表象・説明・道徳的責任
―具体-抽象のパズルをめぐって―*

稲荷森輝一

概要

This study aimed to present a novel explanation for the perplexing nature of moral intuition regarding abstract and concrete descriptions of deterministic actions. Nichols and Knobe (2007) have found that, when deterministic actions are described concretely, most people exhibit compatibilist responses. Conversely, when deterministic actions are described abstractly, most people exhibit incompatibilist responses. The prevailing explanation for this phenomenon is that one of the two responses is an error. However, this study argues that this phenomenon can be understood as a consequence of normal operation of the mechanism that generates moral intuition. The focus is on Björnsson's Explanation Hypothesis (Björnsson & Persson, 2012, 2013; Björnsson, 2015), which posits that differences in explanations provided for deterministic actions in different cases account for the discrepancy in intuition between concrete and abstract cases. This study reinterpreted this proposal by focusing on how actions were represented in concrete and abstract cases and presented a new theory, which is the explanation-representation hypothesis. This theory accounts for both compatibilist and incompatibilist intuition without relying on error accounts.

Keywords: 直観、自由意志、道徳的責任、道徳心理学、実験哲学

1 はじめに

人間は自由意志をもつのだろうか？古代から続いてきたこの哲学的難問は、今や心理学的な問題としての色を帯びてきている。現代自由意志論では、決定論と自由意志（とりわけ道徳的責任に必要な自由意志）との両立可能性が主な争点となってきた。両立論によれば両者は両立し、非両立論によれば両者

* *CAP* Vol. 15 (2024) pp. 110-134. Submitted: 2023.6.8. Accepted: 2024.3.8. Category: 原著論文. Published: 2024.4.10.

は両立しない。本論争においては、直観が重要な役割を果たしてきたと考えられる[*1]。なぜならこの論争では、私たちは決定論的世界の行為者に自由意志・道徳的責任の帰属を認めるような直観（両立論的直観）を有しているのか、それとも決定論的世界の行為者に自由意志・道徳的責任の帰属を認めないような直観（非両立論的直観）を有しているのかが争点の一つとなってきたからだ。たとえばFrankfurt(1969)は、他行為可能性が排除された状況下でも直観的に道徳的責任の帰属が認められる事例（フランクファート型事例）を提示することで両立論の直観的正当化を試みた。これに対し非両立論者は、非両立論の直観性を主張してきた。たとえば代表的な非両立論者である Pereboom (2001; 2014) は、決定論的状況と外的操作によって免責される状況との間にアナロジーが成立することを示す思考実験、「操作論証」を通じて、非両立論の直観性を主張している[*2]。近年この論争は、「実験哲学」と呼ばれる領域にフィールドを拡げてきている。実験哲学とは、質問紙調査などの実験心理学的方法を応用し、哲学的問題に関する人々の直観を調べる研究手法である。自由意志の実験哲学では、決定論的世界を描写したシナリオを与え、決定論に関する人々の直観を調べるといった研究が数多く行われてきた。両立論的直観と非両立論的直観をめぐる論争は、今や経験的な問題としての側面を強めている。

では、実験哲学は直観に関する両立論と非両立論のいずれを支持しているのだろうか。この分野の代表的研究のひとつである Nichols and Knobe (2007) の実験は、決定論的な仕方で生み出された行為（以後これを「決定論的行為」とよぶ）に対する我々の直観が、具体事例と抽象事例において大きく異なることを明らかにした。決定論的宇宙における具体的な非道徳的行為の描写を与えた具体条件では実験参加者の大半が両立論的直観を示したのに対し、具体的な行為を提示せずに責任帰属の可否を尋ねた抽象条件ではほとんどが非両立論的直観を示したのである。

本論文はこの「具体-抽象のパズル」に焦点を当てる。この事象の説明としては、後述する「パフォーマンスエラー説」(Nichols & Knobe 2007) や「バイパス説」(Nahmias & Murray 2011; Murray & Nahmias 2014) をはじめ、具体事例と抽象事例のいずれか一方における回答をエラーとして理解する仮説が示されてきた。パフォーマンスエラー説によれば、具体事例における両立論的直観は、感情的反応というエラー要因が直観産出メカニズムの正常な動作を妨げた結果生じたものである。対してバイパス説によれば、抽象事例における非両立論的回答の多くは、決定論の誤解に起因した見かけ上のものに過ぎないとされる。

こうしたエラー説に対し、本論文は Björnsson の「説明説 (Explanation Hypothesis)」 (Björnsson &

[*1] この点については Cappelen (2012) のように、そもそも哲学者が直観を用いてきたという前提それ自体を否定する議論もあるが、本稿では（少なくとも）自由意志の哲学においては直観が用いられてきたという前提のもと議論を進める。

[*2] Peerboom の操作論証に関する日本語文献としては、稲荷森 (2021) など。

Persson 2012; 2013; Björnsson 2015) に基づく説明に新たな解釈を与えることで、具体-抽象のパズルが直観産出メカニズムの正常な機能の帰結として理解できることを主張する。説明説によれば、具体-抽象のパズルは、具体事例と抽象事例において決定論的行為に適用される行為の説明が異なることに起因しているという。では、このように事例の具体性が行為の説明に影響することで直観の差異をもたらすことは、道徳直観産出メカニズムにとってエラーと言うべきなのだろうか。本論文はこの部分に着目し、説明説が提案する具体-抽象間の直観産出プロセスの理解を表象の差異という観点から再分析することを通じて、二つのプロセスはいずれもエラーを含まないものとして理解できることを明らかにする。

本解釈（説明表象説）によれば、決定論と責任帰属をめぐる具体と抽象のパズルは、事例間における**行為の表象**のされ方の差異に応じて、異なる対象に因果が帰属される仕方で決定論的行為が**概念化**されることに起因している。ここで言う「行為の表象」とは、実験参加者に提示されるシナリオのワーディングなどを意味する。これに対して「表象に応じた（行為の）概念化」とは、そうした表象を提示されたときに参加者の認知メカニズムに生じる表象状態を意味する。行為の表象および表象に応じた概念化はいずれも道徳直観産出にとって必要不可欠であり、直観産出メカニズムにとって内的な要因であるから、具体事例における両立論的直観・抽象事例における非両立論的直観は、いずれも直観産出プロセスにとって外的な要因の影響（エラー）に訴えることなく説明することができる。

本稿は以下の流れで議論を展開する。まず、Nichols and Knobe (2007) の研究内容を簡単に紹介したうえで、具体-抽象のパズルをめぐって考えられうる二つの可能性を整理する (2 節)。上述の通り、彼らは具体事例において両立論的回答が大半を占める一方、抽象事例においては非両立論的回答が大半を占めることを発見した。3 節では、こうした具体-抽象のパズルに関する説明説の理解を概観する。その上で、説明説における具体-抽象間での直観産出プロセスの理解に新たな解釈（説明表象説）を与えることで、具体-抽象における直観の差異は、直観産出メカニズムの正常な機能の結果として理解できることを明らかにする 。最後に、既存の仮説である「パフォーマンスエラー説」(Nichols & Knobe 2007)・「バイパス説」(Nahmias & Murray 2011; Murray & Nahmias 2014) と説明表象説の相違点および本説の優位性を明らかにする (4 節)。

2 具体−抽象のパズル

本節では Nichols and Knobe (2007) が行った実験、および彼らが発見した具体-抽象のパズルについて説明する。その上で、このパズルについて二つの可能な解釈を提示する。

Nichols and Knobe (2007) の内容を説明する前に、Nichols and Knobe (2007) 以前に行われた有名な実験哲学的研究である Nahmias et al. (2005; 2006) の内容について触れておきたい。Nahmias らは以下

の決定論的シナリオを用い、人々の直観を調査した。被験者は自由意志について学んだことのない学部生であった。

> シナリオ：次の世紀、私たちがすべての自然法則を発見し、これらの自然法則と世界のすべてのものの現状から、将来のどの時点でも世界で何が起こるかを正確に推論できるスーパーコンピュータをつくったとします。このスーパーコンピュータは、世界のあり方に関するあらゆることを調べ、それがどのようになるかを 100%の精度で予測することができます。このようなスーパーコンピュータが存在し、ジェレミー・ホールが生まれる 20 年前の西暦 2150 年 3 月 25 日のある時刻における宇宙の状態を調べたとしましょう。するとコンピュータは、この情報と自然法則から、ジェレミーは間違いなく 2195 年 1 月 26 日午後 6 時にフィデリティ銀行を襲うと推論します。いつものように、スーパーコンピュータの予測は正しく、ジェレミーは 2195 年 1 月 26 日午後 6 時にフィデリティ銀行を襲います。(Nahmias et al. 2005, 566)

