能動的推論は〈拡張された心〉仮説の自然化に成功しているのか

一橋大学 神崎祥輝

はじめに

環境との相互作用において心を捉えようとする認知科学の一派は、〈4E 認知科学〉として括られてきた[1]。その中でも、Andy Clark と David Chalmers が提唱する〈拡張された心〉仮説 Theory of Extended Mind（TXM）は、〈心は頭蓋骨を飛び越え世界へと拡張する〉という、ラディカルなテーゼを打ち立てている。

TXM をめぐる議論は、擁護者と批判者の両陣営に二分されている（cf: Adams and Aizawa 2001, 2010; Clark and Chalmers 1998; Clark 2008, 2010; Sterelny 2004, 2010; Menary et al. 2010）。TXM の批判者の中には、Fred Adams と Kenneth Aizawa のように全面的に TXM を棄却しようとする論者もいる一方で、Kim Sterelny のように部分的には TXM の妥当性を認める論者も存在し、TXM が直ちに退けられる荒唐無稽な仮説ではないことを示唆している（cf: Adams and Aizawa 2001, 2010; Sterelny 2010, 480）。しかし、TXM の全面的ないし部分的妥当性をめぐる議論の主要な論点は、〈拡張〉とはいかなるプロセスで、心の〈拡張〉となる環境はどのような条件を満たしていなければならないか、という点に集中し、そもそも〈拡張〉される〈心〉とはいかなるものか、という点が十分に検討されてきたとは言えない。なぜなら、（1.3 で見るように）TXM は心がもつ機能の観点から様々な心的状態や認知プロセスを捉えてはいるが、そもそもそれらの心的状態や認知プロセスがいかなるメカニズムの下でその機能を実現しているのかという点は十分に検討されてこなかったからである（Clark 2022, 6）。しかし、〈拡張〉される対象である心のメカニズムについて明らかにしない限りは、心がどのような機能を担っているのかも不明瞭なままであり、それゆえに心の（機能の）拡張可能性を論ずることは難しいと言えるだろう。

そのような中で近年、心をベイズ主義的な推論機械として捉える能動的推論のフレームワーク Active Inference Framework（AIF）が、〈拡張〉対象である心の様々な働きを物理的かつ統一的に説明可能なものとして、多くの心理学者や認知科学者の注目を集め、感情心理学や記号創発システム論などの幅広い分野にも応用されている（cf: Barrett 2017; 谷口 et al. 2024）。また、AIF に依拠した心理学や認知科学の理論が次々と生まれていることに加えて、能動的推論が実際に行われている脳領域の特定も神経科学研究において進められている（Parr et al. 2022=2022, 93-111; Gordon et al. 2017）。AIF をめぐる以上のような状況の中、TXM の提唱者の Clark も、AIF の（主要な）擁護者の 1 人であり、AIF による TXM の再構成を試みている[2]（Clark 2015, 2022, 2023; White et al. 2024）。ブラックボックスだった心の働きを AIF が物理的に解明する、すなわち心を自然化することで、TXM において何が〈拡張〉されるのかも明確になる。〈自然化〉についての詳細は第 2 節で取り扱うが、本稿において〈心の自然化〉は、神経科学や物理学が探究するような自然法則によって、〈心〉が説明可能となることを意味する（植原 2017, 11-2）。AIF による心の自然化を通じて、拡張対象である〈心〉の研究が自然科学に接続されることに加えて、AIF による TXM の自然化が可能と判明すれば、（第 1 節で説明する）〈拡張〉プロセスや〈拡張〉条件の妥当性を自然科学の知見に照らして検証可能となることも期待できる。

以上のような状況を踏まえた上で、本稿の目的は、AIF による TXM の自然化の射程を明らかにすることである。すなわち本稿では第一に、AIF が妥

当性の高い理論的フレームワークであることを前提とした上で、TXM の妥当性を経験的に判断するための第一歩として、AIF に依拠した TXM の主張の再構成を試みる。その上で、AIF によって TXM の主張を再構成可能な点で、TXM は AIF に基づいて自然化可能であり、TXM の妥当性は今後の AIF やその周辺分野の研究動向も考慮しつつ、経験科学の知見も加味して総合的に判断されるべきであることを提案する。

第 1 節では、TXM の問題関心の背景としての認識的行為や、TXM の主張の各構成要素といった、TXM にまつわる諸概念の全般的な整理を行う。第 2 節では、AIF およびそれを基礎づける自由エネルギー原理 Free Energy Principle（FEP）について概観する。第 3 節では、AIF に基づき TXM を再構成することで、TXM の自然化の成功可否を判断する。最終的に本稿では、TXM を構成する〈拡張〉プロセス、〈拡張〉対象、〈拡張〉条件のすべてが、AIF による自然化可能性に開かれていると結論づける。

