顕在ー潜在ギャップと道徳的責任[*1]*

薄井尚樹

概要

An individual's implicit attitude assessed by indirect measures sometimes conflicts with his/her explicit attitude assessed by direct measures. This phenomenon is sometimes called the explicit-implicit gap. The aim of this paper is to examine the problem with moral responsibility in the context of this gap. First, I will introduce a contemporary form of racism called aversive racism as a typical example of the explicit-implicit gap. Aversive racists explicitly and sincerely acknowledge egalitarian values, whereas they implicitly have negativity towards racial minorities such as Blacks in the United States. It has been shown that aversive racists exhibit a specific pattern of discriminatory behavior under certain circumstances. Based on the framework of aversive racism, I clarify how the problem with moral responsibility for behavior in the context of the explicit-implicit gap can be formulated. Then, I examine whether aversive racists can be morally responsible for their own behavior. In the spirit of attributionism, I here consider two possibilities respectively: One is the possibility that their behavior expresses the values shown in their explicit attitudes. The other is the possibility that it expresses the values shown in their implicit attitudes. As a result, it turns out that both possibilities are denied. Finally, I argue that we can draw three possible conclusions from this discussion. (1) Since the discussion indicates that values shown in either explicit or implicit attitudes of aversive racists are not reflected in their behavior, we cannot attribute moral responsibility to them. In other words, aversive racists are not morally responsible for their behavior at all. (2) The discussion in terms of attributionism is just a starting point for further considerations. It means that we should examine the applicability of other theories of moral responsibility to the explicit-implicit gap. (3) The discussion encourages us to modify attributionism itself. In particular, we should reconsider the

[*1] 本文の強調箇所はゴシック体にした。引用については、原文がイタリック体になっている箇所をゴシック体にした。引用中の[]は引用者による補足をあらわす。引用中の…は引用者による省略をあらわす。

* *CAP* Vol. 13 (2022) pp. 264-284. submitted: 2021.11.28 accepted: 2022.08.09 category: 原著論文 published: 2022.08.30.

nature of the key concept of attributionism, namely the real self.

Keywords: 潜在的態度・道徳的責任

1 はじめに

潜在連合テスト Implicit Association Test（IAT）（Greenwald, McGhee, & Schwartz, 1998）や感情誤帰属手続き Affective Misattribution Procedure (AMP)（Payne, Cheng, Govorun, & Steward, 2005）といった間接的な測定によって査定される個人の潜在的態度 implicit attitude と、直接的な測定（自己報告）によって査定されるそのひとの顕在的態度 explicit attitude は、ときにその対象がおなじであっても食い違うことがある。こうした現象を（ときにそう述べられるように）**顕在ー潜在ギャップ**と呼ぶことにしよう。このようなギャップのありようは、「決定要因としての動機と機会 motivation and opportunity as determinants（MODE）モデル」（Fazio, 1990; Fazio & Olson, 2014; Fazio & Towles-Schwen, 1999; Olson & Fazio, 2009）や「連合的ー命題的評価 associative-propositional evaluation（APE）モデル」（Gawronski & Bodenhausen, 2006, 2011, 2014）などのように、潜在的態度にともなうプロセスと顕在的態度にともなうプロセスを別個に扱う、二重プロセスのモデルによってしばしば説明されてきた。本稿では、とりわけ人種の領域での顕在ー潜在ギャップの事例をとりあげて、そうしたギャップに特有の道徳的責任をめぐる問いを定式化する。そして、その問いを解決することにともなう困難を指摘するとともに、顕在ー潜在ギャップのもとでの道徳的責任のありようについて、今後の研究の方向を示すことを目的とする。

ただし、考察に入る前にひとつ注意しなくてはならないことがある。本来であれば、上記の問題に取り組むためには、ある個人にそのひとの行動の道徳的責任があるのはどのようなときであるのかが、あらかじめわかっていなくてはならない。しかし、それじたいがひとつの大きな哲学的争点であり、それをあわせて論じることは本稿の射程を超えることになるだろう。そこで本稿では、上記の問いに専念するために、あるひとにその行動の道徳的責任がある条件をめぐる議論を避けて、代わりに、近年になって影響力を強めてきた、道徳的責任にかんする説明のひとつ（のあつまり）である、**帰属主義** attributionism[*2] に焦点を絞ることにしよう。この理論によると、「ある行為者になにかの責任があるのは、そのひとにその事柄が適切に**帰属可能**であるばあいだけである」（Levy, 2005, p. 2）[*3]。そして「私の行為が私の本当の自己 real self を表現すると正しく理解されるかぎり――つまり、それが、しかるべき種類のしかたで、私の信念と欲求、価値と関与の産物である…［中略］…かぎり――それは私に適切に帰属可能である」（Levy, 2005, p. 3）。したがって帰属主義によると、ある行為者にある行動の責任があるのは、そのひとのなにがしかの「根本的な評価的姿勢」（Watson, 1996, p. 234）となる本当の自己[*4]（そのひとが何者であるのか）がその行動において表

[*2] 帰属主義という用語は Levy (2005) による。ただし Levy (2005) では帰属主義にたいする批判が展開されている。
[*3] この訳文にかんする査読者の指摘に感謝する。
[*4] 道徳的責任に関連する行為者の心的状態はときに、深い自己 the deep self と呼ばれることもある（たとえば Sripada, 2016）。 本稿では、本当の自己と深い自己を交換可能な用語として扱う。

現されるときだとされる。

帰属主義に焦点を絞る理由は 3 つある[*5]。ひとつは本稿の考察の位置づけにかかわる。John Fischer and Neal Tognazzini (2011)は、道徳的責任の概念を分析するにあたって 15 の問いを段階的に定式化することで、それぞれの問いへの答えにおいて想定される別個の道徳的責任の概念を定式化する。そしてふたりが指摘するように、帰属可能性をめぐる問いはこの分析において他の問いに先行する（Fischer & Tognazzini, 2011, p. 382）。つまり帰属主義は、道徳的責任の概念の分析において他の理論より低い段階に位置づけられる。それゆえ、帰属主義をとりあげることは、他のさまざまな（FischerとTognazzini にしたがえばさらなる問いの段階へと進んだ）道徳的責任の理論が顕在ー潜在ギャップの考察においてどのような主張に着地するにせよ、ひとつの議論の出発点となりうるであろう。第二の理由は帰属主義の独自性にかかわる。帰属主義はしばしば、行為者が意識的に選択あるいは直接的にコントロールした行動だけでなく、自然発生的な行動にも道徳的責任を帰属できる点に、その独自性を求める（たとえば Smith, 2005）[*6]。そしてそうした自然発生的な行動は、自己報告によって測定される顕在的態度よりむしろ、IAT などによって測定される潜在的態度によってうまく予測されることがわかっている（Gawronski, Hofmann, & Wilbur, 2006, p. 492）。このようなつながりをふまえると、顕在ー潜在ギャップのように顕在的態度だけでなく潜在的態度も視野に入れたかたちで道徳的責任を考察するうえでは、帰属主義の立場を採用することに一定の利点が認められるだろう。第三の理由は本稿での論証のかたちにかかわる。以下で見るように複数の論者が、顕在ー潜在ギャップからもたらされる行動の道徳的責任にかんして帰属主義の立場から考察しており、本稿でもこの立場を採用することは、他の論者の見解を検討することを容易にするであろう。

このことを念頭に置いたうえで、本稿は以下のように進行する。まず第 2 節で、顕在ー潜在ギャップの典型的な例として、回避的レイシズム（Dovidio & Gaertner, 2004）という現代的なレイシズムを紹介する。そして、この枠組みをふまえて、顕在ー潜在ギャップのもとでもたらされる行動の道徳的責任をめぐる問いがどのように定式化されるかを明確にする。次に第 3 節で、回避的レイシストに自身の行動の道徳的責任がありうるのかを検討し、既存の立場ではこの問いに肯定的に答えることが難しいことを考察する。最後に、そうした考察からどのような結論を引き出しうるかを論じ、今後の研究の方向を提示する。

2 回避的レイシズム

人種の領域での顕在ー潜在ギャップの典型的な例は**回避的レイシズム** aversive racism（Dovidio & Gaertner, 2004）である。本節では、回避的レイシズムの構造と、それがどのようにして顕在ー潜在ギャップと道徳的責任をめぐる問いをもたらすのかを説明しよう[*7]。