なお、この実験のパイロット実験では、被験者の一部が「ジェレミーは自由意志を有しているからこのシナリオは不可能である」と考える傾向にあることが明らかになっていた。本実験では、この思考が課題文への回答に影響することを防ぐため、質問 1「このシナリオが可能だと思うかどうか」に解答させたうえで以下の文言[*3]を付け加え、決定論的な宇宙における善い行い・悪い行いのそれぞれについて責任帰属直観を調べた (Nahmias et al. 2005, 566-568)。

> 質問 1 にどう答えたかにかかわらず、そのようなスーパーコンピュータが実際に存在し、ジェレミーが銀行を襲うことを含めて、実際に未来を予測できたとします（ジェレミーはその予測について知らないとします）。
>
> ジェレミーが銀行を襲ったとき、彼はそのことで道徳的に非難されるべきだと思いますか？
>
> ジェレミーが子供を助けたとき、彼はそのことで道徳的に賞賛されるべきだと思いますか？

結果、否定的な事例では 83％がジェレミーは非難に値すると判断し、肯定的な事例では、88％が賞賛に

[*3] その後行われた多くの実験哲学的研究においても、実験参加者には同様の指示が与えられることが一般的となっている。

値すると判断した。つまり、大部分の人々の直観的判断は両立論的であった。

ここで注目してもらいたいのは、Nahmias らの実験においては、「ジェレミーが銀行を襲った」・「ジェレミーが子供を助けた」という具体的な行為が提示されているという点である。一方、Nichols and Knobe (2007) は、Nahmias らが用いたのとは異なる仕方で決定論を描写した課題文を用いたうえで、具体条件と抽象条件の二つの条件で参加者の直観を調べた。実験参加者には、条件に応じて以下の課題文が提示された (Nichols & Knobe 2007, 669-70)。

> ある宇宙（宇宙 A）を想像してみてください。この宇宙で起こることはすべて、その前に起こったことのみに起因しています。これは宇宙の始まりから言えることで、宇宙の始まりに起こったことが次に起こったことを引き起こし、それが現在に至るまで続いているのです。たとえば、ある日ジョンは昼食にフライドポテトを食べることに決めました。この決断も、他のすべてのものと同様に、その前に起こったことのみに起因していたことになります。だから、もしこの宇宙のすべてが、ジョンが決断するまで全く同じであったとしたら、ジョンがフライドポテトを食べると決断することは、起こらざるを得ないことであったということです。
>
> 次に、起こることのほとんどすべてが、その前に起こったことのみに起因している宇宙（宇宙 B）を想像してみてください。ただし、人間の意志決定だけは例外です。たとえば、ある日メアリーが昼食にフライドポテトを食べると決めたとします。この宇宙では、人の決断はその前に起こったことのみに起因するわけではないので、たとえメアリーが決断するまで宇宙のすべてがまったく同じだったとしても、メアリーがフライドポテトを食べようと決めることが起こらざるを得なかったわけではありません。彼女は何か別のことを決断することができたのです。
>
> つまり、宇宙 A では、すべての決断は、その決断の前に起こったことのみに起因しており、過去を踏まえると、それぞれの決断はそのようにならざるを得ないということです。それに対して、宇宙 B では、決断は過去のみに起因するものではなく、人間の各決断が実際に起こる通りに起こらざるを得ないということはありません。
>
> 具体条件：宇宙 A（決定論的宇宙）では、ビルという男性が秘書に惹かれるようになり、彼女と一緒になるためには、妻と3 人の子供を殺すしかないと考える。ビルは、火事になっても自分の家から逃げられないことを知っている。出張に出かける前に、彼は地下室に装置を設置して家を燃や

し、家族を殺してしまう。ビルは妻子を殺したことについて、道徳的に完全な責任を負うだろうか？

YES NO

抽象条件：宇宙 A（決定論的宇宙）では、人は自分の行動に完全な道徳的責任を負うことは可能だろうか？

YES NO

抽象的シナリオを提示された群は大多数(86%)が非両立論的回答（NO）をしたが、具体的シナリオを提示された群は大多数(72%)が両立論的回答（YES）をした。つまり、人々は具体条件では両立論的直観を示したものの、抽象条件では非両立論的直観を示したのである。

Feltz and Cova (2014) は、Nichols and Knobe (2007)、Nahmias et al. (2005) および Nahmias et al. (2007) で用いられたシナリオ、並びにこれらを改変したシナリオを用いた諸研究のメタ分析を行った。その結果、抽象事例と具体事例における直観の差異は堅固であることが明らかとなった。加えて、具体事例では両立論的直観が、抽象事例では非両立論的直観が過半数を超えるという傾向は、文化横断的な調査においても確認されている。たとえば Sarkisian et al.(2010) は、インド、香港、コロンビア、アメリカの四地域において、人々の直観を調査した。実験参加者は Nichols and Knobe (2007) の実験と同様の決定論的世界に関する説明および抽象シナリオを読んだ上で、決定論的世界における責任帰属の可否を判断した。結果、アメリカでは 75%、インドでは 72%、香港では 63%、コロンビアでは 68%の人々が、決定論的世界と道徳的責任は両立しないと回答した。つまり、いずれの地域においても、大半の人々の直観は非両立論的であることが示された。一方、20 ヵ国を対象としてより大規模に文化横断的調査を行った Hannikainen et al. (2019) は、Nichols and Knobe (2007) におけるシナリオとほぼ同じ内容の課題文を用い、具体的事例における人々の直観的判断を調査した。なお、本実験では行為者が非難に対するか/罰に値するかに関して七点スケールで回答を求めた。結果、責任帰属に関して、人々の直観は概して両立論的であるという結果が得られた。

このように自由意志の実験哲学では、決定論的世界における行為を具体的に描写した場合には両立論的直観が増加し、行為を抽象的に描写した場合には非両立論的直観が増加することが明らかとなっている。なぜこのようなパラドクシカルな現象が生じるのだろうか？この現象については、大きく分けて以下二つの理解が可能である。

H1 具体事例で増加する両立論的直観・抽象事例で増加する非両立論的直観のうち、少なくともいずれか一方は、何らかのエラーに起因している。

H2 具体事例で増加する両立論的直観・抽象事例で増加する非両立論的直観は、いずれもエラーに起因するものではない。

H1 には、4 節で詳述する「パフオーマンスエラー説」(Nichols & Knobe 2007) や、「バイパス説」(Nahmias & Murray 2011; Murray & Nahmias 2014) が分類される。前者によれば両立論的回答が、後者によれば非両立論的回答がエラーに起因するものであるとして棄却される。一方本論文では、H2 を支持する理論として、説明説の新解釈（説明表象説）を提案する。

本題に入る前に、本稿が問題とする「エラー」の範囲を明確にしておきたい。本稿では、具体-抽象のパズルをめぐって議論されてきた二種類のエラーに焦点を当てる。第一のエラーは、直観産出プロセスにおける、直観産出メカニズムにとって外的な要因の影響である。このエラーを論じるにあたっては、直観産出メカニズムにとって内的要因/外的要因とは何であるかを定義する必要がある。内的要因とは、直観産出メカニズムを構成する要因、つまり直観産出に必要な要因を意味する。たとえば、責任帰属の基準に関する心的表象は、責任帰属直観を産出する心的メカニズムにとって必要不可欠であり、それゆえこのメカニズムにとって内的である。これに対し外的要因とは、直観産出メカニズムに影響しうるにせよ、そのメカニズムにとって不必要な要因のことを意味する。たとえば、疲労感はしばしば道徳判断の厳しさに影響するかもしれないが、私たちは一切疲労を感じていないときであっても道徳判断を下すことができる。ゆえに、疲労感は道徳直観を産出する心的メカニズムにとって不必要であり、外的要因であると考えられる。また、May (2016) など一部の論者によれば、嫌悪感情なども道徳直観にとって心理学的に無関連な外的要因であるとされる。