1. 〈拡張された心〉とは

TXM は、Clark と Chalmers が 1998 年に提唱して以来、TXM の批判者に対する幾度の応答や自己批判による深化を経て、様々なバージョンに分岐した（cf: Clark and Chalmers 1998; Clark 2008; Menary et al. 2010）。本節ではそれらを統一的に把握するために、TXM の基本的な主張内容と関連する重要概念を整理する。ところで TXM は、認識的行為と呼ばれる、特異なタイプの行為が関与する状況を説明するための仮説として導入された。そのためまずは、TXM が取り扱う背景状況としての、認識的行為について概観する。その上で、TXM の実質的な主張内容を、拡張プロセス、拡張対象、拡張条件の三つの要素に分解する [3]。

1.1 認識的行為

人間が行う認知処理の多くは、脳のみではなく、外的環境をも巻き込む。たとえば、ビデオゲームのテトリスをプレイする際には、脳内で想像上の図形を回転させてパズルを解くだけではなく、コントローラを操作して実際の画面上の図形を回転させパズルを解くという形で、環境が巻き込まれている。ここでは、〈図形を回転させてパズルを解く〉という認知負荷が、脳内からコントローラや画面を含む一連の外的環境へと肩代わりされている。それにより、〈脳内で二つのピース同士の一致を判断する〉という困難な認識的課題が、〈画面上で二つのピース同士の一致を判断する〉という比較的容易な認識的課題へと変化する。なぜなら、実際にピースを視覚的に知覚した上で回転操作することで、脳内処理の場合に比べて、個々のピースの形状やピース同士の一致可能性についての、手持ちの情報量が増えるからである。手持ちの情報量が増えるほど、認識的課題の不確実性は低減し、課題の難易度も下がる。このピースの回転操作のような、それ自体は（パズルを解くという）目的に直接的には資さないが、情報を増やす形で間接的に目的に資する行為を、認識的行為と呼ぶ [4]（Kirsh and Maglio 1994, 38）。以上のように、人間の心は、認識的行為を媒介に環境と結合する。

1.2 拡張のプロセス

認識的行為を通じて、認知負荷を脳内から外的環境へと肩代わりさせることができることを確認した。この〈拡張〉プロセスにおいては、脳が担う機能（の一部）が、外的環境へと委託されている。TXM を擁護する John Sutton によれば、その委託のされ方によって、〈拡張〉プロセスを、二つのバージョ

ンに大別できる（Sutton 2010, 193-4）。

一つ目は、機能等価的な〈拡張〉である。機能等価的な〈拡張〉とは、脳の機能が、それと等価な機能を果たす環境によって代替されることである。たとえば、主体が脳内に保持する信念を書き写したノートは、その主体の脳内にある信念と等価な機能を果たすことで脳内の信念を代替しており、〈拡張〉された信念である。Clark と Chalmers が挙げる例では、〈ニューヨーク近代美術館は 53 番街にある〉という脳内の信念も、〈ニューヨーク近代美術館は 53 番街にある〉というノート上のメモも、美術館の住所を主体に知らせる機能を等しくもつので、ノートはその所有者にとって〈拡張〉された心となる（Clark and Chalmers 1998, 13; Sutton 2010, 194-6）。

二つ目は、機能補完的な〈拡張〉である。機能補完的な〈拡張〉とは、脳の機能が、脳と完全に等価ではない機能を果たす環境によって補強されることである（Clark 2003=2015, Chap. 3; Sutton 2010, 204-8）。たとえば、スマートフォンのもつ情報検索機能や情報貯蔵機能は、われわれの脳が本来もつ機能と完全に等価ではないので、脳の機能を代替しているとはいえないが、思考や記憶といった脳が本来もつ機能と結びつき、脳の機能を補強している。この意味で、スマートフォンをはじめとする種々の道具もまた、〈拡張〉された心となりうる [5]。

1.3 拡張の対象

認識的行為を導く心的状態や認知プロセスは知覚や信念、学習や記憶など多岐に渡るため、〈拡張〉対象となる心的状態や認知プロセスもまた多岐に渡る。認識的行為を最初に明示的に概念化した David Kirsh と Paul Maglio は、テトリスのパズルを解くという高次の認知処理を取り扱っていたが、認識的行為はそうした高次の認知処理としての心的状態やプロセスのみに関わるわけではない（Kirsh and Maglio 1994）。〈拡張された心〉仮説においては、知覚、信念、意思決定、感情といったあらゆる心的状態や認知プロセスが〈拡張〉の対象となりうる（Clark and Chalmers 1998, 12）。1.2 で取り上げた Sutton の整理において、外的環境による心の機能の代替あるいは補完プロセスが心の〈拡張〉とされていたように、〈拡張された心〉仮説は心がもつ機能の観点から、心の〈拡張〉可能性を擁護する。それゆえ、機能を有すると一般に解されるすべての心的状態や認知プロセス、すなわち意識を除く心的状態や認知プロセスはすべて〈拡張〉の対象となる。

1.4 拡張の条件

以上では、認識的行為を通じて、環境が〈拡張〉された心となりうることを確認した。しかし、認識的行為の操作対象となるすべての環境が、〈拡張〉された心となるわけではない。Clark や Chalmers を含む TXM 支持者らによる〈拡張〉条件をめぐる論争状況を筆者が整理したところ、環境が〈拡張〉された心とみなされるためには、当該の環境は（以下で順に説明するように）認識的信頼性と現象学的透明性という二つの条件を満たす必要がある [6]（Clark and Chalmers 1998; Clark 2008; Sterelny 2010; White et al. 2024）。