[*5] この論点を明確にするよう助言した匿名の査読者に感謝する。

[*6] 本文で言及することになるように、Smith 自身は自然発生的な行動よりむしろ自然発生的な心的状態に焦点をあてるが、その主張は他の論者によってしばしば行動へと拡張される（たとえば Brownstein, 2018, p. 146; Shoemaker, 2011, p. 606）。

[*7] 以下の回避的レイシズムの特徴づけの記述、およびそのさいの Gaertner & Dovidio (2005)、Dovidio, Gaertner, & Pearson (2017)、Greenwald & Krieger, (2006)、Dovidio, Pearson, & Penner (2019) への言及については薄井 (2020)

John Dovidio and Samuel Gaertner (2004) の特徴づけによると「回避的レイシストは …[中略]… 過去の不正義の犠牲者に同情し、人種の平等の原理を支持し、自身に偏見がないとみなすが、同時に、黒人についてネガティブな感じや信念を所有し、それは無意識的なものかもしれない」(Dovidio & Gaertner, 2004, p. 3)とされる。ここで、前者は顕在的態度であり、自己報告のような直接的な測定によって査定される。さらに回避的レイシストは、定義上、そうした顕在的態度に示される平等主義的な価値を「心から sincerely」(Gaertner & Dovidio, 2005, p. 618)承認しているとされる。他方で、後者は潜在的態度であり、IAT のような間接的な測定によって査定されると考えられている。たとえば Dovidio, Gaertner, and Adam Pearson (2017) は、「回避的レイシズムの枠組みと一貫した」(Dovidio, Gaertner, & Pearson, 2017, p. 274)発見として、Anthony Greenwald and Linda Krieger (2006) に訴える(Dovidio, Gaertner, & Pearson, 2017, p. 274)。それによると、14000 人以上の白人アメリカ人の実験参加者のうち、約 60%はその自己報告においてアフリカ系アメリカ人にたいして偏見のない顕在的態度を示したが、人種 IAT において示される潜在的態度の約 70%はそうした顕在的態度と調和しなかった(Greenwald & Krieger, 2006, p. 958)。このことから、「意識的で顕在的で平等主義的な態度と、非意識的で潜在的な人種的バイアスが結びついた組みあわせが、回避的レイシズムの特徴である」(Dovidio, Pearson, & Penner, 2019, p. 17)という特徴づけがなされることになる。したがって、回避的レイシズムは顕在−潜在ギャップという現象にうまくあてはまると言えよう。

このことをふまえて、回避的レイシストが具体的にどのような行動パターンを示すのかに目を向けることにしよう。ここでは Gordon Hodson, Dovidio, and Gaertner (2002) のおこなった実験をとりあげる。Hodson たちは以下の 4 つのタイプの受験生を設定した。(1)試験のスコアと高校の成績が一貫して高い。(2)試験のスコアと高校の成績が一貫して低い。(3)試験のスコアは高いが、高校の成績は低い。(4)試験のスコアは低いが、高校の成績は高い。そして白人の実験参加者は、これら 4 つのタイプの受験生(ひとりは黒人で、残りは白人)の入学をどれほど強く推薦するかを決めて、さらに入学の決定にとって重要な基準を順位づけるように求められた。結果、特定のタイプの白人の実験参加者には次のような傾向が見られた。それによると、一貫して高い/低い資質をもった黒人受験生にはその資質に即した判断を下したが、他方で、黒人受験生に高い資質と低い資質が混在する条件((3)と(4))では、それぞれの条件において、黒人受験生の不利になるかたちで推薦のための基準の重みづけをシフトした。つまり、黒人受験生の試験のスコアは高いが高校の成績は低いときには前者を低く重みづけたが、他方で、黒人受験生の試験のスコアは低いが高校の成績は高いときには後者を低く重みづけたのである。

このような諸条件での行動の違いは回避的レイシズムのありようからもたらされる予測にうまくおさまると Hodson たちは主張する(Hodson, Dovidio, & Gaertner, 2002, p. 469)。それによると、回避的レイシストの行動は、そのひとが自分のおかれていることに気づく状況のありようしだいで変わるとされる(Dovidio & Gaertner, 2004, pp. 7-8)。Hodson たちの実験における(1)と(2)の条件のように、なされるべき(このばあいは受験生の入学を推薦するための)行動が明白で、それに違反することがあからさまに差別的な行動とみなされるばあい、回避的レイシストは端的にそのような行動には従事しない(Hodson, Dovidio, &

の記述に負う。

Gaertner, 2002, p. 461)。しかし(3)と(4)の条件のように、そうした基準が曖昧なばあいには、事情が複雑になる。Hodson たちは、「[回避的レイシストである]白人の黒人にたいするバイアスは、社会的に適切な規範的応答がそれほど明確に定義されておらず、ネガティブな応答が人種以外の要因にもとづいて正当化されうるときに、もっとも表明されやすい」(Hodson, Dovidio, & Gaertner, 2002, p. 469)と主張する。それゆえ、人種的マイノリティにたいするネガティブな潜在的態度が、推薦のための基準の**重みづけのシフト**の根底にあるとされることになる。しかし同時に、次節で見るように、合理化あるいは正当化そのものに焦点をあてると、平等主義的な価値を示す顕在的態度が、**非人種的な要因**にもとづいて自身の推薦にまつわる行動を合理化しようとする根底にあると考えることもできるかもしれない。

以上の議論をふまえて、本稿で扱う問いは次のように定式化される。回避的レイシストに示される顕在ー潜在ギャップはときに、潜在的態度と顕在的態度のどちらについても、それがかかわると考えられるかもしれない行動をもたらすように思われる。それでは、回避的レイシストにこうした行動の道徳的責任があるとみなすことはできるのだろうか。またそのようにみなすことができるとして、どのようなかたちでそうみなすことができるのだろうか。

3 回避的レイシストの道徳的責任

第1節で紹介した帰属主義の枠組みをふまえると、回避的レイシストにその(Hodsonたちの実験に示される)行動の道徳的責任が**ある**とすれば、ふたつのかたちの説明が考えられるだろう。(1)回避的レイシストにはその行動の道徳的責任があり、回避的レイシストの道徳的責任は、**その回避的レイシストの平等主義的な顕在的態度に示される価値をその行動が表現する**ことによってもたらされる。(2)回避的レイシストにはその行動の道徳的責任があり、回避的レイシストの道徳的責任は、**その回避的レイシストの人種的マイノリティにたいするネガティブな潜在的態度に示される価値をその行動が表現する**ことによってもたらされる。以下では、それぞれの説明の妥当性について順番に考察する。

3.1 第1の説明

第 1 の説明(回避的レイシストにはその行動の道徳的責任があり、回避的レイシストの道徳的責任は、**その回避的レイシストの平等主義的な顕在的態度に示される価値をその行動が表現する**ことによってもたらされる)を検討するうえでは、John Doris (2015) が展開した立場を下敷きにすることが有用だと思われる。そのため、まず 3.1.1 で Doris の主張を簡潔に整理したうえで、3.1.2 においてその Doris の主張を第 1 の説明に適用する。

3.1.1 合理化

まず Doris が道徳的責任にかんしてどのような立場をとるのかを明確にしよう。「典型的には、道徳的責任がともなう行動の具体例には、行為者性の行使がともなう」(Doris, 2015, p. 159)と述べるように、Doris は道徳的責任を行為者性の行使の観点から説明する。そして、そうした行為者性の行使を最終的に、**その行為者の価値の表現**に求める(Doris, 2015, p. 25)。このように、行為者性の行使として生じる行動はそ

の行為者の価値を表現するのであり、ここで私たちはその行動の道徳的責任をその行為者に負わせることができる。したがって、道徳的責任にかんする Doris の立場は帰属主義に分類されうるだろう[*8]。

この立場のもと、Doris は「**不調和** incongruence」(Doris, 2015, p. 51)という現象に注目する。不調和は顕在―潜在ギャップに類似している。つまりどちらも、二重プロセスのモデルのもとで「別々の心理的プロセスがおなじ対象にかんして相異なる出力をうみだす」(Doris, 2015, p. 51)とされる。なかでも彼が特に目を向けるのは「迂回のともなう不調和」(Doris, 2015, p. 52)である。潜在的態度／顕在的態度の区別を用いて述べると、典型的な迂回のともなう不調和では、潜在的態度は、**かりに本人が気づいていたとすれば自分の行動の理由として拒絶したようなもの**であるにもかかわらず、顕在的態度を**迂回**して(つまり顕在的態度と明示的に対立し勝利することなく)その行動をもたらすとされる(Doris, 2015, p. 52)[*9]。

