberg, 2000)．すなわち，素早い回答が求められているが解答の手がかりがほとんどなく，しかも間違えても害がないような状況で合理的とみなされる推論や方略や行動は，それとはまったく異なる状況においては合理的とみなされない可能性がある．単純なヒューリスティック（「適応的道具箱」の中の各道具）のよい面を強調する Gigerenzer らの議論は，環境への適合の観点においては緻密であるが，目標の違いの点では不十分である．道具箱の道具がどれも使えない場合や，あえて使わない方がよい場合についての議論も重要である．

さらには，規範（規範的合理性の基準となるはずのもの）が複数存在するときの規範の違いも，一部は目標の違いによって説明できる．ウェイソン選択課題は，課題を論理主義的に捉えるか Bayes 的に捉えるかによって規範解（正解）が変わると述べた（2.2 節）．この規範の移行は「課題状況の捉え方」の切り替えによるが，そのとき同時に「課題目標」も切り替わっている点を見逃してはならない．すなわち，目標が「反証事例による論理的な偽の例証」から「確証事例による仮説の確信度更新」に切り替わっている．この場合，課題の捉え方の違いと目標の違いは不可分である．

しかも，このケースは，パングロス主義者のいう「規範系の問題」と「解釈の問題」の境界（2.2 節）をもあいまいにする．1990 年代前半にウェイソン選択課題への確率論的アプローチ (Kirby, 1994; Oaksford & Chater, 1994) が現れるまでは，この課題への論理主義的規範の適用が疑われることはほとんどなかった．その頃は，実験参加者が（おそらく Bayes 的規範に基づいて）母音と偶数を選択した場合，本来は反証例を探すべきところ，「文が真であることを確かめようとする傾向を抑制できない」ために起こる「エラー」(Wason, 1966, p. 146) なのか，課題があまりに抽象的すぎたために起きた「単なる認識の失敗」(Cohen, 1981, p. 324) なのか，という解釈の問題の議論になった．しかし，単に実験者と参加者の採用する規範が異なっていたということであれば，解釈の問題というより規範系の問題であり，また，このケースでは目標も異なることから，結局は目標の問題に落ち着くことになる．

以上より，目標の違いによって何が合理的とみなされるかが変わることが明らかになった．つまり，目標の違いが合理性の多義性を生む．ただ，目標が一つに明確に固定しても，合理解（規範解）が一つに定まらないことはある．それは，規範が必ずしも一つではない（2.2 節）からである．しかし，合理解が複数存在すること（合理性の多義性）の理由の多くは，目標の違いによって説明できる．そこで，この次に考えるべきことは，目標が一つでない場合にどうするべきかという問題と，課題と環境を切り分けることの可否とその意味の問題である．次節で前者について考察する．後者については，別の問題とあわせて 7 節で考察する．

4. 目標多重性

ここまでの議論で目標の変動が合理性の多義性を生むことが明らかになったが，目標が複数存在する場合は，その多義性はさらに深刻なものとなる．3.1 節の例で示したジェーンは，二つの目標による葛藤に苦しんでいたことを指摘した．こうした事態は日常では珍しいことではない．それどころか，厳密な意味で目標が単一であるという状況は，まったく非現実的である．しかし，これまでの思考研究の多くは，単一目標を前提とした枠組みでなされており，その前提の不適切さが，人々の行動を不当に非合理的にみせていたということはないだろうか．以下では，その可能性について考えたい．

4.1 漏出と目標分断：通時的多重目標

意思決定研究では，実験室的課題がしばしば用いられる．すなわち，いわば「切り出された」状況における思考や行動が検討される．しかし，現実場面では，意思決定状況は生活の中に連続的に位置づいており，決定結果は，研究者が対象とする事象だけでなく，関連することがらや感情にも影響する．この影響が，決定行動における決定者の潜在的な目標を形成する．

Keys & Schwartz (2007) は，このことを，決定が経験に「漏出（leak）する」と表現した．たとえば，埋没費用錯誤 (sunk-cost fallacy; Thaler, 1980) として知られる現象がある．「ホテルの部屋で 6.95 ドルを払って映画を見たら 5 分で退屈になったとき，見続けるかやめるか」という意思決定課題で，全体の 62% の参加者が「見続ける」と回答したが，この中の大半（79%）は無料ならやめるとした (Frisch, 1993)．すでに払ったお金は返ってこないので，過去の行動によって将来の行動を変えるのは合理的ではないとされる．しかし，見るのをやめれば，そのことによって失敗が顕著化して後悔が強まるかもしれない．すなわち，やめるという決定は，時間の無駄は回避するかもしれないが，決定後の不快感情を高める可能性がある．見続けた人は，お金を払ったという過去の行動が失敗であったことを決定づけるのを避けるという隠れた目標を持っていたのかもしれない．つまり，決定後に現れる後悔を予期し，それを避けるために決定状況の中に別の目標を忍び込ませたといえる．

決定後の予期ではなく，決定過程の中にも複数の目標が入り込んでくる．Medin & Bazerman (1999) は，意思決定理論は認知処理に複数レベルを想定する必要があることを指摘した．次のようなメンタルアカウンティング (mental accounting; Thaler, 1985) の問題については，複数の目標が分断されていると考えられる．たとえば，「二人の学生がカジノを訪れたとし，学生 A はカジノの前で現金 25 ドルを拾い，その後 25 ドルの入場料を払って中に入ったが，学生 B は 25 ドルの入場料を払って中に入った後，カジノの中で現金 25 ドルを拾った」とする．「賭金 25 ドルが 50% の確率で倍になる賭けがあるが，学生 A と B のどちらが参加する可能性が高いか」を考えると，（収支はどちらも同じであるが）B の方が参加しやすいと予想される．その理由については，入場前に得た 25 ドル（A）は入場という目標に充てられるが，すでに目標を達成した後に得た現金（B）は自由になるためと考えることができる．つまり，意思決定理論は，効用の最大化という単一目標だけを想定するのではなく，時間軸上に複数の分断された目標が系列的に存在することを考慮する必要がある．

4.2 行為の意味：共時的多重目標

Medin & Bazerman (1999) が正しく指摘したように，意思決定には，本来の決定という側面以外に，決定という行為がもつ「意味（meaning）」がある．これが，課題に添った「合理的」目標とは別の目標を形成することもある．こうした意味の一例として Medin らがあげたのは，メッセージ伝達性である．たとえば，1 セント（約 1 円）のチップには，サービスに不満足であることだけでなく，チップを置き忘れてはいないことも同時に伝えることによって，前者を顕著化するという目標がある．つまりお金の問題ではない．

これと同様の観点から，Medin & Bazerman (1999) は最後通牒ゲーム (Güth et al., 1982) における行動についても論じている．最後通牒ゲームとは，提案者と応答者の二人で行う実験ゲームで，提案者だけが報酬（たとえば 1000 円）の分配額を提案することができ，応答者はその提案をそのまま受け入れるか拒否するか（拒否すると二人ともゼロ円になる）のいずれかの反応をする．このゲームの合理解は，応答者は自分の受取額が 0 円より大きければ（効用

が正であるので）受け入れ，提案者はそれを見越して「相手に 1 円／自分に 999 円」と配分するというものである．実際には，そのような極端な提案をする提案者は非常に稀であり，応答者の方は提案が不公平になるほど拒否する場合が多くなる．応答者の拒否には，提案者に対する怒りの表出というメッセージ伝達の側面があるとされている (シグナルとしての成立要件の分析については，小林, 2021 を参照)．それを裏づけるように，応答者が提案者に対してメッセージを送れるようにすると，怒りの感情が表出されると同時に，少額の提案に対する拒否が減少する (Xiao & Houser, 2005)．つまり，この課題で提案を拒否する応答者は，効用最大化というゲーム理論の規範からは「非合理的」であるが，応答者には，与えられた目標（獲得金額の最大化）以外の目標，すなわち提案者へのメッセージ伝達という隠れた目標があると考えられる．