一つ目が、その環境が主体に認識的に信頼されていることである。つまり、認識的行為の遂行に利用される環境が、主体による意識的な検閲や懐疑にさらされないということである（Clark 2008, 79; Sterelny 2010, 473-5; White et al. 2024, 10-1; cf: Clark and Chalmers 1998, 17）。たとえば、ノートの頁上のメモ内容がわれわれ自ら記したものであれば、通常はその内容の真偽を追加で検討することなく、記憶の想起や意思決定において利用できる。こ

れはその情報が、主体によって無批判的に利用されることを意味し、認識的信頼性を獲得しているといえる。

二つ目は、その環境が主体にとって、もはや心と独立の環境ではなく、自身の心的状態や認知プロセスの一部として経験されていることである（Clark 2008, 79; White et al. 2024, 12-3; cf: Clark and Chalmers 1998, 17; Sterelny 2010, 475-7）。これを、現象学的透明性と呼ぶ。認識的行為における現象学的透明性は、非認識的行為、すなわち実践的行為における、身体拡張の条件としての現象学的透明性と同様に理解することができる。

身体拡張としての現象学的透明性はたとえば、目の見えない人が白杖をつく場合などにおいて認められる（Sterelny 2010, 475-7; White et al. 2024, 12-3; cf: Gallagher and Zahavi 2008; Head 1920）。目の見えない人が杖を使い始めた当初は、杖の利用習熟度は低く、杖が床に接地する度に、手に衝撃を伝える外的な対象として経験される。しかしだんだんと杖の利用に習熟するにつれて、杖は身体から独立な外的対象ではなく、腕の延長として身体と連続的に経験されるようになる。つまり、杖が〈拡張〉された腕として行為主体に経験される際、杖はもはや道具としては意識されてはいない。これが、身体拡張における現象学的透明性である。

以上は、実践的行為を遂行する際に、身体が環境へと〈拡張〉するケースである。それと同様の構図で、認識的行為を遂行する際に、心は環境へと〈拡張〉する。すなわち、認知主体がノートやコントローラのような道具の利用に習熟するにつれて、道具は心的状態やプロセスから独立な外的対象ではなく、心的状態や認知プロセスを構成する要素の一つとして経験される。つまり、もはや道具としては意識されず、現象学的に透明なものとして経験されるのである。

2. 能動的推論のフレームワークによる心の自然化

前節では、TXM を拡張プロセス、拡張対象、拡張条件の三つの要素に分解した。本節では、Clark が近年支持している、心についての統一的な物理学的フレームワークとしての AIF を取り上げる。AIF は、心のメカニズムの物理的な記述を試みている点で、心の自然化プログラムの一種であると言える。〈自然化〉の内実については様々な立場が存在するが、本稿でいう〈自然化〉は、植原亮による定式化に則り、当該の哲学的学説や理論およびそこでの探究対象が、自然の秩序のうちに捉えられる自然現象の一種であるとみなされることを意味する（植原 2017, 11-2）。植原によれば、とりわけ〈心の自然化〉は、心が備える性質に物理主義的な説明を与えることを通常は意味する（植原 2017, 12）。ここで、〈物理主義的な説明を与える〉ということは、自然法則の支配する物理的基盤の上に説明対象を定位することを指す（植原 2017, 12）。まとめると、〈心の自然化〉とは、神経科学や物理学が探究するような自然法則によって、〈心〉が説明可能となることを意味する。以上を踏まえ、本節では心の自然化プログラムとしての AIF の骨子を確認する。その上で第 3 節では、AIF による TXM の自然化の成功可否を判断する。

Clark が依拠する AIF は、心をベイズ主義的な予測処理の機械として捉える。つまり AIF によれば、心とは、世界の状態について予測すなわち事前に〈推論〉した結果と、予測に沿った行動を通じて得られたデータを加味し事後的に〈推論〉し直した結果との、予測誤差を最小化する確率処理である（Clark 2015, 2022, 2023; White et al. 2024）。

ここで Clark がいう〈推論〉が、〈推論〉についてのわれわれの素朴な理解とはやや異なる点に注意が必要である。つまり、AIF における〈推論〉は、

三段論法のような演繹的な推論ではなく、世界についての確率モデル（ないし仮説）を生成することを指す。また、モデル生成としての推論は、意識的に行われる三段論法とは対照的に、無意識的な認知処理である（Clark 2015, 2）。以下では AIF の文脈に則り、〈確率モデルの無意識的な生成〉という意味で、〈推論〉という語を用いる。