迂回のともなう不調和という現象は、**反省主義** reflectivism と呼ばれる見解に懐疑論を提示する。反省主義とは「**人間の行為者性の行使は、なにを考えるべきでなにをすべきかについての自覚的な反省によって命じられる判断と行動に存する**」(Doris, 2015, p. 19)という主張であり、またそれは、「**人間の行為者性の行使には正確な反省が必要である**」(Doris, 2015, p. 19)という系と結びつけられる。この見解への Doris の批判は次のように進む(Doris, 2015, pp. 64-65)。ある行為者が迂回のともなう不調和を通じてある行動をなすとしよう。そのばあい、かりにその行為者が自分の行動の原因に気づいていたとすれば、そのひとはそれを、自分の行動の理由として拒絶していたであろう。それゆえ、反省主義によると、問題の行動は行為者性の行使ではないことになる。Doris はそうした原因を「**デフィーター** defeaters」(Doris, 2015, p. 65)と呼び、「行為者性の具体例だとされるどれについても、その行動の説明においてデフィーター(あるいは複数のデフィーター)を排除することはできない」(Doris, 2015, p. 65)と主張する。つまり、反省主義を仮定すると、私たちは、どの特定の行動についても行為者性を行使していない可能性を排除できず、それゆえその道徳的責任を問えないことになる。

この問題をふまえて Doris は反省主義を否定し、たとえ正確な自己認識に失敗する(彼はそれを「**自己の無知** self-ignorance」(Doris, 2015, p. 21)と呼ぶ)としても、私たちは行為者性を行使(それゆえ自分たちの価値を表現)し、デフィーターを排除しうると論じる。ポイントは、私たちが、自己の無知の条件にある

[*8] Doris は自身のアプローチを「価値づけ的アプローチ the valuational approach」(Doris, 2015, p. 99)と呼ぶ。

[*9] ここで言及されている気づきの種類に注意しなくてはならない。Doris 自身も簡潔に言及している(Doris, 2015, p. 61)が、Gawronski, Hofmann, & Wilbur (2006) は、態度の気づきを、「(a)人々が特定の態度の起源に気づくこと(**源泉の気づき**)、(b)人々がその態度そのものに気づくこと(**内容の気づき**)、あるいは(c)この態度が他の心理的プロセスにもたらす影響[に気づくこと](**インパクトの気づき**)」(Gawronski, Hofmann, & Wilbur, 2006, p. 486)というように、3つに区別する。潜在的態度のばあい、Hahn, Judd, Hirsh, & Blair (2014) のように、内容の気づきがあることを示すことを示す研究はあるが、Gawronski, Hofmann, & Wilbur (2006) によると、インパクトの気づきは欠けているとされる(Gawronski, Hofmann, & Wilbur, 2006, pp. 491-492:ただし紙幅の関係で検討できないが、この主張については議論の余地がある。たとえば Holroyd (2012, 2015) を参照されたい)。迂回のともなう不調和において問題となる種類の気づきは、潜在的態度が自分の行動の決定へと至るプロセスに影響することにかかわるのだから、インパクトの気づきの一種である(インパクトの気づきのありようについては、たとえば Sullivan-Bissett (2015) も参照されたい)。そしてそのような気づきが実際には欠けているかもしれないとしても、**かりに気づいていたと仮定したら**、潜在的態度が自分の行動の理由となる(自分の行動を正当化する)ことを拒絶したであろう、ということが前半部分では述べられている。この論点を明確にするよう助言した匿名の査読者に感謝する。

ときでさえ、みずからの行動を説明するにあたって、ある種の（非臨床的な）**作話**としての「**合理化** rationalizations」（Doris, 2015, p. 141）に依拠することにある[*10]。自己認識の失敗にもかかわらず、私たちは、とりわけ社会的なやりとりを円滑に進めるために、自分たちの行動を合理化する（Doris, 2015, pp. 139-141）。このようにして「説明や正当化のナラティヴを社会的にやりとりすることで、行為者の価値を表現する行動を支える足場が築かれる。これらの対話は、直接的な反省的コントロールが行き詰まるかもしれない自己の無知の条件において、行為者性の行使をもたらす」（Doris, 2015, p. 129）。つまり、正確な自己認識が欠如しているとしても、私たちは社会的な圧力のもとでの合理化を通じて自分の価値を表現する土台を創造し、それから、それを足がかりにして、**あらたに合理化が原因となることで**（Arpaly, 2018, p. 75）、行動において自分の価値を表現することができるようになるのである。このようにしてデフィーターは排除されることになる。

3.1.2 適用可能性

Doris の主張が第 1 の説明に適用されうるかどうかを考えるにあたって、まず、迂回のともなう不調和という現象が回避的レイシズムの枠組みにうまくあてはまるように見えることに注意しよう。先に見たように、迂回のともなう不調和において、潜在的態度にともなうプロセスは、かりに本人がそれに気づいていたとすれば自分の行動の理由として認めなかったであろうとしても、顕在的態度にともなうプロセスを迂回して行動をもたらす。ここで回避的レイシストの行動の原因は潜在的態度だとしてみよう。そのばあい、回避的レイシストは平等主義を心から承認する顕在的態度をとるのだから、かりに人種的マイノリティにたいするネガティブな潜在的態度にともなうプロセスに気づいていたとすれば、そうした態度を自分の行動の理由として認めることはなかったであろう[*11]。このように、自分の行動の原因でありながら、本人はそれを（かりに気づいていたとすれば）自分の行動の理由として認めないであろうことから、問題の潜在的態度はデフィーターとみなされることになる。このように考えると、回避的レイシズムの枠組みは、迂回のともなう不調和の一種とみなすことができるだろう[*12]。

しかし、Doris の論じる、合理化を通じた価値の表現という考えが、このような図式を妨げるかもしれない。Hodson たちの実験において、回避的レイシストは、自身の行動について、人種とは無関係な要因に訴える**合理化**をおこなったことを思い出そう。ここで次のような解釈が可能かもしれない。回避的レイシストは、

[*10] Doris 自身は、臨床的な事例から区別するために、非臨床的な事例には作話という言葉を用いない（Doris, 2015, p. 141）。

[*11] ここで匿名の査読者から、回避的レイシストや完全な平等主義者であっても、差別的態度が差別的行動を正当化することは認める（たとえば、かりに黒人は能力が低いのならば黒人受験生を低く評価することは合理的であることを認める）のではないか、という指摘がなされた。しかしここで問題になっているのは、差別的態度が**自分の**行動の理由となるかどうかであることに注意しなくてはならない。（1）かりに黒人は能力が低いのならば黒人受験生を低く評価することは合理的であるということと（2）黒人は能力が低いと私は考えるのだから私が黒人受験生を低く評価することは合理的であるということは異なる。そして**自分の**行動の理由となることにかかわるのは（1）ではなく、たとえば（2）のような主観的な説明である（Doris, 2015, pp. 43-44）。回避的レイシストは平等主義的な顕在的態度を抱くのだから、自分の行動の理由にかかわる主張として（2）を認めることはないだろう。

[*12] Doris 自身は回避的レイシズムという用語そのものは用いていないが、迂回のともなう不調和がその枠組みにあてはまると考えているように思われる（Doris, 2015, pp. 57-58）。

平等主義的な価値を心から信じることで、自己認識に失敗するときでさえ、合理化を通じて行動においてそうした価値を表現するための足場を築いてきた。そしてそのような足場のもと、回避的レイシストは、Hodson たちの実験において、個別の事例でそれぞれ、高校時代の学業成績や試験のスコアといった、入学の推薦のための基準をひとつとりあげて、そのような人種とは無関係の理由をもちだす。つまり、自分は人種とは関係なく平等に黒人受験生の入学の（非）推薦についての考察をしているという、自分が以前から顕在的態度として抱いていた平等主義的な価値に沿ったストーリーを個別の事例で描く。そして最終的にそれぞれの事例で、このような顕在的態度にもとづいたストーリーに示される合理化が原因となり、問題の行動（黒人受験生の入学の推薦にかんする決定）が生じたのである。このような解釈が正しいとすると、回避的レイシストはたしかに行為者性を行使しており、そのひとの顕在的態度において示される平等主義的な価値を表現していることになる。このばあい、行為者性の行使を妨げるとされたデフィーター（人種的マイノリティにたいするネガティブな潜在的態度）は、行為者性の行使をもたらす（顕在的態度に示される価値に沿った）合理化に取って代わられることになる。そして、その人々の提供する合理化（またそれを通じたそのひとが抱く価値の表現）に照らして、問題の行動にたいする道徳的責任をその人々に問うことが可能になるであろう。したがって回避的レイシストのおこなう合理化は、Doris の論じるような、迂回のともなう不調和という現象からもたらされる行為者性の行使（ひいては道徳的責任）にかんする懐疑論を避けるたぐいのものだとみなしうるかもしれない。