こうしたメッセージは，自分に対して向けられることもある．Quattrone & Tversky (1984) は，腕を冷水につけることが求められる実験で，心臓が健康な人は長く耐えられると聞いた被験者は，これ以上耐えられないと感じるまでの時間が 1.35 倍も長くなることを明らかにした．心臓の健康は冷圧痛テストの結果に影響を及ぼすが，その逆ではないので，これは「因果推論のエラー」である (Sloman & Fernbach, 2008)．しかし，たとえば，雨にも負けずに毎日ジョギングする自分の姿から自己の意志の強さを確認するように，行動は自己シグナル（self-signaling）になりうる (Bodner & Prelec, 2003)．因果推論に基づく意思決定において，人は規範的な結果効用（outcome utility）だけでなく，診断効用（diagnostic utility）も考慮に入れる．つまり，自己へのメッセージ伝達という隠れた目標も（無意識的に）考慮される．

4.3 自己の多重性

以上の例はいずれも，研究者側が前提とする効用最大化といった規範的な単一の目標とは別に，意思決定者側が暗黙のうちに設定する目標が存在しうることを示唆している．しかも，そうした暗黙の目標は一つとは限らない．たとえば，最後通帳ゲームの応答者の中には「（ゲームで前提とされている）効用最大化を望む自己」に加えて，上述の通り「強欲な相手に対して怒る自己」，さらには「相手に付け込まれないよう防衛的になる自己」「相手（状況によっては実験者や周りの人間も含む）に度量が狭いと思われたくない自己」などの複数の自己が共存している可能性がある．

意思決定者は，多重の自己（multiple selves）を内包している (Medin & Bazerman, 1999)．つまり，規範的な目標を達成する能力がないのではなく，複数の目標を同時に達成しようとするために，結果として規範に必ずしも一致しない行動が発現すると考えることができる．では，合理性が目標依存である（3.3 節）として，目標が複数ある場合の合理性はどう決まるのであろうか．すなわち，実験参加者が，課題の中に私的な目標（研究者が設定した効用最大化という目標以外の多数の目標）を混入させるとき，その参加者にとっての合理性はどう考えたらよいのであろうか．この点については，深く関連する二つの問題，すなわち意識性の問題（5 節）と脱文脈化の問題（6 節）を考察した後（7 節）で議論したい．

目標多重性によって発生する合理性の多義性の問題は，実験者と被験者の課題解釈が異なると言えばその通りであり，よって，その観点に立てば解釈の問題（2.2 節）である．しかし，競合する複数の目標の競合解消問題と捉える方が示唆的である．多くの目標は意思決定者本人にも自覚されていない可能性があるため，目標多重性の問題は，意識的自己と無意識的自己の関係の問題でもある．そこで，次節では，無意識的過程（タイプ 1）と意識的過程（タイプ 2）という二重過程と合理性の関係について議論する．

5. 認知の二重過程

合理性の概念が一つでないことは，認知過程が単一でないことと無関係ではない．認知には，素早くて直感的なもの（タイプ 1）と，遅くて内省的なもの（タイプ 2）の 2 種類があるという考えは，さまざまな研究者によって個別に提唱されてきた [4]．それらは現在，しばしば二重過程理論（dual-process theories）と総称される (e.g., Evans & Stanovich, 2013)．ただ，Kahneman (2011 村井訳 2012) の出版以降，直感的なタイプ 1 が誤りやすく熟考を伴うタイプ 2 が誤り

4) システム 1/2 という呼称は，心が二つあるという仮説を示唆しかねないこと，各処理に単一の処理系を想定しているという誤解を与えかねないことから，Evans & Stanovich (2013) はタイプ 1/2 という呼称を使用することを提唱し，それも広まりつつあるので，本稿もそれに従う．

を正すという見方が世の中に広がりつつある[5]が，こうした考えは根強く存在する誤りである (Evans & Stanovich, 2013)．以下で論じるように，タイプ 1/2 と合理性／非合理性の関係は単純ではない．

5.1 タイプ 2 の合理性

信念バイアス (belief bias; Evans et al., 1983) の研究は，Wilkins (1928) が始めた形式と内容の関係の実験（2.1 節）を系統化したものである．これは，三段論法の結論部分が常識的におかしい（信念に合致しない）ものであれば，論理の誤りが検出されやすい（逆に結論が常識的であれば論理の誤りが見逃されやすい）ことを示したもので，現在の二重過程に関するデフォルト介入説 (default-interventionist accounts; Evans & Stanovich, 2013; Kahneman & Frederick, 2002; Stanovich, 1999) の原点となった．この説によれば，特段の理由がなければ直感的な（誤りやすい）タイプ 1 過程の出力が採用されるが，何らかのトリガーによってタイプ 2 過程が発動されると，タイプ 1 の誤った出力が修正・上書きされる．

デフォルト介入説に対しては，この 10 年ほどでいくつかの異論や拡張が提案されている．たとえば，これまでタイプ 2 が分担すると考えられてきた論理的推論の少なくとも一部は，無意識的に（タイプ 1 で）処理されることが示されている (e.g., De Neys & Pennycook, 2019)．また，信念バイアス研究は信念（内容）から論理性（論理的推論）への影響を明らかにしたが，その逆，すなわち論理性（論理的な正しさ）から信念（結論の信憑性）への影響の方がむしろ大きいことも示されており，その説明として，タイプ 1 と 2 が，系列的ではなく並列的にはたらくとする仮説もある (e.g., Trippas et al., 2017)．こうした知見を踏まえて，論理的推論も（少なくとも一部は）タイプ 1 に依存すると考える必要があり，また，タイプ 2 による介入と考えられていた過程は（少なくとも一部は）タイプ 1 どうしの競合と捉えるべきとされている (Evans, 2018)．

デフォルト介入説は提唱者によって若干の違いはあるが，基本的にタイプ 2 は監視役である．オリジナル版ではタイプ 1 が犯す誤りをタイプ 2 が時系列的に後から正すとしていたが，改訂版では，この考えに並列処理とタイプ 1 内競合の考えが組み込まれた．いずれの場合も，タイプ 2 の振る舞いや機能については，エラー訂正ということ以上に詳しく言及されていないが，その（規範的）合理性は前提となっている．しかし，これとは異なる見方もある．

5.2 タイプ 1 の合理性

Oaksford & Hall (2016) は，タイプ 1 が合理的でタイプ 2 が非合理的と主張する．進化的に古いとされるタイプ 1 過程は，動物の意思決定 (e.g., Monteiro et al., 2013)，知覚・運動意思決定の Bayes 脳理論 (Bayesian brain theory; e.g., Clark, 2013)，論理的直感 (e.g., De Neys & Pennycook, 2019) などの研究結果から，むしろ合理的（規範的）であるとしている．それに対して，ワーキングメモリ（working memory; WM）の制約や言語的なフレーミングの影響を受けるタイプ 2 は，処理の中にエラーが混入しやすい．興味深い仮説として，タイプ 2 は，タイプ 1 過程そのものへはアクセスできず，その「成果物」だけを容量制限のある WM 内に構築するとしている．タイプ 1 の Bayes 的推論過程は連続値的（確率的）であり，そこから推論結果として抽出された WM 内の表象は言語的・離散的である．この仮説は，確率的アプローチ (e.g, Oaksford & Chater, 2007) がタイプ 1 過程を記述し，メンタルモデル理論 (Johnson-Laird, 1983 海保監修 1988) がタイプ 2 過程を記述することを示唆している．この観点は，推論の二大理論の統合の可能性を示唆するもので，最新の演繹推論の確率表象モデル (Hattori, 2016) とも整合的である．また，人工知能の深層学習と記号処理の融合の議論 (松尾, 2022) とも符合する．さらに，言語を通して自分とは異なる他者の推論結果を利用することによって，タイプ 2 による推論エラー訂正の可能性が高まるとし，長い人類の歴史において，選択圧の中で言語が進化した理由も示唆している．

一方，タイプ 1 は合理的でも非合理的でもなく，無合理的（a-rational）とするべきであるという主張もある (Brakel & Shevrin, 2003) これは，合理性と意識の関係をどう考えるかによる（表 1）．合理的と判断されるのは，行動（behavior）ではなく行為（action），すなわち意図的行動（intentional behavior）