第二のエラーは、ある問題にかんする直観産出に必要な条件の欠如である。この種のエラーには、思考実験そのものを正しく理解できていない、といった事態が含まれる。詳しくは4節で扱うが、Nahmias and Murray (2011) および Murray and Nahmias (2014) によれば、少なからぬ人々は決定論的シナリオを正しく理解することに失敗してしまうという。決定論的世界の記述を与えられた人の多くは、決定論的世界では人々の心的状態が因果的効力をもたない、といった誤解を犯してしてしまうのだ。また本稿では取り上げないが、Nadelhoffer et al. (2020) など一部の研究によれば、実験参加者の多くは決定論が別-可能性を排除することを理解できていない可能性がある。決定論が何たるかを理解していなければ、そもそも決定論的行為に関する直観をもつことはできない。こうした思考実験の理解エラーは、直観産出メカニズム

の働きを妨げるというよりは、決定論のような哲学的問題について直観産出メカニズムが機能することそれ自体を不可能にするという意味で、直観産出に必要な条件の欠如をもたらす。

なお、上で挙げた「内的」と「外的」の区別は、あくまで道徳直観を産出する一連の心理学的プロセスに必要か否か、という観点からなされるものである。したがって、直観を産出する主体の内部にある要因がすべて内的要因としてカウントされるわけではない。上記で挙げたように、嫌悪感情などは主体内部の要素でこそあれ、直観産出プロセスにとって外的な要因に分類されうる。同様に、主体外部の要因であるからといって、必ずしも外的要因に分類されるわけではない。たとえば、直観の対象となる行為が提示されること、つまり行為の表象が与えられることは主体外部の要素であるが、直観の産出にとって不可欠な要因である。したがって、本論文の定義に照らすと、行為の表象は内的要因にカウントされることになる。

また、本稿で扱う上記二つのエラーには、客観的な道徳的真理をトラックすることに失敗するといった意味での「エラー」は含まれない。道徳心理学における一部の論者は、行為のフレーミングや提示順序、および嫌悪感などの感情は道徳的真理にとって無関係であり、そうした無関連要因に影響される道徳直観は信頼性を欠くと主張している (e.g. Sinnott-Armstrong 2008)。次節では具体-抽象間におけるフレーミング効果が第一の意味でエラー（直観産出プロセスにおける外的要因の影響）であるかを検討していくが、道徳直観の信頼性を損なうという意味でエラーであるかについては論じない。なぜなら 1 節で述べたように、自由意志の哲学では人々がいかなる直観を有しているのか、という問題それ自体が大きな係争点となっているからだ。ある道徳直観が信頼できるかどうかが重要な問題であることは間違いないが、自由意志論においては人々の直観を明らかにすることそれ自体が大きな哲学的価値を有していると考えられる。

3 エラーに訴えない説明

本節では、Björnsson の「説明説 (Explanation Hypothesis)」 (Björnsson &Persson 2012; 2013; Björnsson 2015) における議論に新たな解釈を与えることで、具体-抽象のパズルが直観産出メカニズムの正常な動作の帰結として理解できることを論じる。本解釈によれば、具体事例と抽象事例で責任帰属直観の差異が生じるのは、具体-抽象で行為が異なる仕方で表象されているのに応じ、行為が異なる仕方で概念化され、それらに（同一の）責任帰属の基準が適用されるからである。説明説によれば、具体事例と抽象事例における直観の差異は、各事例において決定論的行為に異なる説明が適用されることに起因している。しかし後に論じるように、説明説による説明はいずれかの直観がエラーに起因しているという可能性を完全に排除するものではない。つまり Björnsson の議論だけでは、上で挙げた H1/H2 のうちいずれが妥当であるかはなおも明らかでない。本稿の目的はこの論点を掘り下げ、H2 を明確に支持する仕方で説明説の提案を再解釈することにある。以降、まずは「説明説」について説明する (3.1)。そのうえで、

具体事例と抽象事例で異なる行為の説明が適用されるとする説明説の議論を、行為の表象と因果帰属という視点から分析し、H2 を明確に支持する理論として「説明表象説」を提案する (3.2)。

3.1 説明説

説明説は、我々の責任帰属直観に関する一般的な説明を与える理論であり、以下のように定式化される。

> 私たちは、[行為者] A に関連する何らかの動機付け構造を[行為] X の重要な一般的説明の（一部）と見なす場合、[行為者] A は[行為] X に責任があると見なす。 (Björnsson 2015, 6、[ ]内筆者)

つまり、ある行為者の行為が、「一般に責任追及の適切な対象となる種類の動機づけ構造」(Björnsson 2015, 5)によって説明される対象と見なされる場合、私たちはその行為者 A に責任を帰属するということだ。

説明説によれば、他者によるコントロール、脅迫など外部からの圧力、無知などの要因が免責事由となるのは、これらが動機づけ構造に代わって行為に対する重要な説明を与えるものであるからだ。たとえば私が友人との約束を破ったとしよう。それが私の怠慢さや自己中心的性格に起因する場合、私は約束を破ったことについて責任を問われるだろう。なぜなら、そうした私の振る舞いはまさに、私の行為を傾向づける性格特性、すなわち「動機づけ構造」 (motivational structure) によって説明されるものと見なされるからだ。一方、私が誰かに脅迫されていたとか、あるいは何らかの記憶障害ゆえに約束を破ってしまったという場合、私はそのことに対して責任を問われないだろう。なぜならその場合、私の動機づけ構造は約束違反という私の振る舞いを説明するにあたり、さしたる重要性をもたないからだ。

注目すべきは、責任が帰属されるためには単に動機づけ構造が説明上関連するというだけでは不十分であり、それが「重要な説明」を与えるものでなくてはならないということだ (Björnsson &Persson 2013, 11-14)。我々がものごとに対して与える説明は選択的 (selective)である。私たちは事象を説明するとき、人々の関心を引き、説明への関心 (explanatory interest) を十分に満たす要因を挙げる。たとえば、家が燃えているというとき、私たちはその原因としていくつもの要因を挙げることができるが、基本的に家が可燃性であったとか、その周辺に酸素があったといった要因を挙げたりはしない。もちろんそういった要因は家が燃えるという事象に至る因果的連関を構成しており、説明上関連しているにせよ、重要な説明を与えるものではない。むしろ私たちは、「家が雷で打たれたから」といった特定の要因を原因として挙げるだろう。なぜなら、落雷は「家が可燃性である」・「周囲に酸素がある」といった当然の前提とは異なるがゆえ

に、人々の注意を引き付けるものであるからだ。また、私たちは家が燃えた原因として「近くの大気が正負の電荷に分離したから」といった要因を挙げたりもしない。そのような説明を与えられても、我々は「なぜそれが火災を引き起こすのか？」と更に問うことになるだろう。事象の説明は、新たな問いを生むことなく、私たちが求める説明への関心を十分に満たすものでなくてはならないのだ。 ゆえに、何が重要な説明と見なされるかはその人の説明への関心に依存する。たとえば建築士であれば、家が雷で打たれたということを前提したうえで、「（雷に打たれたことで）この家が燃えた」という事象を説明してくれる要因を求め、電気配線や断熱材の素材によって火災を説明するかもしれない。

具体-抽象のパズルに戻ろう。説明説は、具体事例と抽象事例における直観の差異を、私たちが採用する説明のフレームワークの違いという視点から説明することができる (Björnsson & Persson 2013, 26-30; Björnsson 2015, 101-2)。私たちは基本的に、人々の振る舞いに民間心理学的説明を適用し、それらを道徳的責任の帰属対象として認識している。しかし、抽象シナリオにおいては、世界が決定論的であるという前提が導入されることで、欲求や信念、熟慮といった民間心理学的要因が行為の説明としての特権的地位を喪失する。行為者の動機づけ構造は、ある特定の行為を引き起こす一連の因果連関における一要素に格下げされ、その行為が生じたことに関する重要な説明とは見なされなくなるのだ。ゆえに、決定論的世界での振る舞いは道徳的責任の帰属対象から除外される。一方、具体的な行為の内容や行為者の心的状態が明示的に記述された具体事例においては、決定論的シナリオが導入されてもなお、民間心理学的なモデルで行為を説明することが容易である。ゆえに、行為者の動機づけ構造が行為の説明として重要な役割を果たす具体事例では、決定論的行為に対する責任帰属が認められることになる。このように説明説は、具体事例と抽象事例において適用される説明の差異という観点から、両事例間における直観の差異を説明することができる。