このモデル生成を通じた予測誤差の最小化プロセスは、〈自由エネルギー原理 Free Energy Principle（FEP)〉というさらに一般的な、自由エネルギー（FE)の最小化プロセスに基づいて物理的に記述可能であることが、Thomas Parr らによって示されている [7]（Parr et al. 2022=2022）。FE 最小化が予測誤差最小化よりも一般的なプロセスだといえるのは、すぐ後にみるように、FE 最小化の一手段に予測誤差最小化が含まれるからである。

ここで FE とは、熱力学における物理量であり、以下のように定義される [8]（乾・阪口 2020, 付録, 1）。

（FE）＝（内部エネルギー）−（温度）×（エントロピー）

そして、FEP を提唱する Friston らによれば、人間の心のメカニズムも、FE の方程式で表現することができる。ただし、元の FE が熱力学的な物理量であったのに対し、Parr らの FEP では、情報学的な物理量としての〈変分 FE〉を取り扱う。変分 FE は、以下のように定義される（乾・阪口 2020, 付録, 8）。

（変分 FE）＝（エネルギー）−（エントロピー）

以上の FE の定義式と変分 FE の定義式を見比べれば明らかなように、情報量としての変分 FE には温度という変数が関与しないことを除けば、両者は同様の構造を成している。つまりここから、FEP によれば、心が（変分）FE というある種の物理量の最小化を目的としているならば、心とは物理的なアルゴリズムにすぎないことが示唆される。

そして、以上の変分 FE の定義式に適切な変形を施すことで、Clark が心の根本的な機能として捉える予測誤差最小化も、その上位目的である FE 最小化の一手段であることが判明する。Parr らによれば、変分 FE は定義式からの変形によって、以下のようにも表現できる（Parr et al. 2022=2022, 29; 乾・阪口 2020, 付録, 9）。

（変分 FE）＝（ダイバージェンス）＋（シャノンサプライズ）

上記の式の右辺の各項が意味するところを簡潔に説明する。まず、第 1 項の〈ダイバージェンス〉とは、認識確率と真の事後確率の差分である（Parr et al. 2022=2022, 31; 乾・阪口 2020, 21）。つまりダイバージェンスは、実際の生成モデルの精度に応じて生じる予測誤差を意味する。右辺の第 2 項の〈シャノンサプライズ（SS)〉は、感覚信号 s が生じる確率 p(s)のマイナスの自然対数である。つまり、感覚信号 s がめったに生じない感覚信号であるほど、SS の値は大きくなる。

右辺において変数は、以上で説明した二つしか存在しないので、左辺の変分 FE を最小化する手段も、予測誤差の最小化と SS の最小化の二つしか存在しない。予測誤差の最小化とは、予測の精度を上げることである。SS は感覚信号の生起確率の負の対数を取ったものなので、FE 最小化のもうひとつ

の手段は、感覚信号の生起確率を上げること、すなわち予測された感覚信号を、自ら能動的に証拠として観察することである。これは〈予測の自己成就〉と呼ばれ、運動行為を通じて達成される（Parr et al. 2022=2022, 32）。以上をまとめると、FE の最小化手段には、正確な知覚を通じて予測精度を上げる方略と、予測に適合的な行為を通じて予測を自己成就させる方略の二つが存在する。

本節の最後に、ここまでの内容をまとめた上で、知覚から行為に至るまでのあらゆる心的状態やプロセスを、ある種の〈推論〉として解釈できることを示す。ここでいう〈推論〉とは、世界の状態についての予測と、その予測に基づいた行動を通じて得られた感覚データとを照合することで、予測をアップデートしていくプロセスであった。AIF は、FE の最小化メカニズムであり、その最小化手段には予測誤差の最小化と予測の自己成就の二つが存在する。

第一に、予測誤差の最小化からは、知覚が推論として示される。つまり知覚とは、従来の認知科学で理解されていたように、受動的に得られた外的環境についての感覚データを脳内の神経発火パターンとしてボトムアップに集積させたものではない。知覚は実際には、以下のような推論プロセスである。まず知覚主体は、これから入手が予測される感覚データと、その予測上の感覚データを引き起こすものとして推論された原因のペアからなる因果的パターンを脳内のモデルとして事前に生成する。この脳内で生成された因果モデルを、〈生成モデル〉と呼ぶ（Parr et al. 2022=2022, 8）。続いて知覚主体は、対象を知覚することで得られた実際の感覚データを、事前に予測された感覚データと照合させる。最終的に、予測データと感覚データに誤差があった場合は、生成モデルが更新される（Parr et al. 2022=2022, 16; 乾・阪口 2020, 16-7）。このように知覚は、事前に生成された脳内モデルに依拠した、トップダウンの処理過程である。

そして知覚のみならず、信念、学習、記憶といった、知覚に関連するさまざまな心的状態やプロセスも、AIF を介した予測誤差最小化により統一的に説明できる[9]（Parr et al. 2022=2022, Chap. 10）。すなわち AIF では、知覚的側面をもつ様々な心的状態や認知プロセスの機能を、事前に予測された感覚データに、実際に得られた感覚データの誤差を反映して、生成モデルの精度を向上させる機能として、統一的に説明することができる。