5) Kahneman 本人は，次のようにタイプ 1 を「弁護」している点にも留意するべきである．「私たちが犯す誤りの大半はタイプ 1 に端を発するとはいえ，私たちが行う正しいことの大半もタイプ 1 のおかげなのだし，しかも私たちがやることの大半は，そうまちがってはいないのである．私たちの考えと行動は日常的にはタイプ 1 に導かれているが，だいたいにおいて正しい．」(Kahneman, 2011 村井訳 2012, 下巻 p. 270; 原書 p. 416;「タイプ 1」は原著では「システム 1」)

表 1 合理性の二元的分類

合理性	過程	
	無意識的（タイプ 1）	意識的（タイプ 2）
規範的	（N1）	N2
適応的	（A1）	A2

註）何らかのシステムの出力（行動）としての合理性には，規範的（N）／適応的（A）のいずれにも，その出力を生成する過程の意識性（1/2）を想定できるが，無意識的過程の合理性（N1, A1）を定義しない哲学的立場もある．

であるという立場 (e.g., Knauff & Spohn, 2021b, p. 7) に基づけば，意識は合理性の必要条件である．つまり，表 1 の N2 と A2 のみが俎上に載せられる．この論点には 9 節で戻ることとする．

以上のように，二重過程と合理性の関係は単純ではなく，研究者間でコンセンサスが得られているわけでもない．この状況を踏まえて合理性と意識性の間の関係の議論を深めたいが，その前に，二重過程説自体の問題点について指摘しておきたい．

5.3 二重過程を超えて

二重過程の考えは，多くの研究者の支持を集める一方で，一部の研究者から容赦ない批判も浴びせられている．Evans & Stanovich (2013) は，そういった批判を次の 5 点にまとめて反論している．すなわち，(1) 定義があいまいであること，(2) 各過程の属性（無意識的，無意図的など）の分類が不適切であること，(3) 二分法の誤り（処理スタイルの違いは二値的ではなく連続的であるということ），(4) 単一処理モデルでも説明可能であること，(5) 証拠が不十分であること，である．本稿は，二重過程理論について論じることが目的ではないので詳細は省略するが，本論と関係する (1) についてだけ触れておきたい．

二重過程の定義はあいまいではない．Stanovich & Toplak (2012) は，タイプ 1 の属性（自律的，素早い，領域依存，進化的に古い，無意識的，高容量，連合的）とタイプ 2 の属性（熟慮的，遅い，領域一般，進化的に新しい，意識的，容量限定的，規則ベース）のうち，定義的特性として，タイプ 1 の自律性（autonomy），タイプ 2 の WM 使用と認知的分離（cognitive decoupling）をあげた．認知的分離とは，現実と仮想を区別することであり，仮説推論によって，別の可能性を考えてタイプ 1 の推論結果を上書きするために必要な行為である．ただ，この定義による二重過程の区別は，それほど新しいものでも驚くべきものでもない．人間の心に無意識的なものと意識的なものが存在するという指摘は，哲学ではプラトンに，心理学ではフロイトに遡る (Frankish & Evans, 2009)．Pennycook (2018) は，二重過程理論は（反証可能 falsifiable であるが）反駁不能（irrefutable）であるとした．すなわち，シンプルに自律性だけを定義とするのであれば，こうした区別が存在すること自体は多くの人の直感に一致するものであり，したがって証明や反証の対象ではなく前提とするべきものである．つまり，問題はその先にある．予測がないものは理論ではない (Melnikoff & Bargh, 2018a, p. 668) という指摘は本質を突いている．

おそらく，現在の二重過程の考えに決定的に足りないものは二つある．第 1 は，タイプ 1/2 過程の間の相互作用の知見と理論である．タイプ 1/2 過程の相互作用については，洞察問題解決において反直感的な知見がわずかに報告されているものの (たとえば，西田他, 2018)，圧倒的に知見が少ない．プライミングなどの工夫された実験手法によって，タイプ 1 過程については多くの驚くべき知見が蓄積している (e.g., Hassin et al., 2005)．また，タイプ 2 の介入（タイプ 1 の推論結果の訂正）については，よく研究が進んでいる．しかし，タイプ 2 がタイプ 1 の過程に直接アクセスできない（5.2 節）とすれば，タイプ 1 とタイプ 2 の方向性が異なるような場合に，一つの心の中で両者がどのように相互作用するのかについては，もっと多くの知見が必要である．

第 2 は，タイプ 2（意識）のメカニズムや機能についての理論である．タイプ 2 に WM が必要であることや，タイプ 2 がタイプ 1 に介入するといったことは，タイプ 2 がはたらいたときに何が必要とされたか，また結果として何が起こったかという事後的な相関的記述でしかない．タイプ 2 は，どういうときにどのようにはたらくのか，その動きはどのように決まるのか，またその機能は何か，といったことについては，現在の二重過程説では十分に触れられていない．こうした観点から重要と思われる意識の機能の問題については，9 節で論じる．その前に議論の準備として，次節では二重過程と合理性の関係を別の視点から考えておきたい．

6. 自閉スペクトラム症の合理性

Chapman & Chapman (1959) は，人々の不正換位［脚注 1)］の推論を明らかにし，それが，それまで統合失調症という臨床群だけにみられるとされていた述語的同一視の推論と同じであることを指摘した (服部, 2008; 中野・篠原, 2008)．その後の 20–30 年間は，一般の（あるいは教養のある）人々の推論や判断の誤りやすさを明らかにする HB 研究（2.1 節）が主流となったが，この 10 年ほどで，その研究成果のコインの裏側ともいえる知見が急速に蓄積されている．すなわち，臨床群，特に自閉スペクトラム症（autism spectrum disorder，以下 ASD とする）の推論・判断が優れていることを明らかにする多くの研究が現れている (Rozenkrantz et al., 2021)．以下では，そうした知見から合理性について考えたい．

6.1 感情，直感，文脈

ASD の人は，フレーミング (framing; Tversky & Kahneman, 1981) の影響を受けにくいことが明らかにされている (De Martino et al., 2008)．この実験で参加者は，5 千円を受け取った上で，40% の確率で全額を自分のものにできるくじ（60% の確率で何ももらえない）か，必ず 2 千円をもらう（必ず 3 千円を失う）か，どちらかを選ぶといった場面想定法の課題に取り組んだ．この場合，2 千円をもらうこと（利得フレーム）と 3 千円を失うこと（損失フレーム）は同じことを意味するが，確実な利得（として表現された選択肢）は選ばれやすく，確実な損失は避けられやすい（くじが選ばれやすい）．ASD 群ではこうしたフレーミング効果の影響が少ないが，その原因は，感情統合能力の弱さにあることが生理指標データから示唆された．同様に，埋没費用錯誤（4.1 節）についても，ASD の人は影響を受けにくい (Fujino et al., 2019)．ASD 群では，感情を含む直感的推論（タイプ 1）が苦手なので，分析的（タイプ 2）意思決定の傾向が強まると考えられている．

ASD の合理性には，別のメカニズムも考えられる．ASD の人は，誘引効果（atraction effect）の影響も受けにくい (Farmer et al., 2017)．誘引効果とは，同程度の選好の二つの選択肢 A, B があるときに，A に似ているが少し劣る新しい選択肢 C を加えると，A の選好が高まることを指す．さらに，ASD の人は連言錯誤（2.2 節）の影響も受けにくいとされている (Morsanyi et al., 2010)．しかし，誘引効果や連言錯誤には，上述のフレーミングのように感情や報酬は関与していないようにみえる．この場合の ASD の合理性は，全体的文脈情報に対する感受性が低いこと，すなわち，一つ一つの対象を他のものと結びつけず，独立したものと捉えようとする傾向と関係があるとされている．

ただ，Morsanyi et al. (2010) の実験で ASD の連言錯誤が低減したのは，文脈情報があるときだけであった．よって，ASD の人は文脈処理をしないのではなく，文脈情報をタイプ 2 で処理しているために認知資源を消費し，その結果，失敗しやすいという解釈が示されている．それと整合的に，Brosnan et al. (2016) は，ASD の人がタイプ 1 よりタイプ 2 をよく使うことを，認知的熟考テスト (Cognitive Reflection Test, CRT; Frederick, 2005) や合理経験目録 (Rational Experiential Inventory, REI; Epstein et al., 1996) を使って明らかにしている．なお，Farmer et al. (2017) は，ASD の誘引効果の説明としては，全体的文脈に対する感受性の低さより，局所的な情報処理を好む傾向の方が妥当としており，課題によって関与する要因が異なる可能性もある．