3.2 説明説

具体と抽象のパズルに関する上記の説明は以下のように言い換えられる。つまり、具体事例と抽象事例では決定論的行為が異なる仕方で概念化されているがゆえに、それらに同一の責任帰属基準が適用された結果、両事例間で異なる責任帰属直観が導かれるというわけだ。抽象事例における決定論的行為は、まさに決定論的な因果の連鎖によって説明されるものとして理解されるのに対し、具体事例における決定論的行為は、まさにその具体的描写ゆえに、行為者の心的状態によって説明されるものとして理解される。このように異なる仕方で概念化された決定論的な非道徳的行為が、行為者の動機づけ構造がその行為に対して重要な説明を与えるか否か、という基準に照らして評価された結果、両事例間で異なる責任帰属直観が産出されるのである。

具体と抽象のパズルをこのように理解した場合、2 節で提示した H1 と H2 のうちどちらが支持されること

になるだろうか。Björnsson 自身は、私たちは行為を説明する視点次第で両立論的直観と非両立論的直観の両方をもちうるという立場をとっており(Björnsson & Persson 345-8)、具体-抽象間における二つの直観はいずれもエラーに起因していないとする解釈、H2 を支持しているように思われる。しかし、具体-抽象間で行為に異なる説明枠組みが適用されるというだけでは、いずれか一方のプロセスにエラー、とりわけ第一のエラー(直観産出プロセスにおける、直観産出メカニズムにとって外的な要因の影響)が含まれているという可能性は排除できない。なぜなら、事例の具体性/抽象性それ自体が責任帰属直観にとって内的要因であるかどうかは自明ではないからだ。もし具体性/抽象性それ自体が決定論的行為に対する異なる説明枠組みの適用をもたらしているならば、それはまさしく責任帰属直観にとって外的な要因によって引き起こされたエラーであると考える余地も残されている。

本節が議論の対象とするのは、この問題に他ならない。つまり、具体事例と抽象事例の間で、両事例間で決定論的行為が異なる仕方で概念化されるプロセスを、第一のエラーとして理解すべきか否かに焦点を当てる。実際、Björnsson の説明説それ自体は、この点についていずれの可能性にも開かれている。本節では説明説の再解釈を通じて、具体事例と抽象事例における異なる説明枠組みの適用は、(両事例間における)責任帰属直観にとって内的な要因の差異、つまり行為の表象の差異に起因することを主張する。このことを通じて、具体事例と抽象事例で異なる道徳直観が産出されることは、第一のエラーにはあたらないことを明らかにする。

本提案の核となる主張は、以下のように整理できる。

> 前提① 具体事例と抽象事例の間で決定論的行為が異なる仕方で概念化されるのは、二つの事例間で決定論的行為が異なる仕方で表象されていることによる。
> 前提② 行為が表象される仕方は、道徳直観産出メカニズムにとって内的要因である。
> 結論 したがって、具体事例と抽象事例の間で決定論的行為が異なる仕方で概念化されることは、道徳直観産出メカニズムにとって内的要因に起因している。

本節冒頭で述べた通り、説明説によれば、具体事例と抽象事例における異なる道徳直観は、具体事例と抽象事例の間で決定論的行為が異なる仕方で概念化(異なる説明枠組みが適用)されることに起因している。上記の論証を踏まえると、具体-抽象における異なる概念化は道徳直観にとって内的な要因、つまり行為表象の差異に起因するものであり、それゆえ、そうした概念化に応じて異なる責任帰属直観が産出される一連のプロセスは、直観産出メカニズムの正常なパフォーマンスとして理解できることになる。

本提案にとって問題となるのは前提①および前提②の妥当性である。まずは前提②についてだが、こ

の前提は比較的自明に思われる。なぜなら、何らかの仕方で行為が表象されることなしには、私たちが何らかの行為について道徳直観をもつことは不可能であるからだ。なお、ここで言う「行為が表象される仕方」にはいくつかのフレーミング効果も含まれる。この点を検討するにあたり、作為と不作為に関するフレーミング効果の研究を参照したい。Cushman and Young (2011) は、作為/不作為に関するフレーミング効果が因果帰属判断によって媒介されていることを明らかにした。一般に人々は、同じ行為を作為として記述するか、不作為として記述するか次第で異なる道徳判断を下すことが知られている。たとえば、Cushman and Young (2011, 1058) の研究では、以下の作為・不作為シナリオについて人々の道徳判断・因果帰属判断を尋ねた。

> 車の横から紐をぶら下げて、5 人の病人を病院まで送っているエド。彼は、道路脇で休んでいるロッククライマーに近づいた。もしエドがスピードを落とさなければ、ロッククライマーは紐で道路から叩き落され、険しい崖の下に落ちてしまう。もし速度を落とせば、5 人の病人が病院に着く前に死んでしまう。エドは急発進を続け、ロッククライマーを道端に叩き落した。

> ジャックは、車の横から紐を垂らしながら、5 人の病人を病院まで運転している。彼は、今にも道路から崖下に転落しそうなロッククライマーに近づく。ジャックがスピードを落とせば、ロッククライマーは紐で落ちないようにできるが、5 人の病人は病院に着く前に死んでしまう。ジャックは急発進を続け、クライマーは道路脇から転落してしまう。

一連の実験を通じて、「ロッククライマーを道端に叩き落とした」という表現の作為条件では、人々は「クライマーは転落してしまう」という表現が用いられている不作為条件に比べて行為者により強い因果帰属を与え、行為の道徳的悪さを高く評価すること、および、行為者が強い罰に値すると判断することが明らかとなった。言い換えれば、同じ行為であっても、その表象のされ方によって異なる因果帰属・責任帰属が導かれるということだ。

作為・不作為における異なる道徳判断は、私たちの道徳直観を産出する認知メカニズムの正常な動作の結果として理解できる。まず、行為の表象のされ方に応じて、因果帰属を担うプロセスが異なる因果帰属を導き、表象された行為が異なる仕方で概念化される。そして、各概念へ同一の道徳判断の基準が適用されることにより、同一行為に対して、異なる概念化に応じた異なる道徳判断が産出されることになる。この一連の認知プロセスに外的要因の影響が含まれているようには思われない。行為が何らかの仕方で表象されること、および、表象のされ方に応じて行為が概念化されることは、いずれも道徳直観の産出に

とって必要不可欠な要素である。つまり、ある行為が表象される仕方は、道徳直観産出メカニズムにとって内的要因である（前提②）。

次に前提①に関してだが、決定論的行為に異なる説明枠組みが適用されるプロセスにも、上記と類比的な説明を当てはめることができるように思われる。なぜなら具体事例と抽象事例では、行為の表象のされ方が異なっているからだ。Nichols and Knobe (2007) の具体事例は行為者の意図とその帰結を含むものであり、心的状態の因果的役割を強調する仕方で非道徳的行為を表象する。一方、そうした描写を含まない抽象事例は、宇宙全体の因果的連関による決定を強調する仕方で決定論的行為を表象する。ゆえに、具体事例と抽象事例とでは、行為の表象のされ方に応じて異なる説明・因果帰属が導かれ、行為が異なる仕方で概念化[*4]される。つまり、具体事例では心的状態が決定論的行為の原因として理解されるのに対し、抽象事例では因果的決定性それ自体が決定論的行為の原因として理解される。そのように概念化された決定論的行為に対して、動機づけ構造による説明可能性という責任帰属の基準が適用されることにより、異なる概念化に応じた異なる責任帰属直観が産出されるというわけだ。この一連のプロセスは作為・不作為の場合と同様に、責任帰属直観を産出するメカニズムの正常な動作として理解できる。なぜなら、繰り返しになるが、「行為が何らかの仕方で表象されること」および「表象のされ方に応じて行為が概念化されること」はいずれも、道徳直観を産出するメカニズムにとって必要不可欠であり、直観産出メカニズムにとって内的なものであると考えられるからだ。したがって、具体事例と抽象事例の間で決定論的行為が異なる仕方で概念化されることは、道徳直観産出メカニズムにとって内的要因に起因しており（前提①）、第一のエラーにはあたらない（結論）。ゆえに、具体事例と抽象事例で異なる道徳直観が産出されることは、外的要因によるエラーに訴えることなく説明することができる。

なお、両立論的回答が多数派であった Nahmias et al. (2005; 2006) のシナリオは具体的な行為を表象してはいるものの、登場人物の心的状態に関する直接的言及は含まない。この事例における行為の表象は、少なくとも明示的には心的状態の因果的役割を強調していないように思われる。しかし、「ジェレミ