第二に、SS を最小化するための予測の自己成就から、行為もまた推論として示される。生成モデルの正しさを確証するために、モデルと適合的な感覚データを能動的に取得することが、行為なのである。つまり AIF では、行為および行為に関連する、欲求、感情、意思決定などのさまざまな心的状態やプロセスの機能も、SS の最小化として統一的に説明できる[10]。

そして最後に、予測誤差の最小化も SS の最小化も、ともに FE 最小化の手段である。以上のように AIF では、あらゆる心的状態やプロセスの機能が、FEP で統一的に説明される。

3. 能動的推論のフレームワークによる〈拡張された心〉仮説の自然化

AIF に基づけば、FE を最小化する物理的な機械として心を捉えることができる。AIF は TXM にいかなる含意をもたらすのだろうか。第 1 節では TXM を〈拡張〉プロセス、〈拡張〉対象、〈拡張〉条件という三つの要素に分解した。以下では、AIF が各要素の自然化に成功しているのか否かを検討する。本稿の結論としては、これら三つの要素をすべて AIF は自然化可能である。

第一に、〈拡張〉対象としての〈心〉の自然化である。AIF に基づけば、多

様な心的状態や認知プロセスを、予測誤差を最小化するベイズ統計的処理、あるいは予測を自己成就させる運動行為系列のいずれかとして物理的に説明できる。TXM は認識的行為、すなわち〈手元の情報量を増やすことで間接的に目的達成に貢献する行為〉を導く一連の心的状態を説明対象としているのであった。すなわち AIF に基づけば、認識的行為は〈予測誤差あるいは SS を最小化することで間接的に FE 最小化に貢献する行為〉と理解し直すことができる。つまり、AIF の上に立脚した TXM では、拡張対象である〈心〉を、FE を最小化するベイズ統計的処理として、自然科学の範疇に収める自然化に成功しているといえる。

第二に、〈拡張〉プロセスの自然化である。AIF の観点からは、TXM における心の〈拡張〉を、心の根本機能である FE 最小化プロセスの、外的環境への委託として理解できる。委託の仕方には、機能等価的な仕方と機能補完的な仕方の二種類が存在した。以下では、FE 最小化機能が、どちらにおいても委託可能であることを示す。

まずは FE 最小化機能が、機能等価的な〈拡張〉として環境に委託されると想定する。これは言い換えれば、心がもつ FE 最小化機能（の一部）が、それと等価な機能をもつ環境に代替されるということである。AIF において、FE の最小化は、(1)予測精度の向上を通じた予測誤差の最小化あるいは(2)予測の自己成就的な行為を通じた SS の最小化のいずれかによってなされた。

第一に、脳が果たす予測誤差最小化の機能（の一部）が環境に代替されるケースを考える [11]。Clark と Chalmers が挙げていた〈拡張〉された信念の事例は、まさにこのケースに当てはまる。なぜなら、美術館の住所についての脳内信念の代わりにノート上のメモを認知主体が参照することで、予測精度が向上し、最終的に〈美術館に辿り着く〉という予測が実現する可能性が高くなる、すなわち予測誤差が最小化されるからである。つまりこのケースでは、脳内信念が果たしていた予測誤差最小化の機能が、ノートのメモによって代替されている。

また、予測誤差最小化機能の代替と同様に、SS 最小化機能の代替についても考えることができる。たとえば、〈朝 8 時に起きて家を出れば大学の講義に間に合う〉という予測をしている認知主体が朝 8 時に目覚ましをかけるケースを考える。このとき、予測の精度自体は変わらないが、目覚まし時計を使うことで、予測に適合的な行為を取りやすくなっている。すなわち、SS が最小化されている。

続いて、FE 最小化機能の、機能補完的な〈拡張〉について考える。機能補完的な〈拡張〉とは、脳の機能が、脳と完全に等価ではない機能を果たす環境によって補強されることであった。つまり機能補完的な〈拡張〉においては、脳が果たす FE 最小化機能が、脳がもたない機能をもつ環境のサポートによって補強される。たとえば Clark が挙げていた、携帯電話を用いて自身の心拍数をモニタリングするケースが、環境による FE 最小化機能の機能補完的な〈拡張〉に当たるだろう（cf: Clark 2003=2015, 40-1）。なぜなら、われわれ人間の脳は心拍数を(視覚的に)モニタリングする機能をもたないが、そうした機能をもつ携帯電話のような環境のサポートを受けることで、自らの健康状態についての予測精度を向上させることができるほか、〈自分は健康である〉という予測に適合的な行為（たとえば、睡眠時間を増やすなど）を取りやすくなるからである [12]。以上のように、どのバージョンの〈拡張〉プロセスについても、FE の最小化機能が環境に委託されているので、AIF による〈拡張〉プロセスの自然化が成功していると言える。