6.2 社会性，楽観性

さらに，ASD の合理性の別の側面が，最後通牒ゲーム（4.2 節）を使用した研究によって明らかにされている．ASD 群では合理解に沿った行動の割合が高い (Sally & Hill, 2006)．提案者としてはゼロ提案をする者が多く（約 1/3），応答者としてはゼロ提案を受け入れる者が多い（約 1/3）．応答者の受取額が全体額の 30% 以下の提案がなされた場合の拒否率は，神経学的定型（neurologically typical; NT）では 89% であったのに対し，ASD では 68% にとどまった．この結果は，心の理論（theory of mind）の障害，すなわち，いわゆる他者の心を読む能力に問題があるせいで他者の感情や行動がわかりにくいためとされた．ASD でなくても，相手がコンピュータで提案額がランダムに決定される場合にはゼロ提案に甘んじる (Blount, 1995) という結果は，相手に「読む」べき「心」がなければゲームの中に感情を持ち込んでも意味がないことを示している．つまり，感情が切り離されれば「合理解」が受け入れられるといえる．ただし，最近の研究で Jin et al. (2020) は，実行機能テストの結果を踏まえて，ASD

の人は，タイプ 2 ではなくむしろタイプ 1（直感）によって不公平な提案を受け入れている可能性を示唆しており，議論は収束していない．

最後に重要な知見として，ASD においては楽観バイアス (optimism bias; Weinstein, 1980) が少ないことを取り上げておきたい．楽観バイアスとは，出来事の起こりやすさ（生起確率）を，望ましいことは高く，望ましくないことは低く評価する傾向をいう．このバイアスは，自分についてのことがらの場合には，他人の場合と比べて，さらに強いものとなる Kuzmanovic et al. (2019) 点が重要である．ところが ASD 群では，他人についてのバイアスはなく，自分についてのバイアスも弱かった．

以上，ASD 群の判断・推論が NT 群に比べて規範に近くなる事例をみてきた．この合理性を支えるメカニズムについては，まだわからないことも多いが，原因候補とされる要因としては，感情統合が苦手な傾向，局所的情報処理を好む（あるいは全体的文脈情報への感受性が低い）傾向，タイプ 2 過程をよく使用する（あるいはシステム化）傾向などがある．また，それゆえに認知資源が消費されて，課題遂行の失敗につながる場合があるという側面も見逃せない．

7. 合理的になるべきか

これまでの HB 研究（2.1 節）でよく用いられてきた課題において，ASD の人がよいパフォーマンスを示すという前節の知見は，推論・判断・意思決定研究で想定されてきた合理性が，どういう合理性であったのかを浮き彫りにする．そうした課題で要求されていたのは，いわば「切り出された」状況（4.1 節）での思考，すなわち脱文脈化された思考であり，単一目標の思考である．合理性が目標依存（3.3 節）であるとすれば，課題の中での要求は，その課題を使用した研究の「合理性観」を決める．その合理性観は，思考の記述的研究において，あるいは処方的提案に対して，どのような意味をもつであろうか．

脱文脈化思考を求める姿勢は，合理性の概念をかなり限定することになる．それは，言葉（ルール）で表現可能なものだけに限定した合理性であり，多要素の相互依存性や相互作用のないところでの合理性である．連言錯誤の課題（2.3 節）で，リンダが「銀行窓口係」であるより「フェミニストの銀行窓口係」である方が可能性が高いと感じてしまうのは文脈情報による．これが「ピアニストの銀行窓口係」であったなら，たぶん多くの人は間違えないであろう．ビュリダンのロバも，別の文脈に置かれれば愚かではない（3.3 節）．文脈，暗黙の前提・仮定，事前知識は，課題に対する構えや「ものの見方」といった姿勢をある程度規定する．ものの見方が変われば，規範も適応性判断基準も変わりうる．つまり，目標が変わるということであり，合理性の判断基準が変わるということである．しかし，文脈や暗黙の前提といったものは課題文には書かれていない（だからこそ「文脈」である）ので，そういった情報に左右されるということは，合理性の判断基準はかなり恣意的に変わりうるということになる．もちろん，これは課題解決者（実験参加者）の視点に立った合理性の話であり，その視点に基づく基準である．したがって，まさにその点において，研究者側の「硬い」合理性の基準と課題解決者側の（柔軟な？）合理性の基準の間に乖離が発生する（課題解決者の振る舞いが研究者からは「非合理的」にみえる）．

同様に，単一目標思考を求めることも，合理性の概念を限定する．それは，無意識を無視し，感情を切り離し，過去も将来も切り捨てたところでの合理性である．埋没費用（4.1 節）は過去のものであり，それを気にしないことは，その場では効用最大化の観点から望ましい行動かもしれないが，後で後悔を感じて抑うつが高まるかもしれない．最後通牒ゲーム（4.2 節）で，自分が 1000 円のうちの 999 円を受け取ることができたら，経済的効用は大きいが後味は悪いかもしない．そうした「余分なこと」を（無意識的に）考えている意思決定者は，複数のできるだけ多くの目標を，それぞれできるだけ満足のいく水準で同時に達成できるよう，無意識的に行動を調整する問題解決者である．そうした思考や行動に対する合理性の評価はどうあるべきであろうか．

繰り返すが，合理性は目標によって決まる（3.3 節）．よって，合理性の程度は，目標達成度によって評価されるべきである．しかし，目標が複数存在する場合の目標達成度を正当に評価する客観的な手段はない．複数の目標が互いに質的に異なり，それぞれの達成度を数量化することが困難な目標を統合した全体的達成度は，多属性効用理論のように単純に計算することはできない．たとえば，500 円を失うことと，相手を侮辱して感じる心の痛みという二つのことがらを，数値計算のような方法で一つの尺

度上に統合することは容易ではない．しかし，私たちの認知過程は，こうした統合に相当する何らかの処理を，ごく短時間のうちに実行しているはずである．というのも，私たちの体は一つしかない以上，目標が複数あっても取ることができる行動は一つであるため，何らかの行動を起こすときには，その行動にトリガーをかける何らかの統合的な認知処理があるはずである．それは間違いなく，ほとんどがタイプ 1 過程の仕事である．

2.2 節で，課題の定式化が必ずしも一つに定まらないことを示したが，そういった不定性は，課題の性質だけによるものではなく，文脈やものの見方にもよるといえる．文脈やものの見方は，課題，解決者の内部状態（知識・認知機能など），環境（課題と解決者の置かれた状況）の 3 者に依存して決まると考えられる．したがって，課題だけを「切り出し」て定式化し，その課題に取り組む参加者の行動の非合理性を指摘することは，研究方法（アプローチ）としては有効であったとしても，それは研究の最終到達点では決してない．その次のステップが問題である．その切り出しによって見逃されていたものが何であるのか，多重目標を単一化できない（切り離せない）理由があるのならそれが何なのかを追求する必要がある．その心理メカニズムを明らかにすることが，おそらく思考心理学研究の重要な使命であり，本当に望ましい思考や行動についての提言もそれなしには実現しないと思われる．

脱文脈化思考や単一目標思考を目標に設定することによって，その目標に対応した合理性が決まる．しかし，その合理性を満たす思考が望ましいものかどうかは別問題である．よく考案された実験課題は，その課題で求められる思考の合理性の優秀なリトマス試験紙になる．しかし，その試験結果をどう解釈するべきかの答えは，当然ながら試験紙の中にはない．すなわち，その課題が示す合理性をどう意味づけるのか，さらにはその合理性を満たすべきかどうかは，検査結果を解釈する検査者側に預けられた問題である．以下では，こういった問題について議論する．

8. 合理性と創造性

ここまでは，合理性は目標依存として，設定された目標の「適切性」については不問に付してきた．しかし，「人間は合理的か」という問い（1 節）が必然的に惹起する「人間は合理的になるべきか」という問い（7 節）に答えようとするなら，目標設定の問題を避けて通ることはできない．そこで以下では，目標がどう定まるのか，また，どう定めるべきなのかといった問題を考察する．