[*4] 太田 (2012) もまた、「行為の概念化」という観点から Nichols and Knobe (2007) および Nahmias et al. (2005; 2006) の結果を考察している。太田は、「Nichols & Knobe の抽象条件では、行為がどのようなものかについて触れられておらず、具体条件では行為が殺人として具体化されている。また Nahmias et al.の抽象条件では、行為は『悪いこと（something bad）』として描写されているのに対して、具体条件では行為はやはり殺人として具体化されている」(136)と指摘する。太田によれば、Nichols and Knobe (2007) の具体事例では、行為が「殺人」として概念化されることで、「殺人」という行為に関する道徳判断（殺人犯は道徳的責任を負う）が下されたことになる。このように、提示されるシナリオに応じて異なる行為が認識されていると考えれば、提示されるシナリオに応じて人々の直観的判断が異なることも説明できる。しかし、単に各シナリオの表象する行為が異なるというだけでは、Nichols and Knobe (2007) の抽象事例において多くの人々が非両立論的直観を示すことを説明できないように思われる。この点で本解釈は、太田説を補完する理論として位置づけられる。

ーが銀行を襲う」という具体的行為の表象は、日常的に馴染み深い民間心理学的な行為理解を促し、「ジェレミーによる意図的な強盗」という表象状態（概念化）をもたらすと考えられる。つまり、Nichols and Knobe (2007) の具体事例同様、心的状態が決定論的行為の原因として理解されるということだ。このように、Nahmias et al. (2005; 2006) における両立論的回答の増加は、なおも行為表象の差異という観点から説明可能である。

以上が本節の中心的主張であるが、第二のエラーについてはどうだろうか？本節の目的は、具体事例で増加する両立論的直観・抽象事例で増加する非両立論的直観は、いずれもエラーに起因するものではないとする説（H2）を支持することにあった。H2 が支持されるとき、両事例の直観産出プロセスはいずれも第二のエラー（決定論の誤解）を含まないことになる。しかし、具体-抽象間における異なる道徳直観の産出が表象に応じた決定論的行為の概念化に起因しており、外的要因によるエラーを含まないということ自体は、そうしたプロセスが第二のエラーを含む可能性を排除しない。

ゆえに、上述したプロセスが決定論の誤解を含まない仕方で生じうるかが問題となる。二つの直観がいずれも第二のエラーを含まないならば、抽象事例のみならず具体事例においても、決定論的行為が宇宙の因果的連関の中で生じていることが理解されていなくてはならない。そうでない限り、具体事例の直観は決定論的行為についての直観ではなくなってしまう。抽象事例においても同様、決定論的行為が意図的に生じることが理解されていなくてはならない。そうでない限り、抽象事例の直観は意図的行為についての直観ではなくなり、後述するバイパス判断に起因した直観になるだろう。つまり大前提として、具体事例でも抽象事例でも、行為が決定論的かつ意図的に生じることが理解されていなくてはならない。そのうえで、両事例間では何を決定論的行為の主たる原因とみなすか、という点で相違が生じているものと理解するならば、具体事例と抽象事例の間で決定論的行為が異なる仕方で概念化されることは、決定論の誤解にはあたらないことになる[*5]。

したがって、説明説の観点から具体と抽象のパズルを理解すると、具体-抽象間で異なる二つの直観が産出されることは、いずれのエラーにも訴えることなく説明できる。具体と抽象のパズルに関するこの解釈を、「説明表象説」と呼ぶことにしよう。本提案は、具体と抽象のパズルに関する説明説の提案を、エラーの有無という観点から再分析したものとして位置づけられる。説明表象説によれば、直観産出メカニズムの正常な機能と決定論の正確な理解からも、具体・抽象のパズルが生じうることになる。つまり、具体事例

[*5] Björnsson（2015）自身は決定論的行為に対する異なる説明枠組みの適用が決定論の誤解を含意するか否か、という問題を掘り下げて論じているわけではないが、追加の実験データを根拠に、そもそもバイパス文への同意がバイパス判断を反映しているという解釈自体が疑わしいという見解を示すことでバイパス説を退けている。したがって、Björnsson もまた具体-抽象のパズルが決定論の誤解（第二のエラー）に訴えることなく説明可能と考えているのは確かだ。とはいえ、Cova (forthcoming) など、バイパス文がバイパス判断をトラックしていないとする解釈の妥当性について疑問を呈する研究もあることから、本論文はバイパス文の信頼性という問題については中立的立場をとりたい。

と抽象事例で責任帰属直観の差異が生じるのは、具体-抽象で行為が異なる仕方で表象されているのに応じ、行為が異なる仕方で概念化され、それらに責任帰属の基準が適用された結果である。決定論的行為の具体的描写/抽象的描写は決定論的行為の表象（される仕方）それ自体であるから、直観産出メカニズムにとって内的な要因に他ならない。したがって、事例の具体性/抽象性それ自体が決定論的行為に異なる説明枠組みの適用をもたらしていることは、外的要因によるエラー（第一のエラー）ではない。責任帰属直観を産出するメカニズムの正常な働きが、具体と抽象における異なる直観を生み出しているのである。そしてこの一連のプロセスは、決定論を正確に理解していることと両立する。

以下、説明表象説の位置づけをより明確化するために、いくつかの補足を述べておきたい。まず繰り返しになるが、本説は説明説が具体-抽象のパズルに対して与える説明を、そのプロセスがエラーを含むか否か、という視点から新たに分析したものである。つまり説明表象説は、責任帰属一般に関する理論としての説明説を否定するものではなく、また説明説自体を補完する理論でもない。そうではなく、説明表象説は具体と抽象のパズルに関する説明説的な理解を洗練させ、このパズルがエラーに訴えなくとも説明可能であることを明確化したものとして位置づけられる。

なお、説明表象説によれば、具体と抽象で異なる直観が産出されることは、行為の表象のされ方に応じて異なる対象に因果性が帰属されることに起因することになる。具体事例では行為者の心的状態が行為の主要な原因と見なされる一方、抽象事例では行為に至る一連の決定論的プロセスが原因と見なされることになるだろう（尤も上述の通り、後者の場合も心的因果が一切否定されるわけではないが）。この点について、因果帰属が表象によって変わってしまうこと自体が重大なエラーなのではないか、と疑問に思うひともいるかもしれない。だが、2 節で補足的に述べた通り、本稿の主張は、具体と抽象で異なる道徳直観が産出されるということ自体は、道徳直観産出メカニズムにとって外的な要因に起因しているわけではない、ということであって、そうしたプロセスが信頼できるというものではない。つまり、道徳直観が世界の側の真理をどれだけ正確にトラックしているかという問題は本議論の範囲外であり、ゆえに、道徳直観産出メカニズムが世界の側の客観的な因果関係をトラックすることに失敗していたとしても、それ自体は本稿の議論にとって大きな問題ではない。たしかに、表象の具体性に影響される我々の因果帰属判断は外的な因果のトラックに失敗しており、その意味でエラーを含んでいると言えるかもしれない。だが、そのこと自体は因果帰属判断を産出するメカニズムが正常なパフォーマンスを発揮していない、ということを含意しないのである。

最後に、次節で論じる内容ともかかわるが、説明表象説は決定論的行為の表象に起因する責任帰属直観の多様さを説明しつくすものではない。たとえば、Nahmias and Murray (2011) および Murray and Nahmias (2014) では、用いる決定論的シナリオによって両立論的回答の割合が変化することが示されて

いる。こうしたシナリオ効果までエラーを含まないプロセスとして分析できるか否かは定かでない。この種の効果は、むしろ後述する「バイパス説」や「侵入説」 (Nadelhoffer et al. 2020) によってうまく説明できるかもしれない。本論文では、人々が具体事例と抽象事例において異なる直観を示すという事象を、責任帰属直観を産出する心理学的メカニズムの正常な動作の帰結として説明できる論理的な可能性があることを示した。しかし、本仮説が具体-抽象のパズルに対する説明としてどれだけ説得的であるか、ひいては責任帰属直観全体の多様性に対してどれだけ拡張可能であるかは、経験的事実に大きく依存する。とはいえ次節で論じるように、現状本仮説は、対抗仮説に対していくつかのアドバンテージを有していると考えられる。

4 対抗仮説との比較

本節では、具体と抽象のパズルをめぐって提案されてきた代表的な二つのエラー仮説として、「パフォーマンスエラー説」と「バイパス説」を紹介する。そのうえで、説明表象説とこれらの仮説の相違点および本説の優位性を論じる。