第三に、〈拡張〉条件の自然化である。〈拡張〉の条件は、認識的信頼性と

現象学的透明性の二つであった。AIF がこれら二つの条件の自然化に成功しているか否かは、両条件を AIF に基づいて物理的に記述可能か否かにかかっている。そして、〈拡張〉条件の自然化こそが、TXM の自然化の要である。というのも、Clark をはじめに TXM 支持者は、〈拡張〉プロセスや〈拡張〉対象については AIF による再構成を試みているものの、〈拡張〉条件については従来の TXM における条件をそのまま適用し、AIF による再構成はなされていないからである（White et al. 2024, 15; Constant et al. 2022, 388; cf: Clark 2022, 5)。しかし、White や Constant らも認めるように、〈拡張〉条件も AIF によって説明される必要がある（White et al. 2024, 15; Constant et al. 2022, 388)。本稿では、これら二つの条件がともに、AIF による説明可能性に開かれていることを示す。

まず、認識的信頼性は、FEP を用いて物理的に記述することが可能である。認識的信頼性とは、認識的行為の遂行に利用される環境が、主体による意識的な検閲や懐疑にさらされないことであった。AIF に基づけば、利用対象となる環境が意識の検閲や懐疑にさらされないのは、その環境を利用することで得られると認知主体が予測した感覚データと、その結果実際に獲得された感覚データが一致し、予測誤差最小化が達成されているからであると考えられる。つまり、TXM の〈拡張〉条件の一つである認識的信頼性は、AIF に基づいて説明することができると考えられる。

現象学的透明性もまた、AIF による自然化可能性が見込まれる。現象学的透明性とは、当該環境が主体にとって、心と独立の外的環境ではなく、自身の心的状態や認知プロセスの一部として経験されていることであった。第 1 節で言及した通り、現象学的透明性は心の〈拡張〉だけでなく、身体の〈拡張〉においても認められる。身体〈拡張〉においては、当該の外的環境は行為主体にとって、身体と独立の外的環境ではなく、自身の身体の一部として経験される。この状態は、FEP の下では、SS の最小化として理解可能ではないだろうか。SS の最小化は、予測の自己成就、つまり認知主体が事前に立てた予測に適合的な運動行為の遂行によって達成されるのであった。すなわち、身体〈拡張〉においては、環境を利用する場合であっても、環境を利用しない場合（純粋に自らの身体のみを行使する場合）と同様の予測モデルが適用できるように、運動行為が微調整されている可能性がある [13]。

身体〈拡張〉において SS の最小化がなされているとすれば、同様に、心の〈拡張〉においても SS の最小化がなされていると考えられる。身体の〈拡張〉が実践的行為を伴うのに対し、心の〈拡張〉は認識的行為を伴う。つまり、認識的行為も行為である以上は、ボタン操作やノートの頁をめくるといった、しかるべき身体行為を伴う。したがって、心の〈拡張〉においては、環境を利用する場合であっても、環境を利用しない場合（純粋に

脳状態や脳内プロセスのみにアクセスする場合）と同等の予測モデルが適用できるように、身体行為が微調整されている可能性がある [14]。

以上のように、二つの〈拡張〉の条件も、AIF による自然化可能性に開かれていることが示唆された。本節の議論を踏まえると、TXM を構成する三つの要素である、〈拡張〉の対象、〈拡張〉のプロセス、〈拡張〉の条件はすべて、〈FE の最小化〉という観点から心の機能を捉える AIF によって自然化することができると言えるだろう。

おわりに

本稿では、TXM の自然化の手立てとして AIF を位置づけた上で、AIF による TXM の自然化の成功可否を明らかにすることを目的とした。上記の目

的を達成するために、以下の手順を取った。第 1 節では、そもそも TXM において何が自然化されなければならないのかを明らかにするために、問題関心の背景としての認識的行為や、TXM の各構成要素について概観的なアプローチを採用した。第 2 節では、AIF およびそれを基礎づける FEP が、いかにして〈心の自然化〉プログラムといえるのかを検討した。第 3 節では、AIF に基づき TXM を再構成し、AIF による TXM の自然化の成功可否を判断した。最終的に、TXM を構成する各要素は AIF の射程に収まり、自然化が可能であると結論づけた。

しかし、本稿の到達点には限界もある。第一の点は、数理モデルや実証データを用いた具体性の不足である。本稿では TXM の各構成要素の、AIF に基づく自然化可能性を概観する形で検討したため、あくまでも自然化の〈スケッチ〉に留まり、各構成要素の自然化の具体的な仕方までは説明できていない。TXM の各構成要素を AIF に基づいて数理モデルとして記述できた際に、TXM の自然化がさらに進んだと言える。第二の限界は、AIF そのものが抱える課題に由来する。すなわち本稿では、TXM が依拠する素朴な（常識心理学的な）機能を、AIF に基づく FE 最小化機能として再構成することを試みたが、FE 最小化機能が実現されている物理的基盤（脳神経系など）については言及できていない。これらの点が今後の研究課題として残っており、AIF の今後の発展とともに取り組まれていくべきである。

参考文献

Adams, F. and Aizawa, K. (2001) “The Bounds of Cognition”, *Philosophical Psychology*, 14-1, 43-64.

———. (2010) “Defending the Bounds of Cognition”, In Menary, R et al. *The Extended Mind*, MIT Press.