しかし、このような解釈からは、回避的レイシストの道徳的責任について直観的に奇妙な帰結がもたらされるように思われる[*13]。問題の回避的レイシストの行動（黒人受験生にたいする入学の推薦をめぐる決定）についてのいまの解釈だと、回避的レイシストは、自身の顕在的態度に示される平等主義的な価値に沿って合理化をおこない、そうした合理化を通じて、その価値はその回避的レイシストの行動において表現される。回避的レイシストはこのようにして行為者性を行使し、自分の価値を表現しているのだから、その回避的レイシストには自分の行動の道徳的責任があることになるとされる。しかし、なぜレイシストであるにもかかわらずその行動が平等主義的な価値を表現するという、見たところ直観に反する結論に至るのだろうか。

問題は次のようなものだと思われる。たしかに、入学をめぐる決定を**個別に**見たばあい、先に見たように、顕在的態度に示される平等主義的な価値にしたがってその決定がなされたと述べることができるかもしれない。しかし Hodson たちの実験は、そうした個別のケースを別個に考えても意味をなさない。回避的レイシズムのありようを明らかにするものとしての Hodson たちの実験が示すのは、個別のケースを**まとめてみると**実は黒人受験生に不利なかたちで入学規準がシフトしていたという、**全体の行動パターン**である。そして全体のパターンとして見ると、入学の推薦のための基準を黒人受験生に不利なかたちで決める行動が（それがなんらかの価値を表現するとして）平等主義的な価値を表現すると考えることは奇妙だと思われる。それゆえ、合理化を土台として顕在的態度に示される価値が表現されるという、Doris の示唆する一般的な図式は、すくなくとも回避的レイシストの行動（パターン）の説明としては、私たちの直観的な考えに明らかに反し、さらに Hodson たちの実験の意義を捉えていないため、ふさわしいものではないと思われる

[*13] 以下のふたつの段落にわたり論点を明確にするよう助言した匿名の査読者に感謝する。

のである。

回避的レイシストの行動にかんする私たちの直観の根底には、その行動における道徳的責任を検討するにあたって、平等主義的な価値を示す顕在的態度ではなく、人種的マイノリティにたいするネガティブな潜在的態度と問題の行動とのあいだの因果関係に焦点をあてるべきだという考えがあるように思われる。つまり、そうした直観は、回避的レイシストの顕在的態度よりむしろ潜在的態度に示される価値が問題の行動において表現されているという考えを意味するであろう。次にこの説明に向かうことにしよう。

3.2 第 2 の説明

第 2 の説明によると、回避的レイシストには（Hodson たちの実験に示される）行動の道徳的責任があり、回避的レイシストの道徳的責任は、**その回避的レイシストの人種的マイノリティにたいするネガティブな潜在的態度に示される価値をその行動が表現する**ことによってもたらされる。以下ではこの説明の妥当性を検討する。

研究者のなかには、帰属主義を潜在的態度に適用し、潜在的態度（あるいはその根底にある評価的判断）が行為者のなにがしかの価値を示し、潜在的態度が原因となる行為においてそうした価値が表現されるという考えを展開してきた人々がいる（たとえば Brownstein, 2016, 2018; Faucher, 2016; Holroyd & Kelly, 2016; Smith, 2018）。この主張にもとづいて、（Hodson たちの実験に示される）回避的レイシストの行動は潜在的態度が原因であり、その行動において人種的マイノリティにたいするネガティブな潜在的態度に示される価値が表現されている、というかたちで第 2 の説明がなされるかもしれない。そのばあい、Doris の言葉を用いると、回避的レイシストは行為者性を行使しているのだから、そのひとには自分の行動の道徳的責任があることになる。さらに先に見た第 1 の説明と異なり、この説明にしたがうと、回避的レイシストの行動が人種的マイノリティにたいするネガティブな潜在的態度に示される価値を表現することになるであろう。このことは私たちの直観的な考えとうまく合致する。

しかしこの立場にも問題があるように思われる。ここでは、個別の論者の主張をそれぞれ独立に検討するよりむしろ、この立場が一般に直面するふたつの問題を提起しながら、必要に応じて個別の主張に言及することにしたい[*14]。

3.2.1 迂回による表現

第一の問題は、先に見たように、回避的レイシズムの枠組みが、Doris の唱える迂回のともなう不調和の図式にあてはまるように思われることに関係する。潜在的態度にともなうプロセスと顕在的態度にともなうプロセスが、おなじ対象にかんして異なる結果をもたらす（顕在ー潜在ギャップ）という想定のもとで、潜在的態度にともなうプロセスが顕在的態度にともなうプロセスを迂回するかたちで行動の原因になるとき、迂回のともなう不調和という図式は成り立つ。この図式のもとでは、顕在的態度に示される価値を行為者が抱いているために、潜在的態度はデフィーター、つまり、かりに本人がそれに気づいていたとすれば、自分の行動の理由として認めなかったであろうものとみなされることになる。回避的レイシズムにあてはめる

[*14] 本文では Faucher (2016) を検討していないが、少なくとも 3.2.2 の批判はあてはまると思われる。

ならば、回避的レイシストは平等主義的な価値に顕在的態度を通じて心から関与しているのだから、かりに自分の行動の原因が人種的マイノリティにたいするネガティブな潜在的態度だとわかったとしても、それを自分の行動の理由とは認めないであろう。このように行動の理由として行為者に認められないのだから、デフィーターは、行動の原因にはなるが、行為者の価値を表現しないものだとされる。先に見たようにDoris は、合理化に訴えることで、迂回のともなう不調和という図式を避けて、さらにデフィーターの存在を排除しようとした。しかし第 2 の説明は、人種的マイノリティにたいするネガティブな潜在的態度が原因で行動が生じると考えるのだから、回避的レイシストが顕在的態度に示される平等主義的な価値を抱いている以上、迂回のともなう不調和という図式を避けることができない。したがって、この説明のために必要とされるのは、たんに、行為者は自分の潜在的態度が原因となる行動において（ときにそれと対立する顕在的態度とおなじように）その態度に示される価値を表現しうる、と主張することだけではない。それに加えて、**迂回のともなう不調和の図式のもとで**（つまり行為者が顕在的態度に示される価値を心から抱いているにもかかわらず）、なぜ潜在的態度がデフィーターでありながら、行動においてその態度に示される価値が表現されうるのかを明らかにしなくてはならない[*15]。

このふたつの課題の違いを次のように述べなおすことができるだろう[*16]。3 つの事態を考えてみよう。（1）本当の自己はときに、両立不可能な価値を示す顕在的態度と潜在的態度のそれぞれから構成されうる。（2）潜在的態度は、かりに行為者がそれに気づいていたとすれば、（それに対応する顕在的態度とは対照的に）自分の行動の理由として認めなかったであろう。（3）潜在的態度が原因となる実際の行動においてその態度に示される価値は表現される。前者の課題は、どのようにして**（1）と（3）を満たす事態**が生じうるのかを問う。それにたいして後者の課題のばあい、どのようにして**（1）から（3）のすべてを満たす事態**が生じうるのかを問う点で、前者よりも特定的なものである。

Michael Brownstein (2016, 2018) は、前者の課題に対処しうるが、後者の課題には対処できていないように思われる。Brownstein は、道徳的責任が問われるさいに行為者に帰属される（つまり本当の自己として同定される）評価的姿勢を、**ケア**という独自の心的状態として特徴づけて、「潜在的態度はケアとみなされ**うる**」（Brownstein, 2018, p. 120）と論じる。さまざまな論者がケアのありようにかんして論じているが、Brownstein は自身のケアにかんする説明を「感情的説明 affective account」（Brownstein, 2018, p. 108）と呼び、「なにかについてケアすることは、そのひとのケアの対象と連動してある特定の複雑な情動を感じる傾向性があることである」（Brownstein, 2018, p. 108）と論じる[*17]。たとえば、自分の飼っているハムスターについてケアするとは、そのハムスターのありようとつながったかたちで情動を感じる傾向性があることにほかならない。ケアにかんするこうした特徴づけのもと、前者の課題に関係するのは、Chandra Sripada (2016) の「X をケアし、Y は X と両立不可能だと信じ、そして Y をケアすることは不合理ではない」