まずはその緒として，Bostrom が提示した直交命題（orthogonality thesis）を取りあげたい．Bostrom (2014 倉骨訳 2017) は，人類の知力を凌ぐ人工知能が出現する可能性について論ずる中で，そうした人工知能が設定する目標について，次のような命題を提唱した (翻訳書 p. 228; 原書 p. 130)．

> 【直交命題】知能と最終到達目標は直交的である．いかなるレベルの知能といかなる最終到達目標との間においても，知能と最終到達目標の組み合わせは，基本的に，可能である．

続いて思考実験として，ペーパークリップ超知能（paperclip-maximizing superintelligence; 以下, PMS）の例をあげている．これは，ペーパークリップの生産量を最大化するという目標を与えられ，地球全体を，続いて宇宙までペーパークリップに変えてしまう人工知能である．

PMS は知的であろうか．与えられた目標を忠実に，しかも効率的に達成しようとするという意味では合理的かもしれないが，自分の推論や行為の含意（派生効果）を理解していないという意味では，超知能どころか，むしろ無能にみえる．PMS は，まさに「切り出された」状況（4.1 節）の中だけで最大限の合理性を発揮する極端なケースである．つまり，本論文のここまでの議論を踏まえると，PMS は（まったく知的ではないが）（超？）合理的であることになる [6]．この例でわかるように，目標はシステムの外側にある．合理的エージェント自身は，目標を遂行するだけで自ら目標を設定したり変更したりすることはなく，その合理性評価は，目標の達成度や達成の効率のみでなされる．すなわち，合理性の概念はシステムの中で閉じている．

ところが，このように目標設定と目標達成を完全に切り離してしまうことによって，違和感も発生する．たとえば，自国の要求の実現を所与の目標とした場合，その目標達成のために隣国に対して一方的に武力攻撃をすることは，合理的といえるだろうか．

6) これとは正反対に，Bostrom (2014 倉骨訳 2017) は，PMS は合理的ではないが（超）知的であるとしている (翻訳書 p. 228; 原書 p. 130)．本論文との見方の相違点を改めて強調しておきたい．

あるいは，自らの権力拡大を目標として，宗教を隠れ蓑に大量無差別殺人をすることはどうであろうか．これらも，突出した目標を徹底的に追求するという事例であるようにみえる．つまり，見方によっては，切り出された状況で求められる推論に似ている．よって，ここまでの議論を踏まえれば，PMS と同様，合理的ということになるであろう．しかし，合理性という言葉をこのように使用することに問題はないであろうか．英語の合理性（rationality）という言葉は，理性（reason）と語源が同じ（ラテン語の ratio）であり，心的能力の観点からは両者はもともと同義であったとされる (Broome, 2021)．日本語においても，両者の意味は似ている．もちろん，これは単に言葉の定義の問題であるので，定義はどのようなものとすることもできる．しかし，専門用語の定義が，一般的な言葉の意味から大きく乖離することは避けなければならない．そうであるならば，設定された目標の適切さ，目標に付与された優先順位や重みづけの適切さなども，一定程度は合理性の概念の中に取り込まれるべきであろう．

適切な目標の設定や調整は，目標追求型思考だけでは決して実現できない．Ohlsson (2011) は，K. Gödel の不完全性定理（incompleteness theorem）と J. Fodor の思考の言語（language of thought）の例をあげて，演繹の生成性と限界を論じている (p. 55)．演繹的形式記号システムは，原理的にすべての可能な真理の部分集合しか生成できないため，最初に与えられた記号システムを意味論的に拡張することはできない．よって，人間なら誰でも行っているようなこと，すなわち，枠組みや視点を変えて認知的表象力を段階的に高めるといったこと（創造性）は，演繹システムではモデル化することができない．

枠組みの転換をもたらす非演繹的・非目標志向的な思考には，創造的な特性がある．創造性は，芸術や科学や発明に直結するものだけではない．たとえば，思考の一段階として孵化（incubation）とよばれる現象が古くから知られている (e.g., Wallas, 1926 松本訳 2020)．これは，当該の問題から一時的に離れているときに，ふとよいアイデアが浮かんでくるといった現象で，誰もが経験することである．実際，問題を解決するという目標を継続的に強く追求するだけでははなく，むしろ，気分転換をする（目標をあきらめる）ことや別の問題に取り組む（別の目標に切り替える）ことが結果的に有効な場合もある．このような本来の目標志向的ではない行動を，一時的にしろ，また意識的にではないかもしれないが，別の目標として設定するためには，対象レベルを超えたメタレベルの（意識的あるいは無意識的な）思考[7)]が必要である．こうした思考こそが，おそらく理性あるいは知性の本質的な特性であり，その実現のためには，Hofstadter (1979 野崎他訳 1985) が正しく指摘したように，「システムの外に跳び出す（jump out of the system）能力」(翻訳書 p. 667; 原書 p. 678) が必要である．

9. 創造性と意識的自己

ここまでの議論で明らかになったのは，合理性は目標に依存する概念であり（3 節），目標多重性が合理性の多義性の主な要因であること（4 節），認知の二重性もこの問題に関係していること（5 節）である．さらに，合理性という概念の妥当性を担保するためには目標追求の側面を考えるだけでは不十分で目標設定の側面も考慮するべきことが提起され，よって合理性の中に創造性の要素が必然的に取り込まれることが示された（8 節）．このことは，合理性という「冷たい」と思われがちな概念が，実は人間臭い「熱い」概念でもあるべきことを意味する．以下では，この点について議論する．

9.1 合理性が愚かにみえるとき

PMS（8 節）が愚かにみえる理由の一つは，メタ的な視点がないことである．すなわち，PMS には自らの課題解決事態を一段上から眺める視点がない．それは，目標が単一であることと密接に関係している．単一目標の課題は，課題を脱文脈化することが容易であり，むしろそれが，余分な情報を排除するという意味で望ましい．目標が単一であれば，その目標だけを見据えればよいため，メタ的な視点は不要である．しかし，もし別の目標，たとえば他者のウェルビーイングに配慮するといった目標が同時に存在するならば，課題を別の視点からも眺める必要がある．

さらに難しいのは目標発見である．目標設定が課せられるというのは，単に与えられた目標候補の中から自分の目標を選ぶことではない．目標自体を

7) こうした思考がもつ特性を，服部 (2017) はメタ合理性，またその概念を拡張して服部 (2018) はパラ合理性とよんだ．しかし，合理性の概念をさらにあいまいにすることを避けるため，本稿ではこれらの語は使用しない．

自ら創り出さなければならない．たとえば，私たち人間は，生まれたときに目標が与えられているわけではなく，人生を生きていく中で，自ら目標を見つけ，目標を定め，またそれを更新していく．こうした目標設定のためには，以下に示すように，もう一つ PMS に足りないものが必要となる．

目標の発見と設定のためには，メタ的視点を持つことが必要であるが，それだけでは十分ではない．Damasio (1994 田中訳 2010) は，生物的欲求と情動（＝内外の情報から引き起こされる心的現象の評価プロセスと身体的変化の組み合わせ）が，「とくに個人的かつ社会的領域における合理的な行動にとっては本質的に重要である」(翻訳書 p. 299; 原書 p. 192) とした．前頭前野腹内側部（vmPFC）を損傷した患者は，二つの面会候補日の費用便益分析は著しく冷静に行うが，結局，30 分経っても日を決められなかった．Damasio は，この事例を「純粋理性の限界（limits of pure reason）」の好例 (翻訳書 p. 301; 原書 p. 193) として紹介している．おそらく，目標を発見し，候補となる複数の目標の優先順位や重みづけを決めるといったことをするためには，大局的で客観的な視点と主観的な視点を往還する必要があり，そのためには，文脈や感情が重要となる．ここで，文脈情報や感情の処理が苦手な ASD の人が「切り出された」課題において合理性を発揮するという事実（6 節）が思い起こされる．「単一目標の合理的な実行」と「合理的な目標の設定」には，二律背反的な側面があるのかもしれない．