Barrett, L. F. (2017) *How Emotions are Made: The Secret Life of the Brain*, Houghton Mifflin Harcourt.

Chalmers, D. (2022) *Reality +: Virtual Worlds and the Problems of Philosophy*, W W Norton & Co Inc. (高橋則明訳 (2023)『リアリティ＋（下）：バーチャル世界をめぐる哲学の挑戦』, NHK 出版.)

Clark, A. and Chalmers, D. (1998) “The Extended Mind”, *Analysis*, 58-1, 7-19.

Clark, A. (2003) *Natural-Born Cyborgs: Minds, Technologies and the Future of Human Intelligence*, Oxford University Press.（呉羽真・久木田水生・西尾佳苗訳（2015)『生まれながらのサイボーグ：心・道具・知能の未来』, 春秋社.）

———. (2008) *Supersizing the Mind: Embodiment. Action. And Cognitive Extension*, Oxford University Press.

———. (2010) “Memento’s Revenge”, In Menary, R et al. *The Extended Mind*, MIT Press.

———.(2015) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford University Press.

———. (2022) “Extending the Predictive Mind”. *Australasian Journal of Philosophy*, 102(1), 119–130.

———. (2023) *The Experience Machine: How Our Minds Predict and Shape Reality*, Penguin Books.

Constant, A, Clark, A, Kirchhoff, M, Friston, K. (2022) “Extended Active Inference: Constructing Predictive Cognition beyond Skulls”, *Mind & Language*, 37(3), 373-394.

Friston, K. (2009) “The free-energy principle: a rough guide to the brain?”, *Trends in Cognitive Science*, 13-7, 293-301.
Gallagher, S. and Zahavi, D. (2008) *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*, Routledge.（石原孝二・宮原克典・池田喬・朴嵩哲訳（2011）『現象学的な心：心の哲学と認知科学入門』, 勁草書房.）
Gordon, N, Koenig-Robert, R, Tsuchiya, N, van Boxtel, JJ, Hohwy, J. (2017) Neural markers of predictive coding under perceptual uncertainty revealed with Hierarchical Frequency Tagging. *eLife*, 6: e22749.
Head, H. (1920) *Studies in Neurology*, Volume 2. Oxford University Press.
Iriki, A., Tanaka, M. and Iwamura, Y. (1996) Coding of Modified Body Schema during Tool Use by Macaque Postcentral Neurons, *NeuroReport*, 7, 2325-2330.
Kirsh, D. and P. Maglio. (1994) On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science* 18: 513-59.
Menary, R. (ed.) (2010) *The Extended Mind*, MIT Press.
Parr, T., Pezzulo, G., Friston, K. (2022) *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*, MIT Press.（乾敏朗訳（2022）『能動的推論：心、脳、行動の自由エネルギー原理』, ミネルヴァ書房.）
Sterelny, K. (2004) “Externalism, Epistemic Artefacts and the Extended Mind”, In Richard Schantz et al., *The Externalist Challenge*, De Gruyter. pp. 239-254.
———. (2010) “Minds: extended or scaffolded?”, *Phenom Cogn Sci* 9, 465-481.
Sutton, J. (2010) “Exograms and Interdisciplinarity: Historythe Extended Mind, and the Civilizing Process”, In Menary, R et al. *The Extended Mind*, MIT Press.
植原亮（2017）『自然主義入門：知識・道徳・人間本性をめぐる現代哲学ツアー』, 勁草書房.
谷口忠大編（2024）『記号創発システム論：来るべき AI 共生社会の「意味」理解に向けて』, 新曜社.
乾敏朗, 阪口豊（2020）『脳の大統一理論：自由エネルギー原理とはなにか』, 岩波書店.

[1] 〈環境〉という概念は非常に多義的であり、脳と対置した場合の脳外環境としての〈外的環境〉は、身体、自然環境、道具、社会制度、文化などを幅広く含みうる。TXM における拡張先としての〈環境〉もまた、論者によってこれらを幅広く含みうる（Clark and Chalmers 1998; Clark 2008; Sutton 2010）。本稿では、Clark と Chalmers の 1998 年の議論に則り、〈環境〉を〈道具〉として限定的に定義する。

[2] ただし、本稿以降で触れる通り、Clark が進める AIF による TXM の再構成プロジェクトには不十分な点もいくつか存在する。本稿の目的はそうした点を指摘し、代替案ないし改善案を示すことにある。

[3] 本論で TXM の主張内容を三つの要素に分類した理由は、(1) TXM をめぐる既存の議論への依拠という外在的理由と、(2) 本論の第 3 章で AIF による TXM の自然化の射程を図るための準備作業という内在的理由の二つに分けられる。以下でそれぞれについて簡潔に説明する。(1) Clark and Chalmers（1998）、Clark（2008）、Sterelny（2010）、Sutton（2010）な