[*15] この箇所についての匿名の査読者の助言に感謝する。

[*16] このように論点を明確にし、以降の議論を整理するよう助言した匿名の査読者に感謝する。

[*17] こうした感情的説明については、Brownstein 自身が言及するように（Brownstein, 2018, p. 108）、Jaworska (2007)、Shoemaker (2003)、Sripada (2016) の影響をうけている。ただし、 Jaworska (2007) や Shoemaker (2003) がケアを二次情動に制約するのにたいし、Brownstein はそのような制約を設けない（Brownstein, 2016, pp. 780-781; Brownstein, 2018, pp. 110-116）。

（Sripada, 2016, p. 1226）という考えからもたらされる、本当の自己にかんする「**モザイク的な**考えかた」（Sripada, 2016, p. 1225）である。この考えかたによると、本当の自己は複数の両立不可能なケアから構成されうることになる。そして Sripada は、そのほうが「人間行為者であることの豊かでときに骨の折れる複雑さ」（Sripada, 2016, p. 1226）をうまく捉えることができると考える。こうして上記の（1）が満たされることになる。またケアは特定の行動にたいして動機的な影響を行使し（Sripada, 2016, pp. 1215-1222）、そのため、複数のケアのどれもが本当の自己を構成しながらそのなかで（行動の動機としての）**優先順位がある**、あるいは、本当の自己を構成する複数のケアが**葛藤する**という事態がもたらされうる（Sripada, 2016, p. 1226; Sripada, 2016, footnote 31, p. 1221）。このことは（3）にあるような、潜在的態度が原因となる行動においてその態度に示される価値が表現される可能性を確保する。こうした本当の自己にかんするモザイク的な考えかたにもとづいてBrownsteinは「あるひとのBEIB［潜在バイアスの行動上の表現 *b*ehavioral *e*xpression of *i*mplicit *b*ias］とそのひとの反省的な平等主義的判断は、どちらもそのひとの深い自己を反映すると考えることができる」（Brownstein, 2016, p. 779）と論じる。

以上の議論は確かに、（1）と（3）を満たす事態がどのようにして生じるかを説明することができる。しかし、（1）から（3）のすべてを満たす事態がどのようにして生じるかという、迂回のともなう不調和という現象がともなう後者の課題に答えることはできない。先述のように Sripada は（3）を説明するにあたって、優先順位というメカニズムや葛藤という現象に言及する。しかし優先順位や葛藤という考えは（2）に示される考えと整合的ではない。第一に、（2）にあるように、回避的レイシストは顕在的態度に示される平等主義的な価値を心から承認するため、潜在的態度にかりに気づいていても自分の行動の理由とはみなさないであろう。したがって潜在的態度は顕在的態度よりも動機としての優先順位が高いわけではない。つまり、かりに潜在的態度が原因で行動が生じるとしても、優先順位を理由に、その態度に示される価値が行動に表現されていると述べることはできない。第二に、（2）にあるように、潜在的態度はかりに行為者がそれに気づいていたとすれば自分の行動の理由として拒絶するであろうものなのだから、その態度はなんらかのかたちで顕在的態度との葛藤に勝利することによって、行動の原因となるわけ**ではない**。じっさいDorisが述べるように、迂回のともなう不調和では「往々にして葛藤の経験がまったくない」（Doris, 2015, p. 52）。つまり潜在的態度は、そのような葛藤をともなわずに、顕在的態度を**迂回する**ことで、行動の原因となるのでなくてはならない。このふたつの論点をまとめると、迂回のともなう不調和という現象のもとでは、潜在的態度は、顕在的態度とくらべて優先順位が低いにもかかわらず、顕在的態度との対立を避けて、行動の原因となる。だからこそ、回避的レイシストは、葛藤をともなわずに平等主義的な価値を顕在的に承認しながら、同時に、人種的マイノリティへのネガティブな潜在的態度が原因となる行動にかかわるのである。それではなぜ、潜在的態度に示される価値が、顕在的態度に示される価値よりも優先順位が高いわけでもなければ、顕在的態度と対立して勝利するわけでもないのに、迂回のともなう不調和のもとで行動において表現されうるのだろうか。潜在的態度をケアとみなし、本当の自己にかんするモザイク的な考え、優先順位というメカニズム、葛藤といった現象に訴えるだけでは、この問いに答えるための説明的なリソースが不足しているように思われる。それゆえ（1）から（3）のすべてを満たす事態がどのようにして生じうるかは説明されえない。

Angela Smith もおなじ問題に苦しむように思われる。まず Smith の道徳的責任にかんする見解を紹介することからはじめよう。Smith もまた、「行為者の自身の行為と態度の責任を、それらが道徳的行為者としての**そのひとのありよう**を表現するという事実に（それが事実であるばあい）根拠づける」（Smith, 2008, p. 368）と考える点で、帰属主義者に分類される。そして Smith は、私たちの評価的判断（のあつまり）が本当の自己を構成すると考える。たとえば Smith はこう述べる。「これらの判断［評価的判断］は、ひとまとめにされると、私たちがそれを通じて世界を眺めるような、基本的な評価的枠組みを作りあげる」（Smith, 2005, p. 251）。それでは、私たちの営みはいつ本当の自己を反映する（つまり行為者に帰属しうる）のだろうか。Smith によると「ある心的状態が…［中略］…あるひとに帰属可能であるのは、その状態がそのひとの根底にある評価的判断に…［中略］…合理的につながれるばあいである」（Smith, 2005, p. 262）とされる。そのようなつながりは「特定の評価的判断を心から抱くとすれば問題の心的状態は生じる**べきである**（あるいは**べきではない**）ようなかたちで、私たちの心的状態のいくつかは特定の判断にリンクされている」（Smith, 2005, p. 253：強調引用者）というように表現される。たとえば、Smith 自身の例（Smith, 2005, p. 270）を借りると、自分は Y に不当に扱われていると X が評価的に判断するならば、原理的には Y にたいする怒りが生じる**べき**である。このようなつながりのもと、Y にたいする怒りが X に生じるとき、その怒りは X に帰属可能である（その怒りは X の本当の自己を反映する）ことになる。そして、そうした（怒りのような）態度と「私たちの評価的判断との合理的なつながりのために、その態度は、理由あるいは正当化が適切に求められうるような種類の状態である」（Smith, 2005, p. 251）と論じられる。つまり、帰属可能な心的状態にかんしては（たんなる因果ではなく正当化の観点から）「なぜ」という問いが意味をなすのでなくてはならず（Smith, 2012, footnote 6, p. 578）、そうした心的状態の根底にある評価的判断は「そのひとが直接的に道徳的に釈明可能 answerable だと私たちが考える判断」（Smith, 2005, p. 251）だとされる。Smith は以上のような見解を、**合理的関係説** rational relations view と呼ぶ。また注意すべきこととして、 Smith 自身は評価的判断と心的状態とのあいだの合理的な関係に焦点をあてるが、この関係は心的状態から行動へと拡張されうると、他の論者によってしばしば論じられる（たとえば Brownstein, 2018, p. 146; Shoemaker, 2011, p. 606）。

以上が Smith の見解の簡単な要約だが、それをふまえて、あらためて、第 2 の説明に求められる最初の課題を思い出そう。それによると、両立不可能な複数の価値をともないうる本当の自己のもと、行為者は、自分の潜在的態度が原因となる行動において、その態度に示される価値を表現しうると主張しなくてはならない。まず合理的関係説は、顕在ー潜在ギャップのように、ある個人に複数の対立する評価的判断が存在することを否定しない。Smith は、ある態度が本当の自己に帰属可能かどうかを検討するさいに「私たちは他者の明示的な自認に絶対的な権威を与えるわけではない」（Smith, 2004, p. 341）と主張し、本人が意識的に抱いているわけではなく、むしろ明示的に否認さえする態度もまた、そのひとの根底的な評価的判断を反映しうると論じる（Smith, 2004, pp. 341-343）。この論点を潜在的態度に拡張すると、潜在的態度には「私たちが正当化するよう求められうることが理解できる、そういった「合理的」評価が依然としてともなうように思われる」（Smith, 2018, p. 777）のであり、そうした評価は、たとえ本人に意識されないとしても、あるいは本人が明示的に否認するとしても、じっさいになされうることになる（Smith, 2012, footnote 10, pp.