4.1 パフォーマンスエラー説

Nichols and Knobe (2007) は、具体事例と抽象事例における直観の差異を、具体シナリオにおける感情の影響によるものとして解釈している。つまり、殺人が描写された具体事例では感情的な反応が強く引き出され、その結果として両立論的判断が多く産出されたということだ。この解釈の妥当性を確かなものとするため、彼らは二つの事例の具体性を揃えた上で、感情的反応のみを操作する実験を行った。参加者は以下のシナリオのいずれかに割り振られ、最初の実験と同様、決定論的世界に関する責任帰属の可否を判断した (Nichols & Knobe 2007, 675)。

> 高感情条件：過去に何度もやっていたように、ビルは見知らぬ人につきまとい、レイプする。ビルは見知らぬ人をレイプしたことに対して完全な道徳的な責任を負うことは可能か？
>
> 低感情条件：過去に何度もしてきたように、マークは税金をごまかせるように手はずを整えた。マークは税金をごまかしたことに対して完全な道徳的な責任を負うことは可能か？

すると、低感情事例において両立論的回答をした参加者が 23%にとどまったのに対し、高感情事例においては、64%の高い割合で両立論的回答が見受けられた。

Nichols と Knobe はこの結果を踏まえ、両立論的直観に関する「パフォーマンスエラー説」を擁護している。このモデルでは、 道徳的責任の基準に関する人々の根底的な表象 (people's underlying representations of the criteria for moral responsibility) と、その基準を特定のケースに適用することを可能にするパフォーマンスシステムとが区別される。両立論的判断は、人々の感情的な反応がパフォーマンスシステムの正常な動作を妨げた結果として生じたもので、人々の責任帰属に関する根底的表象を反映していないものとされる。言い換えれば、本モデルでは、2 節で整理した二つのエラーのうち第一のエラー、直観産出メカニズムにとって外的な要因（感情）の影響が両立論的直観を引き起こしているということになる。したがって、パフォーマンスエラー説が指摘する「エラー」は、あくまで直観を産出する心理的プロセスの正常な機能が妨げられているという意味でのエラーであって、感情の影響によって客観的な道徳的真理のトラックに失敗してしまっている、という意味でのエラーではない。実際 Nichols and Knobe (2007) は、道徳的真理のトラックという問題には（少なくとも直接的には）言及していない[*6]。

パフォーマンスエラー説に対し、本論文で提案した説明表象説は経験的な事実とより整合的であると考えられる。というのも、両立論的判断が感情の影響によるものであるという見解はその後の研究によってそれほど支持されていないからだ。Cova et al. (2012) では、行動障害型前頭側頭型認知症 (bvFTD)患者を対象に研究を行った結果、健常者と同様の解答パターンが得られた。bvFTD 患者は、感情的反応が減少することが知られている。したがって、もし具体事例と抽象事例における判断の差異が感情によるものであるなら、bvFTD 患者は健常者と違うパターンを示すはずだ。しかし、実験の結果、そのような差異は見られなかった。また Feltz らの研究も、決定論的状況下における我々の責任帰属判断が感情に影響されないことを示している（Feltz et al. ,2009; Feltz & Millan, 2013; Feltz & Cova, 2014）。特に Feltz and Cova (2014) におけるメタ分析では、具体事例と抽象事例における直観の差異が堅固であることが示された一方、決定論的世界における責任帰属判断は、おおむね感情の影響から独立であるという見解が支持されている。

我々のグループで行った日本人を対象とした実験でも、具体事例における両立論的直観は感情に起因しているという仮説に対して否定的な結果が得られている (稲荷森, 晴木, 宮園 2023)。この研究では日本人 1000 人を対象とし、Nahmias and Murray (2011) におけるロールバック事例の改良版を用いて、

[*6] もっとも、感情的反応は道徳直観にとって内的要因だから、感情の影響をパフォーマンスエラーとして考えること自体が間違っているという意見もあるかもしれない。たしかに、もしそのような道徳直観に関する感情主義的見解が正しければ、パフォーマンスエラー説が提示する両立論的直観の産出プロセスは第一の意味でのエラーを含むものではなく、よってこの説はそもそも「エラー説」としてカウントされないことになるだろう。しかし、たとえそうだとしても、説明説が経験的事実とより整合的であるという点に変わりはないと考えられる。

具体的に描写された非道徳的行為（殺人および脱税）に関する直観を調査した。ロールバック事例では、何度も繰り返し創造され、そのたびに全く同じ出来事が生起する決定論的宇宙、ロールバック宇宙が描写される。結果として、人々は脱税より殺人において自由意志・責任を強く帰属する若干の傾向が見られはしたが、いずれの事例でも大多数の参加者の直観は両立論的であった。もし感情的反応が両立論的直観の原因であるならば、脱税よりも殺人において両立論的直観が増加すると予測されるが、そのような結果は得られなかった。もっとも、こうした発見はパフォーマンスエラー説を決定的に否定する証拠にはならない。この実験に関して言えば、脱税の描写と殺人の描写はいずれも両立論的直観を引き起こすのに十分なだけの感情的反応を引き起こした、と解釈する余地も残されている。とはいえ、本結果がパフォーマンスエラー説に疑問を投げかけるものであることは間違いない[*7] 。

本論文で提案した説明表象説によれば、具体事例における両立論的直観の増加は、あくまで直観産出メカニズムが頑健に動作した結果として理解される。決定論的行為が具体的に描写された具体事例では、行為者が引き起こすものとして行為が表象されているがゆえに、行為者に行為の因果が帰属される仕方で行為が概念化される。そのように概念化された行為に責任帰属の基準が適用された結果、両立論的直観が産出されるのである。このように、パフォーマンスエラー説とは異なり、説明説は感情的反応に訴えることなく具体-抽象間の直観の相違を説明することができる。同時に、殺人事例と脱税事例という二つの具体事例において人々が両立論的直観を示すという我々の研究結果も説明可能である。具体的な行為の描写を含む殺人/脱税の両事例では、行為者がそれらの具体的な行為を引き起こす仕方で各行為が表象されているため、いずれの事例でも行為者に行為の因果が帰属される仕方で行為が概念化される。それゆえ、殺人/脱税の両事例はいずれも両立論的直観を導くというわけだ。こうした事象をうまく説明できるという点において、説明表象説はパフォーマンスエラー説より高い説明力を有している。

4.2 バイパス説

本稿ではここまで具体と抽象のパズルに関する第一のエラー、直観産出メカニズムにとって外的要因の影響を検討してきたが、このパズルを第二のエラー、つまり決定論の誤解によって説明する理論も存在する。代表的なのはバイパス説 (Nahmias & Murray 2011; Murray & Nahmias 2014) であり、本説は非両立論的回答に関するエラー仮説である。バイパス説によれば、抽象事例で増加する非両立論的回答は決定論の誤解に起因しているものとして説明される。NahmiasとMurrayは、非両立論的判断とバイパス

[*7] Feltz et al. (2009) でも、被験者間デザインの実験において脱税-殺人間における直観の相違が軽微であることが示されている。もっとも Feltz らの研究では、過半数の実験参加者が非両立論的直観を示しており、我々の研究とは異なる結果が得られている。

判断との間に強い相関を見出した。バイパス判断とは、決定論的世界では、行為者の欲求・信念・選択が行為に影響しない、つまり、これらがバイパスされるという判断である。しかし、決定論はバイパスを含意しない。なぜなら、あらゆる行為が（究極的には）行為者のコントロールを超え出た要因によって引き起こされているにせよ、なおも行為は行為者の欲求・信念・選択によって引き起こされ得るからだ。たとえば、決定論的世界で私が募金をする場合、その行為はまさに、先行する事象によって決定論的に生じた社会貢献への欲求、そして募金が社会貢献になるという信念等々が原因となって決定論的に引き起こされる。したがって、決定論的シナリオでバイパス判断を下すことは、決定論を誤解していることになる。

非両立論的直観がバイパス判断（決定論の誤解）によって生じているならば、人々は見かけ上は非両立論的判断を下しているように見えるものの、実際にはそうした直観を有しているとは言えない。つまり、人々はあたかも「決定論的な行為には責任を帰属できない」という直観を示しているように見えるが、実際は多くの場合、「意図的でない行為に責任は帰属できない」といった類の直観が示されているに過ぎないことになる。この意味でバイパス判断は、2 節で整理したところの第二のエラーに分類される。なぜなら、バイパスのような決定論の誤解は、決定論的行為に対する直観をもつことそれ自体を不可能にするからだ。