どの TXM の先行研究においてすでに、〈拡張のプロセス〉、〈拡張の対象〉、〈拡張の条件〉に（必ずしも明示的ではないが）分解して TXM の主張内容が検討されているため、その形式を踏襲した。(2) 拡張のプロセス、拡張の対象、拡張の条件という三つの構成要素に TXM の主張を分解することで、それらすべてを AIF に基づいて再構成可能か否かが判別可能となる。たとえば、Clark（2022）では本稿と同様に AIF による TXM の再構成が試みられているが、拡張の条件については〈信頼性〉と〈容易なアクセス可能性〉という従来の TXM における条件が再利用されており、環境がこれら二条件を満たすことが AIF ではどのように説明されるのかが不明瞭なままになっている（cf: Clark 2022）。拡張の対象についても Clark（2022）では同様の課題が残っている。本稿ではこうした課題を検討するために、以上の三要素に TXM を分解した。

[4] 目的に直接資する行為のことを、実践的行為と呼ぶ（Clark 2022, 3）。実践的行為の例としては、(林檎を食べたいという目的に資する）林檎を掴む、という行為などが挙げられる。

[5] 本稿では機能補完的な〈拡張〉の例としてスマートフォンを挙げたが、Clark や Chalmers が〈拡張〉のケースとして挙げている、スマートゴーグルによる知覚の〈拡張〉や、BMI（Brain Machine Interface）による思考の〈拡張〉なども、機能補完的な〈拡張〉に含まれるだろう（Clark 2003=2015; Chalmers 2022=2023）。

[6] TXM が初めて提唱された Clark and Chalmers (1998)、およびその注釈書的な扱いである Clark (2008)においては、信頼性、容易なアクセス可能性、自動的承認、過去の意識的承認の四つが〈拡張の条件〉として提示されている（Clark and Chalmers 1998, 17; Clark 2008, 79）。これら各条件の是非を問うことは本稿の射程を超えるが、〈拡張の条件〉をめぐるその後の論争や Clark らの最新見解を踏まえると、Clark や Sterelny をはじめとする TXM 支持者の多くが認める〈拡張の条件〉はそれら四つのうち、信頼性と透明性の二つになると思われる（cf: Sterelny 2010; White et al. 2024）。

[7] FEP を初めて提唱したのは Karl Friston であるが、本稿では FEP および AIF の主要文献と広くみなされている Parr et al.（2022）を主に参照している。なお、Friston もこの文献の共著者の一人である。Friston がはじめて AIF について総説した著作としては、Friston（2009）を参照。

[8] FEP の物理学的な定式化と数学的変形については、Parr et al.（2022）の第 2 章および乾・阪口（2020）の付録を参照。

[9] ここで挙げた様々な心的状態やプロセスが AIF によってどのように説明されるのかは、Parr et al.（2022）の第 10 章および乾・阪口（2020）を参照。

[10] 行為に関連する心的状態やプロセスの AIF による説明可能性についても、Parr et al.（2022）の第 10 章および乾・阪口（2020）を参照。

[11] 予測誤差最小化機能の機能等価的な〈拡張〉のケースを AIF に基づいて再構成する試みは、Clark や Constant らも論じている（Clark 2022, 6; Constant et al. 2022, 387-90）。この点に関する本稿の見解も、Clark や Constant らと概ね共通する。ただし、Clark や Constant らが予測誤差最小化機能の機能等価的な〈拡張〉プロセスの AIF に基づく再構成に留まるのに対し、本稿では SS 最小化機能の機能等価的な〈拡張〉プロセスや、機能補完的な〈拡張〉プロセスの AIF に基づく再構成も試みている。

[12] 心拍数の視覚的なモニタリングは脳のもつ機能ではないが、心拍数のモニタリング自体は脳（延髄）の機能であるため、本稿の例は機能補完的な

〈拡張〉に該当しない、という反論が予想される。その場合は、完全記憶能力やエコロケーションなど、スマートフォンのような環境が果たすことができるが、人間の脳が（通常は）果たせないような、より極端なケースに置き換えて論ずることができるだろう。

[13] 本稿では環境の利用時と非利用時における SS の値が等価になりうることを FEP の数理モデルを用いて物理的に示すことはできないが、その傍証となる神経科学の実験を挙げることはできる。入來篤史らによれば、飼育実験環境下のニホンザルを、熊手を用いて手の届かない距離にある餌を取るように訓練した結果、身体表象機能を司る大脳皮質頭頂葉の神経細胞の活動パターンが、熊手を腕の一部として同化させるように変化したという（Iriki et al., 1996）。

[14] 本稿では、現象学的透明性の AIF に基づく再構成について、環境の非利用時の予測モデルをデフォルトに、環境利用時にも同様の予測モデルを認知主体が使っていると想定した。しかし、それとは逆転的な構図で SS の最小化を考えることもできる。すなわち、環境の非利用時であっても、環境の利用時と同様の予測が利用できるように、環境を利用する場合と同様の身体行為があえて動員されている可能性である。たとえば、そろばん習熟者は実際にそろばんを使わずとも、そろばんを弾くように指を動かすことで、複雑な暗算を実行できる。これは、そろばんを利用せずとも、そろばん利用時と同様の予測を立てることで、SS が最小化されているケースとして想定できる。