581-582; Smith, 2018, p. 778）。したがって顕在ー潜在ギャップが生じるとき、合理的関係説によると、顕在的態度と潜在的態度の根底にあるそれぞれの評価的判断が対立することになるだろう。しかし合理的関係説では、「態度が対立するばあい、私たちは「そのひと」が本当はどこに立っているのかを決めようとする必要はないし、そうしようとすべきではない」（Smith, 2004, p. 348）、あるいは、「こうした対立する評価の両方がそのひとに正しく帰属可能である」（Smith, 2005, p. 255）とされる。こうして（1）が満たされることが示される。そして、所与の場面でどの態度がじっさいに表現されるかは、どれがじっさいにそのひとの応答のパターン（あるいは行動のパターン）を「構造化し動機づける」（Smith, 2004, p. 342）かによって定まる。このことは（3）の可能性を確保する。このようにして、合理的関係説は、第2の説明の最初の課題に対処しうるように思われる。

それでは、（1）から（3）のすべてを満たす事態がどのようにして生じうるかという、もうひとつの課題についてはどうだろうか。すなわち、迂回のともなう不調和のもとで、デフィーターとされる潜在的態度に示される価値が行動において表現されうると考えられるのだろうか。合理的関係説によると、行為者が評価的判断に顕在的に気づいている必要はないのだから、かりに問われたとして、行為者が潜在的態度（と合理的につながっている評価的判断）を自分の行動の理由として否認することじたいには問題はない。合理的関係説にとって大事なのは、本人がそれを明示的に否認するか、あるいは意識しているかどうかではなく、先に述べたように、私たちが「なぜ」と問うことが**意味をなす**かどうかであり、そのように問うことが意味をなすということは、そのひとの行動や態度と評価的判断とのあいだに合理的なつながりがあることを意味する。しかし、顕在的態度と潜在的態度がたんに相反するかたちでそれぞれ独立に存在するのではなく、後者にともなうプロセスが前者にともなうプロセスを**迂回する**という関係にあること、つまり（2）の結果として生じる事態は、そのように問うことが意味をなすということに疑いを差し挟む。回避的レイシストは、平等主義を心から主張しながら、その潜在的態度から差別的な行動に従事する。合理的関係説によると、そのひとに問題の行動の道徳的責任があるためには、その行動と評価的判断とのあいだに合理的なつながりがなくてはならない。しかし、ここでの行動と評価的判断とのつながりは直接的なものではなく、それと一貫しないもうひとつの（平等主義的な主張の根底にある）評価的判断を迂回したかたちをとる。このとき、私たちが回避的レイシストに「なぜ」と問うことは意味をなしえないように思われる[*18]。ここで、回避的レイシストの行動と潜在的態度の根底にある評価的判断は、平等主義的な顕在的態度とその根底にある評価的判断とのあいだの合理的関係をカッコに入れて、それを迂回するかたちをとることでしか、つながることができないことに注意しよう。つまり、このつながりはそれだけで独立して成立するものではない。そうした非独立的なつながりは、率直に「べき」というかたちであらわされる合理性を損なうだろう[*19]。結果として（1）から

[*18] 以下の論点を明確にするよう助言した匿名の査読者に感謝する。

[*19] Levy (2011)は、合理的関係説によると顕在ー潜在ギャップのもとで私たちには潜在的態度の責任はないことになると論じる。「私が自分の人生の何年もかけて自分の潜在的なレイシズムを根絶させようとしてきたときに、そうした自分の潜在的なレイシズムを正当化するよう私に求めることは意味をなさない」（Levy, 2011, p. 25; Brownstein, 2018, p. 147も参照）。また Smith (2018) の本文の後に付属している“NOTES AND QUESTIONS”でも類似した論点が提示されている（Smith, 2018, pp. 780-781）。ここでの私の批判は、これらとおなじ論点を異なる角度からなしていると考えられるかもしれない。

(3)をすべて満たす事態がどのようにして生じうるかは説明されえない。つまり合理的関係説は、迂回のともなう不調和のもとで、潜在的態度に起因する行動が行為者の本当の自己を表現するということを説明することができないと思われるのである。

最後に、Jules Holroyd and Daniel Kelly (2016) の主張を見ることにしよう。Holroyd と Kelly は、潜在的態度が本当の自己(ふたりの表現だと「性格ベースの評価の適切な標的」(Holroyd & Kelly, 2016, p. 123))であるためには、それがなんらかのかたちで**コントロール可能**なものなくてはならない、と主張する(Holroyd & Kelly, 2016, p. 107)。それをふまえてふたりは、Andy Clark(2007)の生態的コントロール ecological control という考えに注目する。この考えによると、私たちは、自分の行動を部分的に構成するが直接的にはコントロールすることのできない「**サブパーソナルな**構造と下位システム」(Holroyd & Kelly, 2016, p. 118)を、反復練習などで較正したり、あるいはそれが有効なかたちではたらくように外部環境を整えたりすることによって、自身の価値にしたがった行動をもたらすように間接的にコントロールすることができるとされる。Holroyd と Kelly は経験的発見をもとに、このようなかたちのコントロールが潜在的態度に適用できると論じる(Holroyd & Kelly, 2016, pp. 120-124)。したがって潜在的態度は(間接的に)コントロール可能なのだから、ふたりの主張する本当の自己とコントロール可能性のつながりをふまえると、潜在的態度はそのひとが何者であるかをあらわしうると結論されることになる(Holroyd & Kelly, 2016, p. 130)。

しかし、この議論もまた、いま考察している第二の問題を解決するには不十分だと思われる[*20]。というのも、迂回のともなう不調和をめぐる問題は、そもそも潜在的態度がコントロールできるかどうかという論点に直接的にはかかわらないからだ。Holroyd と Kelly が主張するように、潜在的態度が(顕在的態度とおなじように)なんらかのかたちでコントロール可能であることを認めることにしよう。しかしそれはせいぜい(1)と(3)を満たす事態がどのようにして生じうるかにしか答えていない。ここでもやはり問題となるのは、どのようにして(2)も同時に満たすかということである。迂回のともなう不調和のもとでは、潜在的態度は、それと対立する顕在的態度に示される価値のために、かりに本人が気づいていたら自分の行動の理由とはみなさなかったであろうものだとされる。それにもかかわらず、そのような潜在的態度が原因で行動が生じたばあい、なぜその潜在的態度はデフィーターでありながら、そこから生じる行動は本当の自己をあらわすと述べることができるのだろうか。潜在的態度がコントロール可能であると主張するだけでは、この問題は依然として残るように思われる。というのも(1)から(3)のすべてを満たす事態がどのようにして生じるかを説明するためには、Holroyd と Kelly はこの問題が(生態的)コントロール可能性と独立した論点**ではない**ことを示さなくてはならないからだ。つまり、ある態度が(Holroyd と Kelly が述べるようなかたちで間接的に)コントロール可能であるとすれば**それだけで**、かりに本人がその態度を自分の行動の理由として認めないとしても、その態度に示される価値が行動に表現されうるのに**十分である**のはなぜかを説明しなくてはならない。そして生態的コントロールという考えかただけではその役割を果たしえないように思われるのである。

3.2.2 通時的な安定性

潜在的態度と行動とのあいだの因果関係にかんする事実を無視して、潜在的態度に示される行為者

[*20] 以下の論点を明確にするよう助言した匿名の査読者に感謝する。

の価値がどのようにして行動において表現されるかという問いに取り組むことは難しい。しかし他方で、回避的レイシストが抱く人種的マイノリティにたいするネガティブな潜在的態度が原因で差別的な行動が生じることになるという因果関係に訴えるだけでは、その行動においてその価値が表現されることを示すのに十分ではない。Neil Levy (2011) は「ある態度が原因である行為が生じ、さらに**しかるべき種類のしかたで**生じるとき、その行為はその態度を表現する」(Levy, 2011, p. 248:強調引用者)また「行為の内容と行為者の態度とのあいだの関係性が偶然的なものだとすれば、その行為はこれらの態度を反映するか、あるいはその態度に(たんに)マッチするかのいずれかである」(Levy, 2011, pp. 248-249)と論じる[*21]ことで、ある行動がある態度(に示される価値)を表現することと、そうでないことを区別する。それゆえ、たんに態度と行動とのあいだで因果関係が成立しているだけでは、それが本当の自己をあらわすと述べるのに十分ではない。それでは、私たちにはいつ、行為者の態度が原因でしかるべき種類のしかたで行動が生じ、結果としてその態度に示される価値が表現されることがわかるのだろうか。