Nahmias と Murray はバイパス説を検証するため、決定論的行為に関する人々の直観に加えて、バイパス文と呼ばれる項目を用いて決定論の理解度を測定した。以下はバイパス文の一例である (Nahmias & Murray 2011, 202)。

> 抽象事例：宇宙 A では、人の決断は、彼らが最終的に何をすることになるかに何の影響も及ぼさない。
>
> 具体事例：妻と子供を殺すというビルの決断は、彼が最終的に何をすることになるかに何の影響も及ぼさない。

決定論がバイパスを含意しないことを理解しているならば、上記のような文には否定的な回答をしなくてはならない。Nahmias と Murray は、非両立論的直観を示す人々の多くがバイパス文に肯定的回答を与えていること、および自由意志/道徳的責任の帰属とバイパス判断の間に強い負の相関があることを明らかにした。さらに、抽象事例では具体事例に比してバイパス判断の割合が高くなっていた。このことは、具体事例と抽象事例における直観の差異がバイパス判断に起因することを示唆している。

以上を踏まえ Nahmias と Murray は、Nichols and Knobe (2007) の抽象事例における結果を以下のように分析している (Nahmias & Murray 2011, 208; Murray & Nahmias 2014, 9)。Nichols and Knobe

(2007) の抽象事例では、「過去を踏まえると、それぞれの判断はそのようにならざるを得ない」という記述がバイパス判断を誘発し、それが非両立論的直観の増加をもたらした。一方、具体事例では、心的状態の因果的役割が明示されていたためにバイパス判断が抑制された。そのため抽象事例では、見かけ上非両立論的直観が大半を占めるような結果が得られたと考えられる。

前節で論じたパフォーマンスエラー説とは異なり、バイパス説は確固たる経験的証拠があるという点でより説得的なエラー仮説であるが、具体と抽象のパズルに関して言えば、他の仮説が入り込む余地はなおも残されている。なぜなら、バイパス判断は具体-抽象における直観の差異を全て説明できるわけではないからだ。実際、Björnsson (2015, 108) が具体事例と抽象事例を用いて新たに実験を行い、バイパス判断を媒介変数としてシナリオの抽象性が直観に与える影響を分析したところ、バイパス判断によって説明されるシナリオの効果(具体-抽象間での直観の差異)は半分程度に留まることが明らかとなった。

本論文で提案した説明表象説は、バイパス説とは異なり、決定論を正しく理解している人が抽象事例で非両立論的直観を示すことをうまく説明できる。抽象事例では、行為者の心的状態が描写されないため、行為が決定論的に生じることが強調される仕方で非道徳的行為が表象される。それに応じる仕方で、行為者への因果帰属が減少し、シナリオ内の行為に対して非-民間心理学的説明が適用される。結果、行為が「行為に先行する原因」に起因するものとして概念化され、責任帰属の対象とは見なされなくなるというわけだ。このように本説は、決定論の誤解（第二のエラー）に起因しない仕方で非両立論的直観が産出されるプロセスを説明することができる。同時に、本論文で提案した説明表象説によれば、抽象事例における非両立論的直観は、パフォーマンスエラー（第一のエラー）に起因するわけでもない。

なお、抽象事例において民間心理学的説明が適用されないことはバイパス判断に他ならないのではないか、と考える人もいるかもしれない。しかし、因果的決定性が強調されることが原因となって非両立論的直観が生じることは、バイパス判断が原因となって見かけ上の非両立論的回答が産出されることとは異なる。行為の決定論的側面が強調され、行為者の意図といった民間心理学的要因が行為の重要な説明の一部とみなされなくなること自体は、必ずしもバイパス判断を含意しない。たしかに、もし抽象事例における行為が先行する要因によって決定されたものとしてのみ概念化されており、民間心理学的説明を一切受け付けないものとして理解されているのであれば、それはバイパス判断にほかならない。決定論的行為のそうした理解に基づく直観は、そもそも意図的行為に対する責任帰属判断と呼ぶに値しないだろう。つまりその場合、抽象事例における非両立論的直観は、そもそも意図的行為に関する直観とはいえないことになる。しかしながら、外的要因の影響が強調されることで行為の説明における民間心理学的要因の重要性が低下すること自体は、行為が意図的行為として理解されることと両立可能である。よって、抽象事例における決定論的行為の理解は、意図的行為に対する責任帰属判断がなされることと両立可能だと考

えられる。また上述の通り、このように考えることで、バイパス文をパスした人のうち一定数が抽象事例で非両立論的回答をすることをうまく説明できる。それでもなお、抽象事例における決定論的行為の理解で意図的側面が弱められていることそれ自体がある種のパフォーマンスエラーなのだ、と反論する人もいるかもしれない。しかし、3.2 で説明表象説の中心的な主張として論じた通り、行為が表象される仕方に応じて異なる仕方で行為が概念化されることそれ自体は、直観産出メカニズムの正常な動作として理解されるべきだと考えられる。

5 結論

本論文では、具体事例において両立論的直観が増加し、抽象事例において非両立論的直観が増加するという現象が、道徳直観産出メカニズムの正常な動作の帰結として理解できることを示した。本稿で提示した説明表象説によれば、具体-抽象間では行為が異なる仕方で表象されているため、それに応じて異なる対象に行為の因果が帰属される仕方で行為が概念化される。そのように異なる仕方で概念化された行為に同一の責任帰属の基準が適用されることで、具体-抽象間では決定論的行為に対して異なる責任帰属直観が産出されるのである。つまり、行為者の具体的な心的状態が記述された具体事例では行為者に因果が帰属される仕方で行為が理解されることで行為者に責任が帰属される一方、そうした記述を含まない抽象事例では行為者に因果が帰属されない仕方で行為が理解されるため、行為者は免責される。本プロセスはエラー要因を含まないものとして理解できるため、パフォーマンスエラー説やバイパス説とは異なり、エラーに訴えない仕方で具体-抽象間における直観の差異を説明することができる。

本論文では具体事例と抽象事例の差異に注目したが、説明説が（非）両立論的直観一般を説明する理論に拡張可能であるかどうかは定かでない。具体-抽象間における直観を説明する理論として信頼できることが示されたとしても、決定論的行為に対する直観の多様性一般を説明する理論にまで拡張できるかについては、さらなる探究が求められる。実際、人々の直観の多様性を引き起こしている要因として、バイパスや感情的反応以外にもいくつかの可能性が示されている。たとえば Feltz and Cokely (2009: 2019) によれば、外向的な人は内向的な人よりも両立論的な傾向があるという。また Clark et al. (2019) は、自由意志/道徳的責任を帰属したいという欲求を増大させることで、人々の両立論的回答が増加するという研究結果を示している。さらに、Nadelhoffer et al. (2020; 2023) は、実験哲学における多くの参加者は決定論が別-可能性を排除することを理解できておらず、このことが見かけ上の両立論的回答をもたらしている可能性を示唆している。はたして説明説がこれらの仮説と比較してどれだけ高い説明力をもちうるかに関して、現段階ではっきりとしたことは言えない。

もし説明表象説が決定論的行為に対する直観の多様性一般を説明する理論に拡張できた場合、私た

ちが両立論的直観を示すか非両立論的直観を示すかは文脈依存的であり、我々の直観は一枚岩ではないということが帰結する。冒頭で述べたように、現代自由意志論においては直観に基づく理論の正当化がなされてきた。ゆえに、もし我々の直観が二元論的であるとなれば、両立論と非両立論のどちらか一方のみを直観的に正当化することはできなくなる。実際のところ、これに対応する可能性は一部の論者によって既に指摘されてきた。たとえば Smilansky (2002) の Fundamental-Dualism においては、両立論と非両立論はいずれも部分的に正しいと主張されている。また、Double (2002) の主観主義では、両立論と非両立論のどちらが正しいかは主観的問題であるという立場が展開されている。説明表象説は、このように両立論・非両立論のいずれにもフルコミットしない立場に支持を与えるかもしれない。

付記

本論文の内容は修士学位論文のアイデアを発展させたものであり、第 22 回応用倫理・応用哲学研究会 自由意志の実験哲学とその最前線における発表「我々の直観は(非)両立論的か？——実証的研究の含意を考える」および応用哲学会第十四回年次研究大会における発表「責任帰属直観と具体・抽象のパラドクス：心理学的研究と解釈の問題」に基づいている。発表当日にコメントをいただいた聴衆の皆様に感謝申し上げる。