9.2 ウェルビーイングと自己の物語

目標設定に強く関係するのがウェルビーイングである．Quattrone & Tversky (1984) の実験では，自己シグナルが隠れた目標になっていたことを指摘した（4.2 節）．このケースは一種の自己欺瞞であり，規範的観点からは合理的な因果推論ではないが，もし自己シグナルが満足感や自己効力感をもたらし，その結果，幸福感が得られる (Taylor & Brown, 1988) とすれば，この行動は，「気持ちよく生きる」という生物としての基本的な目標に合致するという意味で適応的観点からは合理的とみることもできる．楽観バイアス（6 節）についても，同様のことがいえる (Bortolotti & Antrobus, 2015; Sharot, 2011)．楽観バイアスは，精神的・身体的健康や社会的成功と関連があることが明らかにされている．楽観主義者は寿命が長く，給与が高く，ビジネスで成功していること，その逆に，将来に対してポジティブな期待をもたないことは軽度のうつや不安と関連があることが明らかにされている (ただし，Makridakis & Moleskis, 2015 も参照)．なお，ASD の人に楽観バイアスが少ないことから，ここでも ASD の合理性とのトレードオフが認められる．

ウェルビーイングを考えるときに必然的に関わってくるのが，自己や意識の問題である．私たちの幸福感や快楽（あるいはその逆の苦痛）といったことの評価は，瞬間瞬間の心地よさを積み上げた総和で決まっているのか，出来事を後から振り返ったときの印象（記憶）で決まっているのか．この問題を Kahneman (2011 村井訳 2012) は，経験する自己（experiencing self）と記憶する自己（remembering self）という二つの乖離する自己の問題として表現している．Kahneman et al. (1993) は，低温昇圧試験の実験で，14°C の冷水に 60 秒間手を入れる場合（短時間条件）と，それと同じことを経験した後で続いて 15°C の冷水に 30 秒間手を入れる場合（長時間条件）では，後半に少し苦痛が和らぐ長時間条件の方が苦痛経験時間は長いにもかかわらず，むしろ好まれることを示した．この結果から，出来事の評価が，経験する自己ではなく記憶する自己によって決まること，しかも，その評価に対して持続時間はほぼ無関係で，ピーク時と終了時の苦痛の平均で決まることがわかった．なぜ人がそのような評価をするかは明らかではないが，他の動物においても同様のことが確認されていることから，快楽や苦痛の総量は進化生物学的に重要ではない可能性も指摘されている (Kahneman, 2011 村井訳 2012, 下巻 p. 223; 原書 pp. 383–384)．経験の評価を記憶に頼る，すなわち自己の物語に依拠することは，おそらく，以下で論じる生物，特に人間の未来志向性や，さらには意識の機能と関係しているように思われる．

3.1 節で，後悔の予期が合理性の決定要因として適切かという問題提起をしたが，ここでやっとこの点について考察することができる．後悔とは，出来事を振り返ったときの感情であるため，それを予期するのにも記憶が必要である．よって，予期後悔の発生は，記憶する自己のはたらきによる．合理性が目標によって決定し，ウェルビーイングが行動の重要な目標の一つでありうるなら，しかも，記憶する自己が主に幸福感を決定し，その自己に予期後悔を

考慮する傾向があるとすれば，予期後悔を合理性の決定要因とすることには十分な根拠がある．とはいえ，この説明はややトートロジー的である．その響きを低減するためには，記憶する自己が幸福感の評価に対して専制的権限をもつことについて，もう少し粒度の細かい進化生物学的説明が望まれる．上で自己シグナルや楽観バイアスの適応的合理性について触れたが，後悔感情の漏出（4.1 節）にも同様の意味があるように思われる．そう考えれば，たとえば埋没費用錯誤は，気にしても意味のない過去にこだわった（非合理的）行動ではなく，むしろ，将来の幸福を見越した上での現在の未来志向的な（合理的）行動とみることができる．

9.3 意識の機能

Kahneman が区別した経験する自己と記憶する自己は，Damasio (1999 田中訳 2018) の中核自己（core self）と自伝的自己（autobiographical self）の区別や，Gallagher (2000) の最小自己（minimum self）と物語的自己（narrative self）の区別に対応すると思われる．おそらく，記憶する自己は，経験する自己より進化的に新しい．その発生には，何らかの進化生物学的メリットがあったと考えるのが自然である．そこで思い起こされるのは，意識の機能についての次の議論である．

Fodor (2004) は，「フロイト以来，非常に複雑な心理的プロセスが無意識的に行われうること」が明らかになっており，「私たちの意識がすることは意識がなくても同じようにできる」のに「なぜ神はわざわざ意識を作ったのか」(p. 31) と問うた．これに対して Humphrey (2011 柴田訳 2012) は，Fodor は「完全に誤った角度から問題を眺めている」とし，意識の役割は「それがなければできないようなことをできるようにするのではなく，それがなければしようとは思わないことをするようにやる気を出させること」(翻訳書 p. 93; 原書 p. 71) とした．その上で，意識の機能として，(1) 意識を持つことを満喫する，(2) 自分が意識を持って生きている世界を愛する，(3) 意識を持っている自己を尊ぶ，という三つのレベルを提案した (翻訳書 p. 97; 原書 p. 74)．第 3 のレベルの機能は，人間とそれにごく近い動物だけにあるもので，Humphrey は，人間が自分自身をかけがえのないものとして尊ぶこと，死の脅威を感じること，不滅の魂を信じること，二元論的な考えや本質主義的な考えを抱くことなどが，このレベルの意識の機能から説明できることを示唆している．

このレベルの意識を持つものは，意識を持つという事実，意識を持っている自己，自らの自由意志といったものを特別のものとして扱う．高いレベルの意識を持つことと意識を特別のものと感じて尊ぶことは，どちらが原因でどちらが結果でもない．両者が渾然一体化して，将来の何かに対する強い志向性，自発性，創造性といった（熱い）特性を生じさせると考えられる．そう考えれば，合理性の判断において人が意識的な思考を特別視することの意味がわかる．二つの目標の間で葛藤するジェーン（3.1 節）の合理性が，強く意識されている方の目標だけで決まると考えるのも，意志を伴う行動だけが合理性のスコープに入るとされる哲学的信念（5.2 節，表 1）も，こうした第 3 のレベルの意識の機能と一体化していると考えられる．

バイアスや錯誤と呼ばれる心理現象やその理論で，意識のこのレベルに関係するものは無数にあると思われる．すべてを挙げることはできないが，特に気になるものとしては，スポットライト効果 (spotlight effect; Gilovich et al., 2000)，透明性の錯誤 (illusion of transparency; Gilovich et al., 1998)，コントロールの錯誤 (illusion of control; Langer, 1975)，保有効果 (endowment effect; Kahneman et al., 1990)，単純所有効果 (mere ownership effect; Beggan, 1992)，マイサイドバイアス (myside bias; Perkins, 1985)，後知恵バイアス (hindsight bias; Fischhoff, 1975)，知識の呪縛 (curse of knowledge; Birch & Bloom, 2004)，存在脅威管理 (terror management; Solomon et al., 1991) などがある．さらに，「思考の図と地」と呼ばれる現象 (服部, 2014; Hattori et al., 2016) も，このレベルの意識の機能に強く依存していると考えている．

意識がなければ，熱烈に何かに打ち込むこともない．つまり，意識は創造性の源でもある．Humphrey (2011 柴田訳 2012) は，高度な知性を持つ地球外生物がいたとしても，意識を持たなければ物事に関心を抱かないため，わざわざ地球へ来ることはないとした (翻訳書 p. 266; 原書 p. 208)．同様に服部 (2018) は，意識や自己のない人工知能は自発的に何か語り始めることがなく，本当の知性を発揮することはないと論じた (p. 777)．本誌第 29 巻第 2 号の誌上討論で展開された「創造的自己」の議論 (石黒他, 2022) も，こうした考えと整合的であるように

思われる．これまでの認知科学は，規範を物差しとしながら，巧みな実験によってさまざまな認知現象を個別のバイアスとして同定してきた．しかし，個別的な現象を非合理的なバイアスと同定することは研究の最終到達点ではない．それらを統合的に理論化しなければならない．そのためには，意識や自己といった概念がキーとなるかもしれない．

謝　辞

本論文の草稿に対して，清河幸子氏，特集担当エディタの眞嶋良全氏と本田秀仁氏より有益なコメントを得た．本研究は，科研費 21K18567 の助成を受けた．以上，ここに記して謝意を表す．

文　献

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates.