ここで帰属主義を採用する論者はしばしば、ただの因果関係に加えて、原因となる態度(に示される価値)の**通時的な安定性**に訴える。たとえば Brownstein は、本当の自己を「行為者の安定したそして同一性を根拠づける態度をあらわす機能的概念」(Brownstein, 2016, p. 767)と特徴づけたうえで、上記の Levy の指摘に答えるかたちで、「持続的で、さまざまな文脈で生じ、さらに一連の思考や感じなどにつながれてもいる行為パターン」(Brownstein, 2016, p. 775)に注目することで、私たちは行為者の価値(ケア)を推論できると主張する(Brownstein, 2016, p. 775; Brownstein 2018, p. 125)。また Smith は、自身が評価的**判断**という用語を用いる理由として、「特定の状況に特定のしかたで応答する、継続的でわりあい安定した傾向性であって、ただ一度きりの査定ではないことを明確にしたい」(Smith, 2005, footnote 27, p. 251)ということを挙げる。Doris も「価値はある欲求と結びつけられて、その欲求は、なんらかの程度の強さ、持続、究極性、非代替性を示しながら、計画において決定的な正当化の役割を果たす」(Doris, 2015, p. 28)と述べて、価値の通時的な性格を強調し、通時的に一定の行動や認知のパターンを示すことではじめて、行為者の価値は明らかにされると論じる(Doris, 2015, pp. 162-164)。そして Holroyd & Kelly (2016) もまた、 Maria Merritt (2000) が状況主義を念頭に置いて提示する、徳にかんする Hume 的な考えに頼って、性格の安定性を主張する(Holroyd & Kelly, 2016, p. 125)。その考えによると、性格の安定性は、

[*21] ここで Levy による用語の使用法について注意が必要である。Levy は、「ある態度を**表現すること**、ある態度を**反映すること**、ある態度に**マッチすること**とのあいだの区分」(Levy, 2011, p. 248) を設ける。その区分は以下のように説明される。

> ある行為がある態度を表現するのは、その態度が原因でその行為が生じ、さらにしかるべき種類のかたちで生じるばあいである。…[中略]…その行為がその行為者の態度が原因で生じるが、しかるべきかたちで生じるわけではないとすれば、それはその行為者の態度を**反映する**。その行為はその行為者の態度が原因で生じるわけではないが、それにもかかわらず、その態度が(しかるべきしかたで)生じていた**とすれば**もっていたであろう内容をその行為がもつとすれば、その行為はその態度に**マッチする**(Levy, 2011, pp. 248-249)

本稿ではこれまで、ある行動がある態度を「反映する」「表現する」「あらわす」といった言い回しを用いてきたが、ここでの Levy の論点は、そのような言い回しが上記の区分における「表現」に対応したものでなくてはならない、ということである。

周囲の社会的な関係性や環境からのサポートによる「**持続的な社会的貢献**」（Merritt, 2000, p. 374）によってしばしば達成される。そして Holroyd と Kelly は「私たちの見解は、Maria Merritt (2000) が性格にかんする Hume 的なモデルと呼ぶものにピッタリ合う」（Holroyd & Kelly, 2016, p. 125）と主張する。

したがって、いずれにせよ、かりに潜在的態度（に示される価値）が本当の自己を構成するのだとすれば、潜在的態度は（すくなくともある程度の）通時的な安定性を示さなくてはならないだろう[*22]。しかし、近年の経験的研究はこのことを否定しているように思われる。ここでは Bertram Gawronski, Mike Morrison, Curtis Phills, and Silvia Galdi (2017) のおこなった縦断的研究のひとつを紹介しよう。その実験では、約1ヶ月離れたふたつの時点で、おなじ被験者が抱いている人種（黒人と白人）にかんする潜在的態度と顕在的態度が測定された。ここで潜在的態度の測定としては、黒人の顔／白人の顔と快語（よい、快い、好ましい、すてきな、親切な）／不快語（わるい、不快な、嫌いな、嫌な、不親切な）を交互にペアにしたかたちでの人種 IAT が用いられた。他方で、顕在的態度についてはふたつの測定が用いられた。そのひとつでは、 IAT にあらわれた黒人と白人の顔にたいして抱く感じを1（とても否定的）から7（とても肯定的）までの 7 ポイントの尺度で評定するよう求められた。もうひとつでは、アフリカ系アメリカ人と白人アメリカ人が IAT で対となっている快語と不快語（たとえば、よい―わるい）を両極とする7ポイントの尺度のどこにあてはまると個人的に考えるかを示すよう求められた。そしてこうした測定をおこなった結果、異なるふたつの時点での顕在的態度は、おなじく異なるふたつの時点での潜在的態度よりも高い相関関係を示すことがわかった。換言すれば、潜在的態度は顕在的態度とくらべると時間的に不安定であることがわかったのである。こうしたことをふまえて Gawronski たちは「潜在的測定における個人差は顕在的測定における個人差ほど通時的に安定していない」（Gawronski, Morrison, Phills, & Galdi, 2017, p. 310）と結論する。さらに Gawronski たちは、それまでに出版された潜在的態度の縦断的研究をサーヴェイし、この結論をサポートすることを見いだしている（Gawronski, Morrison, Phills, & Galdi, 2017, pp. 307-308）。

このように潜在的態度（に示される価値とされるもの）は、先に見た論者たちが想定するほど通時的に安定していないように思われる[*23]。それゆえ、通時的な安定性という観点から、潜在的態度が本当の自己を構成するという主張を確立することは（すくなくとも顕在的態度よりもさらに）難しいだろう。それどころか、かりに潜在的態度が原因で差別的な行動が生じるとしても、それは、たとえばうっかりひとを傷つけてしまうばあいとおなじように、**一時的**なものであって、行為者の本当の自己を表現するわけではないとさえ考えられるかもしれない。

この潜在的態度の通時的な不安定性（テスト―再テストの信頼性の欠如）という論点にたいして、Brownstein (2018) は IAT に「文脈的制約」（Brownstein, 2018, p. 218）を加えることで通時的な安定性を高めることができると主張する（Brownstein, 2018, pp. 217-218）。しかし IAT の心理測定的な性質をどのようにして高めるかという問題と、「行為者の安定したそして同一性を根拠づける態度をあらわす機能的概

[*22] 3.2.2 における、潜在的態度には通時的な安定性が欠如しているため本当の自己を構成できないという論点、Gawronski, Morrison, Phills, & Galdi (2017) の実験結果にたいする言及、それをふまえた Brownstein への批判（ただし 3.2.2 の最終段落を除く）は、薄井 (2019) の記述に負う。

[*23] この論点について慎重な記述をするよう指摘した匿名の査読者に感謝する。

念」（Brownstein, 2016, p. 767）というBrownstein自身の提示する本当の自己の定義を潜在的態度が満たすかという問題は、直接的にはリンクしない。じっさい、Brownsteinが述べるようなかたちでIATによって測定される潜在的態度の通時的な安定性を高めることができるとしても、そのばあい、（潜在的態度によって構成される）本当の自己とされるものは文脈依存的なものとなるだろう。しかし、そうしたものを本当の自己と呼びうるのだろうか。たとえば、潜在的態度が文脈依存的であるにもかかわらず本当の自己を構成するとすれば、Brownstein の本当の自己の定義で言及されている同一性はどのように理解されるべきなのだろうか。こうした問題についてはさらなる考察の余地があるだろう。そして同様の議論は、状況的要因を考慮に入れる Holroyd & Kelly (2016) の立場にもあてはまるように思われる[*24]。

4 おわりに

以上の議論からどのような結論を引き出すことができるだろうか。ここで少なくとも3つの可能性が考えられる。それらを順番に見ることで本稿を終わることにしよう。

第一に、回避的レイシストは**その行動においていかなる価値も表現しておらず**、したがってその回避的レイシストには自身の行動の道徳的責任はない、と結論されるかもしれない。換言すれば、回避的レイシストの行動は、平等主義的な顕在的態度（に示される価値）も、人種的マイノリティにたいするネガティブな潜在的態度（に示される価値）も表現していないのである。このばあい、回避的レイシストはこれまで見てきたような独自の心的メカニズム（とりわけ迂回のともなう不調和）のために**道徳的責任をともなわないレイシスト**とみなされることになるという（常識的にレイシズムと道徳的責任とのあいだに強いつながりがあることをふまえると）いささか奇妙な考えを受け入れざるをえないだろう。