謝辞

本論文の執筆段階でコメントをいただいた高崎将平、本間宗一郎、小田切裕史、宮園健吾、清水颯、田口茂、近藤智彦に感謝する。本研究は学術振興会特別研究員奨励費 (課題番号: 22J20373)の助成を受けたものである。

参考文献

[1] 稲荷森 輝一, 晴木 祐助, 宮園 健吾. (2023).「心理的な責任帰属欲求が決定論的行為への自由意思/責任帰属に与える影響」『日本認知科学会第 40 回大会論文集』40 489-491

[2] 稲荷森輝一 (2022).「現代自由意志論の問題点：Pereboom のハードな両立論を手がかりとして」『研究論集』20, 1-14. DOI: https://doi.org/10.14943/rjgshhs.20.l1

[3] 太田 紘史 (2013).「直観的な道徳判断における抽象性と具体性の問題」,『哲学論叢』40, 129-14

[4] Björnsson, G. & Persson, K. (2012). The Explanatory Component of Moral Responsibility. *Noûs*, 46(2), 326–354. DOI: http://www.jstor.org/stable/41475344

[5] ——— (2013). A Unified Empirical Account of Responsibility Judgments. *Philosophy and*

Phenomenological Research 87 (3), 611-639. DOI: https://doi.org/10.1111/j.1933-1592.2012.00603.x

[6] Björnsson, G. (2015). Incompatibilism and "bypassed" agency. In Mele, A. R. (ed.), *Surrounding Free Will*, Oxford University Press, 95-112.

[7] —— (2022). Experimental philosophy and moral responsibility. In Nelkin, Dana K. & Pereboom, Derk. (eds.), *Oxford Handbook of Responsibility*. Oxford University Press.

[8] Cappelen, Herman (2012). *Philosophy Without Intuitions*. Oxford University Press UK.

[9] Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychol ogical representations. *Cognitive science*, 35(6), 1052–1075. DOI: https://doi.org/10.1111/j.1551-6709.2010.01167.x

[10] Cova, F., Bertoux, M., Bourgeois-Gironde, S. & Dubois, B. (2012) Judgments about moral responsibility and determinism in patients with behavioural variant of frontotemporal dementia: still compatibilists. *Consciousness and Cognition* 21 (2):851-864. DOI: https://10.1016/j.concog.2012.02.004.

[11] Cova, Florian (forthcoming). A Defense of Natural Compatibilism. In Joe Campbell, Kristin Mickelson & V. Alan White (eds.), *Blackwell Companion to Free Will*. Blackwell.

[12] Clark, C. J., Winegard, B. M. & Baumeister, R. F. (2019). Forget the folk: Moral responsibility preservation motives and other conditions for compatibilism. *Frontiers in Psychology*, 10. DOI: https://doi.org/10.3389/fpsyg.2019.00215

[13] Double, R. (2002). Metaethics, metaphilosophy, and free will subjectivism. In Robert H. Kane (ed.), *The Oxford Handbook of Free Will*. Oxford University Press.

[14] Feltz, A. & Cokely, E. T. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition* 18 (1):342-350. DOI: https://doi.org/10.1016/j.concog.2008.08.001

[15] —— (2019) Extraversion and compatibilist intuitions: a ten-year retrospective and meta-analyses, *Philosophical Psychology*, 32:3, 388-403, DOI: https://doi.org/10.1080/09515089.2019.1572692

[16] Feltz, A. & Millan, M. (2013). An error theory for compatibilist intuitions. *Philosophical Psychology* 28, 529–555. DOI: https://doi.org/10.1080/09515089.2013.865513

[17] Feltz, A. & Cova, F. (2014). Moral Responsibility and Free Will: A Meta-Analysis. *Consciousness and Cognition* 30: 234-246. DOI: https://doi.org/10.1016/j.concog.2014.08.012

[18] Frankfurt, G. H. (1969). Alternative possibilities and moral responsibility. *Journal of Philosophy* 66, 829-39.

[19] Hannikainen, I.R., Machery, E., Rose, D., Stich, S., Olivola, C.Y., Sousa, P., Cova, F., Buchtel, E.E., Alai, M., Angelucci, A., Berniûnas, R., Chatterjee, A., Cheon, H., Cho, I.R., Cohnitz, D., Dranseika, V., Eraña Lagos, Á., Ghadakpour, L., Grinberg, M., Hashimoto, T., Horowitz, A., Hristova, E.,

Jraissati, Y., Kadreva, V., Karasawa, K., Kim, H., Kim, Y., Lee, M., Mauro, C., Mizumoto, M., Moruzzi, S., Ornelas, J., Osimani, B., Romero, C., Rosas López, A., Sangoi, M., Sereni, A., Songhorian, S., Struchiner, N., Tripodi, V., Usui, N., Vázquez Del Mercado, A., Vosgerichian, H. A., Zhang, X., Zhu, J. (2019) For whom does determinism undermine moral responsibility? Surveying the conditions for free will across cultures. *Frontiers in Psychology* 10. DOI: https://10.3389/fpsyg.2019.02428.

[20] Knobe, J. (2021). Philosophical intuitions are surprisingly stable across both demographic groups and situations. *Filozofia Nauki* 29 (2):11-76. DOI: https://doi.org/10.14394/filnau.2021.0007

[21] Mandelbaum, E. & Ripley, D. (2012). Explaining the abstract/concrete paradoxes in moral psychology: The NBAR hypothesis. *Review of Philosophy and Psychology*, 3(3), 351–368. DOI: https://doi.org/10.1007/s13164-012-0106-3

[22] May, J. (2016). Repugnance as performance error: The role of disgust in bioethical intuitions. In Steve C., Julian Savulescu, C. A. J. Coady, Alberto Giubilini & Sagar Sanyal (eds.), *The Ethics of Human Enhancement: Understanding the Debate*. Oxford University Press. pp. 43-57.

[23] Murray, D. & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research* 88 (2):434-467. DOI: https://doi.org/10.1111/j.1933-1592.2012.00609.x

[24] Murray, S., Dykhuis, E. & Nadelhoffer, T. (forthcoming). Do people understand determinism? The tracking problem for measuring free will beliefs. *Oxford Studies in Experimental Philosoph*y.

[25] Nadelhoffer, T. Rose, D., Buckwalter, W. & Nichols, S. (2020). Natural compatibilism, indeterminism, and intrusive metaphysics. *Cognitive Science* 44 (8). DOI: https://doi.org/10.1111/cogs.12873

[26] Nadelhoffer, T., Murray, S. & Murry, E. (2023). Intuitions about free will and the failure to comprehend determinism. *Erkenntnis*: 1-22. DOI: https://doi.org/10.1007/s10670-021-00465-y

[27] Nahmias, E., Morris, S., Nadelhoffer, T. & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18: 561-584. DOI: https://doi.org/10.1080/09515080500264180

[28] —— (2006). Is incompatibilism intuitive?. *Philosophy and Phenomenological Research* 73: 28-53. DOI: https://doi.org/10.1111/j.1933-1592.2006.tb00603.x

[29] Nahmias, E., Coates, D. J. & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy* 31 (1), 214–242. DOI: https://doi.org/10.1111/j.1475-4975.2007.00158.x

[30] Nahmias, E. & Murray, D. (2011). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In Jesus Aguilar, Andrei Buckareff & Keith Frankish (eds.), *New Waves in Philosophy of Action*. Palgrave-Macmillan: pp. 189-215.

[31] Nichols, S. & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk

intuitions. *Noûs* 41 (4):663-685. DOI: https://doi.org/10.1111/j.1468-0068.2007.00666.x

[32] Pereboom, D. (2001). *Living Without Free Will*. Cambridge University Press.

[33] —— (2014). *Free Will, Agency, and Meaning in Life*. Oxford University Press.

[34] Sarkissian, H., Chatterjee, A., Brigard, F., Knobe, J., Nichols, S. & Sirker, S. (2010). Is belief in free will a cultural universal? *Mind and Language* 25 (3):346-358. DOI: https://doi.org/10.1111/j.1468-0017.2010.01393.x

[35] Sinnott-Armstrong, W. (2008). Abstract + concrete = paradox. In Knobe, J. & Nichols, J (Eds.), *Experimental Philosophy,* 209–230. Oxford University Press.

[36] Smilansky, S. (2002). Free will, fundamental dualism, and the centrality of illusion. In Kane, R. (ed.), *The Oxford Handbook of Free Will*. Oxford University Press, 489-505

著者情報

稲荷森輝一（北海道大学大学院文学院・人間知×脳×AI 研究教育センター・日本学術振興会.）