Beggan, J. K. (1992). On the social nature of nonsocial perception: The mere ownership effect. *Journal of Personality and Social Psychology*, *62* (2), 229–237. https://doi.org/10.1037/0022-3514.62.2.229

Birch, S. A. J., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences*, *8* (6), 255–260. https://doi.org/10.1016/j.tics.2004.04.011

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology*, *96* (1), 85–94. https://doi.org/10.2307/1422211

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, *63* (2), 131–144. https://doi.org/10.1006/obhd.1995.1068

Bodner, R., & Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. In I. Brocas, & J. D.Carrillo (Eds.), *The psychology of economic decisions, Volume 1: Rationality and well-being*. Oxford University Press.

Bortolotti, L., & Antrobus, M. (2015). Costs and benefits of realism and optimism. *Current Opinion in Psychiatry*, *28* (2), 194–198. https://doi.org/10.1097/YCO.0000000000000143

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. (ボストロム, N.　倉骨 彰 (訳) (2017). スーパーインテリジェンス：超絶 AI と人類の命運　日本経済新聞社)

Brakel, L. A. W., & Shevrin, H. (2003). Freud's dual process theory and the place of the a-rational. *Behavioral and Brain Sciences*, *26* (4), 527–528. https://doi.org/10.1017/S0140525X03210116

Broome, J. (2021). Reasons and rationality. In M. Knauff, & W. Spohn (Eds.), *The handbook of rationality* (pp. 129–136). MIT Press.

Brosnan, M., Lewton, M., & Ashwin, C. (2016). Reasoning on the autism spectrum: A dual process theory account. *Journal of Autism and Developmental Disorders*, *46* (6), 2115–2125. https://doi.org/10.1007/s10803-016-2742-4

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology*, *58* (3), 220–226. https://doi.org/10.1037/h0041961

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4* (3), 317–331. https://doi.org/10.1017/S0140525X00009092

Colman, A. M. (1995). *Game theory and its applications in the social and biological sciences*. Butterworth-Heinemann.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. Quill. (ダマシオ, A.　田中 三彦 (訳) (2010). デカルトの誤り：情動, 理性, 人間の脳　筑摩書房)

Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Vintage. (ダマシオ, A.　田中 三彦 (訳) (2018). 意識と自己　講談社)

De Martino, B., Harrison, N. A., Knafo, S., Bird, G., & Dolan, R. J. (2008). Explaining enhanced logical consistency during decision making in autism. *Journal of Neuroscience*, *28* (42), 10746–10750. https://doi.org/10.1523/JNEUROSCI.2895-08.2008

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28* (5), 503–509. https://doi.org/10.1177/0963721419855658

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgement and choice. *Annual Review of Psychology*, *32* (1), 53–88. https://doi.org/10.1146/annurev.ps.32.020181.000413

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71* (2), 390–405. https://doi.org/10.1037/0022-3514.71.2.390

Evans, J. St. B. T. (1993). Bias and rationality. In K. I. Manktelow, & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 6–30). Routledge.

Evans, J. St. B. T. (2018). Dual process theory: Perspectives and problems. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 137–155). Routledge.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11* (3), 295–306. https://doi.org/10.3758/BF03196976

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Psychology Press. (エヴァンズ, J. St. B. T.・オーバー, D. E.　山 祐嗣 (訳) (2000). 合理性と推理：人間は合理的な思考が可能か　ナカニシヤ出版)

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8* (3), 223–241. https://doi.org/10.1177/1745691612460685

Farmer, G. D., Baron-Cohen, S., & Skylark, W. J. (2017). People with autism spectrum conditions make more con-

sistent decisions. *Psychological Science*, *28* (8), 1067–1076. https://doi.org/10.1177/0956797617694867

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1* (3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Fodor, J. A. (2004, March 4). You can't argue with a novel. *London Review of Books*, *26* (5), 30–31.

Frankish, K., & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In J. St. B. T. Evans, & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–29). Oxford University Press.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19* (4), 25–42. https://doi.org/10.1257/089533005775196732

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes*, *54* (3), 399–429. https://doi.org/10.1006/obhd.1993.1017

Fujino, J., Tei, S., Itahashi, T., Aoki, Y., Ohta, H., Kanai, C., Kubota, M., Hashimoto, R., Nakamura, M., Kato, N., & Takahashi, H. (2019). Sunk cost effect in individuals with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *49* (1), 1–10. https://doi.org/10.1007/s10803-018-3679-6

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, *4* (1), 14–21. https://doi.org/10.1016/S1364-6613(99)01417-5

Gigerenzer, G., Todd, P. M., & The ABC Research Group. (1999). *Simple heuristic that make us smart*. Oxford University Press.

Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*, *78* (2), 211–222. https://doi.org/10.1037/0022-3514.78.2.211

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of other's ability to read one's emotional states. *Journal of Personality and Social Psychology*, *75* (2), 332–346. https://doi.org/10.1037/0022-3514.75.2.332

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109* (1), 75–90. https://doi.org/10.1037/0033-295X.109.1.75

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3* (4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Harman, G. (1995). Rationality. In E. E. Smith, & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (2nd ed., pp. 175–211). MIT Press.

Hassin, R. R., Uleman, J. S., & Bargh, J. A. (Eds.). (2005). *The new unconscious*. Oxford University Press.

服部 雅史 (2008). 推論と判断の等確率性仮説：思考の対称性とその適応的意味　認知科学, *15* (3), 408–427. https://doi.org/10.11225/jcss.15.408

服部 雅史 (2014). 思考の図と地：フレーミングによる肯定・否定の非対称性　立命館文学, *636*, 131–147. https://doi.org/10.34382/00006478

Hattori, M. (2016). Probabilistic representation in syllogistic reasoning: A theory to integrate mental models and heuristics. *Cognition*, *157*, 296–320. https://doi.org/10.1016/j.cognition.2016.09.009

服部 雅史 (2017). 合理性と目標多重性：限定合理性と二重合理性を超えて　日本認知科学会第 34 回大会発表論文集, 111–113. https://www.jcss.gr.jp/meetings/jcss2017/proceedings/pdf/JCSS2017_OS02-1I.pdf

服部 雅史 (2018). 人工知能は創造的認知の何を語るか：思考の二重性と合理性に基づく一考察　人工知能, *33* (6), 771–779. https://doi.org/10.11517/jjsai.33.6_771

Hattori, M., Over, D. E., Hattori, I., Takahashi, T., & Baratgin, J. (2016). Dual frames in causal reasoning and other types of thinking. In N. Galbraith, E. Lucas, & D. Over (Eds.), *The thinking mind: A festschrift for Ken Manktelow* (pp. 98–114). Routledge.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. Basic Books. (ホフスタッター, D. R.　野崎 昭弘 他 (訳) (1985). ゲーデル, エッシャー, バッハ：あるいは不思議の環　白楊社)

Humphrey, N. (2011). *Soul dust: The magic of consciousness*. Princeton University Press. (ハンフリー, N.　柴田 裕之 (訳) (2012). ソウルダスト：〈意識〉という魅惑の幻想 紀伊國屋書店)

石黒 千晶・清水 大地・清河 幸子 (2022). 誌上討論「『創造的自己』をめぐって」編集にあたって　認知科学, *29* (2), 266–269. https://doi.org/10.11225/cs.2021.020

Jin, P., Wang, Y., Li, Y., Xiao, Y., Li, C., Qiu, N., Weng, J., Fang, H., & Ke, X. (2020). The fair decision-making of children and adolescents with high-functioning autism spectrum disorder from the perspective of dual-process theories. *BMC Psychiatry*, *20* (1), Article 152. https://doi.org/10.1186/s12888-020-02562-8

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge University Press. (ジョンソン＝レアード, P. N.　海保 博之 (監修) (1988). メンタルモデル：言語・推論・意識の認知科学　産業図書)

Kahneman, D. (2011). *Thinking, fast and slow*. Penguin. (カーネマン, D.　村井 章子 (訳) (2012). ファスト&スロー：あなたの意思はどのように決まるか？ (上・下)　早川書房)

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4* (6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, *98* (6), 1325–1348. https://doi.org/10.1086/261737

Keys, D. J., & Schwartz, B. (2007). "Leaky" rationality: How research on behavioral decision making challenges nor-

mative standards of rationality. *Perspectives on Psychological Science*, *2* (2), 162–180. https://doi.org/10.1111/j.1745-6916.2007.00035.x

Kirby, K. N. (1994). Probabilities and utilities of fictional outcome in Wason's four-card selection task. *Cognition*, *51* (1), 1–28.