第二に、回避的レイシストに道徳的責任があるという考えを維持するならば、これまでの議論は**帰属主義という道徳的責任の理論の限界を示す**ものだと考えられるかもしれない[*25]。第 1 節で述べたように、帰属可能性をめぐる問いは、道徳的責任の概念の分析において他の立場に関連する問いに先行するとされた（Fischer & Tognazzini, 2011, p. 382）。そのことをふまえると、顕在ー潜在ギャップの道徳的責任を考えるうえで帰属主義に焦点を絞ることは、議論の**出発点**としては有用であろう。しかし概念の分析において先行することは、かならずしも他の道徳的責任の理論からもたらされる結論を制約するわけではない（Fischer & Tognazzini, 2011, p. 382）。それゆえ本稿の議論は、顕在ー潜在ギャップをめぐる道徳的責任の問題における帰属主義の限界を示すと同時に、他の道徳的責任の理論を適用する可能性に目を向けることを示唆するものだと考えられるかもしれない。

第三に、顕在ー潜在ギャップの存在は**帰属主義そのもののありかたを捉え直す**ことを私たちに促すと考えられるかもしれない。これまで見てきたように、帰属主義が回避的レイシストに道徳的責任を帰属する

[*24] 匿名の査読者から、文脈依存的なのは潜在的態度（そのもの）ではなくその発現のしかたにすぎないのではないか、という指摘がなされた。しかし、通時的に不安定なのが潜在的態度の発現だとしても、（行動において表現される）本当の自己が通時的に安定していると想定するかぎり、潜在的態度が本当の自己を構成することは難しいという論旨は変わらないと思われる。

[*25] この論点を指摘した匿名の査読者に感謝する。なお「限界」という表現は匿名の査読者によるものである。

さいの大きな障害のひとつは、本当の自己とデフィーターの境界設定にある。たとえば、従来の帰属主義は本当の自己をあまりに厳しく理解しているために、潜在的態度が本当の自己を構成する余地がなくなってしまったと論じることができるかもしれない。したがって帰属主義を捉え直すひとつのやりかたとしては、本当の自己という考えを標準的な帰属主義が想定するよりも緩やかに捉えて、迂回のともなう不調和という現象のもとでも道徳的責任の帰属が（帰属主義の枠組みのなかで）可能であるように再定式化することかもしれない[*26]。

このように本稿の議論からは以上の3つの結論に至る可能性が示唆される。とはいえ、いずれの結論を検討するにせよそれじたいが大きな課題であり、ここでは今後の研究方針として素描するにとどめて、稿をあらためて論じることにしたい。

付記

本稿の一部は、2022 年度哲学若手研究者フォーラム（2022 年 7 月 23 日、オンライン）においておこなわれたワークショップ「行為・知識・責任：法と哲学の対話」のなかでの発表「無知と作話」にもとづく。

謝辞

2 名の匿名の査読者の指摘・助言に感謝する。本研究は JSPS 科研費 JP21K00031 の助成を受けたものである。

参考文献

[1] Arpaly, N. (2018). Comments on *Talking to Our Selves* by John Doris. *Philosophy and Phenomenological Research, 97*(3), 753–757.

[2] Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology, 7*(4), 765–786.

[3] ———. (2018). *The Implicit Mind: Cognitive Architecture, the Self, and Ethics.* Oxford University Press.

[4] Clark, A. (2007). Soft selves and ecological control. In D. Ross, D. Spurrett, H. Kincaid, & G. Stephens (Eds.), *Distributed Cognition and the Will: Individual Volition and Social Context* (pp. 101–122). MIT Press.

[5] Doris, J. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency.* Oxford University Press.

[6] Dovidio, J., & Gaertner, S. (2004). Aversive racism. *Advances in Experimental Social Psychology,*

[*26] 潜在的態度の道徳的責任という観点から本当の自己という考えを捉え直す試みとして、たとえば薄井 (2019) を挙げることができる。

36, 1–52.

[7] Dovidio, J., Gaertner, S., & Pearson, A. (2017). Aversive racism and contemporary bias. In C. Sibley & F. Barlow (Eds.), *The Cambridge Handbook of the Psychology of Prejudice* (pp. 267–294). Cambridge University Press.

[8] ———., Pearson, A., & Penner, L. (2019). Aversive racism, implicit bias, and microaggressions. In G. Torino, D. Rivera, C. Capodilupo, K. Nadal, & D. Sue (Eds.), *Microaggression Theory: Influence and Implications* (pp. 16–31). Wiley.

[9] Faucher, L. (2016). Revisionism and moral responsibility for implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy vol.2* (pp. 115–134). Oxford University Press.

[10] Fazio, R. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology, 23*, 75–109.

[11] ———., & Olson, M. (2014). The MODE model: Attitude-behavior processes as a function of motivation and opportunity. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process Theories of the Social Mind* (pp. 155–171). Guilford Press.

[12] ———., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual-Process Theories in Social Psychology* (pp. 97–116). Guilford Press.

[13] Fischer, J. M., & Tognazzini, N. A. (2011). The physiognomy of responsibility. *Philosophy and Phenomenological Research*, *82*(2), 381–417.

[14] Gaertner, S., & Dovidio, J. (2005). Understanding and addressing contemporary racism: From aversive racism to the common ingroup identity model. *Journal of Social Issues*, *61*(3), 615–639.

[15] Gawronski, B., & Bodenhausen, G. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731.

[16] ———., & ———. (2011). The associative–propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44*, 59–127.

[17] ———., & ———. (2014). The associative-propositional evaluation mode: Operating principles and operating conditions of evaluation. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process Theories of the Social Mind* (pp. 188–203). Guilford Press.

[18] ———., Hofmann, W., & Wilbur, C. (2006). Are “implicit” attitudes unconscious? *Consciousness and Cognition*, *15*(3), 485–499.

[19] ———., Morrison, M., Phills, C., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300–312.

[20] Greenwald, A., & Krieger, L. (2006). Implicit bias: Scientific foundations. *California Law Review, 94*(4), 945–967.

[21] ———., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.

[22] Hahn, A., Judd, C., Hirsh, H., & Blair, I. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*(3), 1369–1392.

[23] Hodson, G., Dovidio, J., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin, 28*(4), 460–472.

[24] Holroyd, J. (2012). Responsibility for Implicit Bias. *Journal of Social Philosophy, 43*(3), 274–306.

[25] ———. (2015). Implicit bias, awareness and imperfect cognitions. *Consciousness and Cognition, 33*, 511–523.

[26] ———., & Kelly, D. (2016). Implicit bias, character, and control. In A. Masala & J. Webber (Eds.), *From Personality to Virtue: Essays on the Philosophy of Character* (pp. 106–132). Oxford University Press.

[27] Jaworska, A. (2007). Caring and internality. *Philosophy and Phenomenological Research, 74*(3), 529–568.

[28] Levy, N. (2005). The good, the bad and the blameworthy. *Journal of Ethics and Social Philosophy, 1*(2), 2–16.

[29] ———. (2011). Expressing who we are: Moral responsibility and awareness of our reasons for action. *Analytic Philosophy, 52*(4), 243–261.

[30] Merritt, M. (2000). Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice, 3*(4), 365–383.

[31] Olson, M., & Fazio, R. (2009). Implicit and explicit measures of attitudes: The perspective of the MODE model. In R. Petty, R. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the New Implicit Measures* (pp. 19–62). Psychology Press.

[32] Payne, B., Cheng, C., Govorun, O., & Stewart, B. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277–293.

[33] Shoemaker, D. (2003). Caring, identification, and agency. *Ethics, 114*, 88–118.

[34] ———. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics, 121*, 602–632.

[35] Smith, A. (2004). Conflicting attitudes, moral agency, and conceptions of the self. *Philosophical Topics, 32*(1/2), 331–352.

[36] ———. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics, 115*, 236–271.

[37] ———. (2008). Control, responsibility, and moral assessment. *Philosophical Studies, 138*(3), 367–392.

[38] ———. (2012). Attributability, answerability, and accountability: In defense of a unified account. *Ethics*, *122*(3), 575–589.

[39] ———. (2018). Implicit bias, moral agency, and moral responsibility. In G. Rosen, A. Byrne, J. Cohen, E. Harman, & S. Shiffrin (Eds.), *The Norton Introduction to Philosophy* (2nd ed., pp. 773–381). W. W. Norton and Company.

[40] Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*, *173*(5), 1203–1232.

[41] Sullivan-Bissett, E. (2015). Implicit bias, confabulation, and epistemic innocence. *Consciousness and Cognition*, *33*, 548–560.

[42] Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*, 227-248.

[43] 薄井尚樹. (2019).「潜在的態度は「本当の私」なのか」『倫理学年報』68, 201-214.

[44] ———. (2020).「ロボット／AI は差別の対象となりうるのか」『人文論叢：三重大学人文学部文化学科研究紀要』37, 15-26.

著者情報

薄井尚樹（関西大学）