Knauff, M., & Spohn, W. (Eds.). (2021a). *The handbook of rationality*. MIT Press.

Knauff, M., & Spohn, W. (2021b). Psychological and philosophical frameworks of rationality: A systematic introduction. In M. Knauff, & W. Spohn (Eds.), *The handbook of rationality* (pp. 1–65). MIT Press.

小林 佳世子 (2021). ゲーム理論からみた怒りの感情の役割：最後通牒ゲームの受諾者を題材として　認知科学, *28* (3), 445–457. https://doi.org/10.11225/cs.2021.033

Kuzmanovic, B., Rigoux, L., & Vogeley, K. (2019). Brief report: Reduced optimism bias in self-referential belief updating in high-functioning autism. *Journal of Autism and Developmental Disorders*, *49* (7), 2990–2998. https://doi.org/10.1007/s10803-016-2940-0

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32* (2), 311–328. https://doi.org/10.1037/0022-3514.32.2.311

Makridakis, S., & Moleskis, A. (2015). The costs and benefits of positive illusions. *Frontiers in Psychology*, *6*, Article 859, 1–11. https://doi.org/10.3389/fpsyg.2015.00859

Manktelow, K. I. (2004). Reasoning and rationality: The pure and the practical. In K. I. Manktelow, & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 157–177). Psychology Press.

Manktelow, K. I. (2012). *Thinking and reasoning: An introduction to the psychology of reason, judgment and decision making*. Psychology Press. (マンクテロウ, K. 服部 雅史・山 祐嗣 (監訳) (2015). 思考と推論：理性・判断・意思決定の心理学　北大路書房)

松尾 豊 (2022). 知能の 2 階建てアーキテクチャ　認知科学, *29* (1), 36–46. https://doi.org/10.11225/cs.2021.062

Medin, D. L., & Bazerman, M. H. (1999). Broadening behavioral decision research: Multiple levels of cognitive processing. *Psychonomic Bulletin & Review*, *6* (4), 533–546. https://doi.org/10.3758/BF03212961

Melnikoff, D. E., & Bargh, J. A. (2018a). The insidious number two. *Trends in Cognitive Sciences*, *22* (8), 668–669. https://doi.org/10.1016/j.tics.2018.05.005

Melnikoff, D. E., & Bargh, J. A. (2018b). The mythical number two. *Trends in Cognitive Sciences*, *22* (4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001

Monteiro, T., Vasconcelos, M., & Kacelnik, A. (2013). Starlings uphold principles of economic rationality for delay and probability of reward. *Proceedings of the Royal Society B: Biological Sciences*, *280* (1756), Article 20122386. https://doi.org/10.1098/rspb.2012.2386

Moro, R. (2009). On the nature of the conjunction fallacy. *Synthese*, *171* (1), 1–24. https://doi.org/10.1007/s11229-008-9377-8

Morsanyi, K., Handley, S. J., & Evans, J. St. B. T. (2010). Decontextualised minds: Adolescents with autism are less susceptible to the conjunction fallacy than typically developing adolescents. *Journal of Autism and Developmental Disorders*, *40* (11), 1378–1388. https://doi.org/10.1007/s10803-010-0993-z

中野 昌宏・篠原 修二 (2008). 対称性バイアスの必然性と可能性　認知科学, *15* (3), 428–441. https://doi.org/10.11225/jcss.15.428

西田 勇樹・織田 涼・服部 雅史・V. カストルディ・L. マッキ (2018). 洞察問題解決におけるアイデア生成と抑制機能　認知科学, *25* (1), 100–114. https://doi.org/10.11225/jcss.25.100

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101* (4), 608–631. https://doi.org/10.1037/0033-295X.101.4.608

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, M., & Hall, S. (2016). On the source of human irrationality. *Trends in Cognitive Sciences*, *20* (5), 336–344. https://doi.org/10.1016/j.tics.2016.03.002

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. Cambridge University Press. https://doi.org/10.1017/cbo9780511780295

Pennycook, G. (2018). A perspective on the theoretical foundation of dual process models. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 5–27). Routledge.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, *77* (5), 562–571. https://doi.org/10.1037/0022-0663.77.5.562

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, *46* (2), 237–248. https://doi.org/10.1037/0022-3514.46.2.237

Rozenkrantz, L., D'Mello, A. M., & Gabrieli, J. D. E. (2021). Enhanced rationality in autism spectrum disorder. *Trends in Cognitive Sciences*, *25* (8), 685–696. https://doi.org/10.1016/j.tics.2021.05.004

Sally, D., & Hill, E. (2006). The development of interpersonal strategy: Autism, theory-of-mind, cooperation and fairness. *Journal of Economic Psychology*, *27* (1), 73–97. https://doi.org/10.1016/j.joep.2005.06.015

Sharot, T. (2011). The optimism bias. *Current Biology*, *21* (23), R941–R945. https://doi.org/10.1016/j.cub.2011.10.030

Simon, H. A. (1957). *Models of man: Social and rational*. Wiley.

Simon, H. A. (1991). Cognitive architectures and rational analysis: Comment. In K. VanLehn (Ed.), *Architectures for intelligence* (pp. 25–39). Lawrence Erlbaum Associates.

Sloman, S. A., & Fernbach, P. M. (2008). The value of rational analysis: An assessment of causal reasoning and learning. In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 485–500). Oxford University Press.

Smedslund, J. (1990). A critique of Tversky and Kahneman's distinction between fallacy and misunderstanding. *Scandinavian Journal of Psychology*, *31* (2), 110–120. https://doi.org/10.1111/j.1467-9450.1990.tb00822.x

Solomon, S., Greenberg, J., & Pyszczynski, T. (1991). A

terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol.24, pp.93–159). Academic Press. https://doi.org/10.1016/S0065-2601(08)60328-7

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Elrbaum.

Stanovich, K. E., & Toplak, M. E. (2012). Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, *11* (1), 3–13. https://doi.org/10.1007/s11299-011-0093-6

Sternberg, R. J. (2000). Damn it, I still don't know what to do! *Behavioral and Brain Sciences*, *23* (5), 764–765. https://doi.org/10.1017/S0140525X00003447

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103* (2), 193–210. https://doi.org/10.1037/0033-2909.103.2.193

Thaler, R. (1980). Towards a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1* (1), 39–60. https://doi.org/10.1016/0167-2681(80)90051-7

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science*, *4* (3), 199–214. https://doi.org/10.1287/mksc.4.3.199

Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition*, *45* (4), 539–552. https://doi.org/10.3758/s13421-016-0680-1

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211* (4481), 453–458. https://doi.org/10.1126/science.7455683

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90* (4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Viale, R. (Ed.). (2021). *Routledge handbook of bounded rationality*. Routledge.

Wallas, G. (1926). *The art of thought*. J. Cape. (ウォーラス, G. 松本 剛史 (訳) (2020). 思考の技法 筑摩書房)

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12* (3), 129–140. https://doi.org/10.1080/17470216008416717

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Penguin.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, *23* (1), 63–71. https://doi.org/10.1080/00335557143000068

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39* (5), 806–820. https://doi.org/10.1037/0022-3514.39.5.806

Wilkins, M. C. (1928). The effect of changed material on ability to do formal syllogistic reasoning. *Archives of Psychology*, *102*, 1–88.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, *102* (20), 7398–7401. https://doi.org/10.1073/pnas.0502399102

服部 雅史（正会員）

1996 年北海道大学大学院文学研究科博士後期課程単位取得退学．博士（文学）．1997 年より立命館大学文学部．現在，同大学総合心理学部教授．英国カーディフ大学心理学部，米国ブラウン大学認知言語心理科学部，仏国 ÉPHÉ の各客員研究員を歴任．推論，判断，問題解決を研究する．日本認知心理学会，日本心理学会，日本基礎心理学会，Psychonomic Society ほか会員．