はじめに

本書を手に取った人は「自由エネルギー原理」という、ミステリアスで難解そうだが、強力でなんでも説明してくれそうな原理の名前に引かれたのだろうか。それとも自由エネルギー原理という名前は聞いたことがあって「脳の大統一理論」になりうる理論だ、という評判も知っていたから本書を開いたのだろうか。

実際、いかにも難解そうで、いかにもすべてを説明するかのような響きをもつこの原理は、さまざまな宣伝文句と結び付けられてきた。「脳の大統一理論」、「進化論以来の生物学分野の統一理論候補」、「すべての『モノ』の理論」と言ったような具合だ。本原理の提唱者であり、そしてその原理に基づく諸理論を展開している中心人物であるカール・フリストンはかつてこう言っていた。「脳中の変化はすべて予測誤差の最小化を目的としているのだ。ニューロンの発火もそれらの間のつながりもそうだし、目の動きから日常生活での私たちの選択にいたるまでそうなのだ」(Huang, 2008)。現代を代表する心の哲学者のアンディ・クラークは、自由エネルギー原理と基本的なアイデアを共有する予測処

理の枠組みとその基本的な主張についてこう言っている。「知覚、行為、身体化、そして人間の意識経験の本性、を理解するための新たな統合的な枠組みは待ち望まれているが、予測処理のこれら四つの主張（著者注：予測の観点から知覚と行為を統一的に説明できるという一連の主張）はそれにつながる最初の一歩だ」（Clark, 2015）。

長らく心理学や神経科学など心に関わる科学は、物理学におけるアインシュタインの相対性理論やマクスウェルによる電磁気学のような「数式で書かれた」統一理論を持たなかった。心の科学にも数学的な理論が全くないわけではなかったが、どれも限定的な適用領域でのみよく機能するものであった。たとえば心理物理学では刺激強度と感覚強度の関係を記述するウェーバー＝フェヒナーの法則やスティーヴンスのべき法則が提案され、記憶研究ではワーキングメモリの容量や構成要素、あるいはアイコニックメモリ（視覚情報をごく短時間保持する記憶）の持続時間などが実験的に体系化されてきた。これらはどれも有用な理論枠組みであったが、それが適用されるのは限られた領域であって、心のはたらき一般がしたがうべき法則や原理といったもの、特に数式で書かれたもので広く受け入れられたものはなかったのだ。

物理学は理論の統一を経て実際に一段と進化を遂げたように思われる。単に、同じ理論で語れることによって見通しが良くなるということ以外にも、新しい予測が生まれやすくなる。実際、マクスウェルの電磁気学は電気と磁気を統一することで、電磁波の存在とその速度（光速と一致する）を予測し、そこから「光は電磁現象である」という結論を導いた。そしてこの予測は、実際にヘルツにより検証されたのだ。彼は電磁波を生成・検出し、その反射や屈折が光と同様であることを示したのだ。

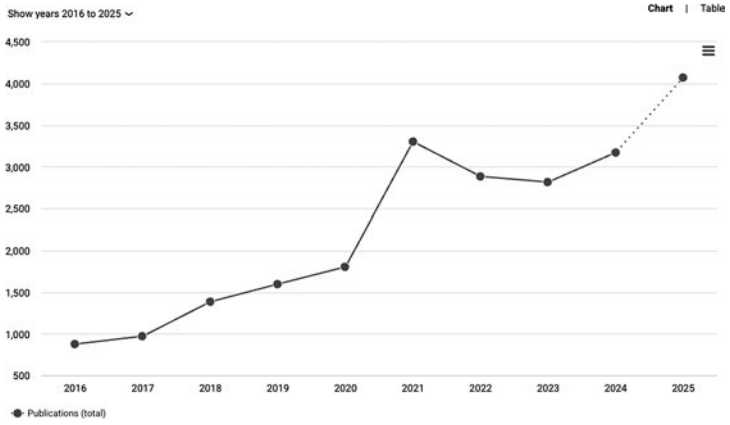


図1　Digital Science社が提供する研究情報プラットフォームであるDimensionsで、検索式（“free energy principle” OR “active inference”、Search in: Full data、指標：Publications）で年別件数を集計したもの（2025/12/19に取得）。

もし自由エネルギー原理が、心や脳の統一理論となるのならば、心の科学においても同様の飛躍的な進歩が期待できると考えられるだろう！　実際にこの原理は、知覚・行動・学習を、ばらばらの機構の寄せ集めではなく、同じ形式のもとで記述しようとする（Friston, 2010）。さらに、行為を含む枠組みとしての能動推論を通じて、「環境の中で生き延びるとはどういうことか」という生命一般の問題と、脳・心の問題とを同じ数理の言葉で接続する野心をもつ。心理学・神経科学・AI・ロボティクスまで橋渡しできる可能性がある点は、この理論の最大の魅力だ。

実際、そのような期待を背景に、多くの科学者、哲学者が自由エネルギー原理や関連する理論を用いる分野は増え続けており、それに関連した論文の数も増え続けている（図1、図2）。

しかし自由エネルギー原理の影響力が増すにつれて、それに対する批判や懐疑的な声も増えてき

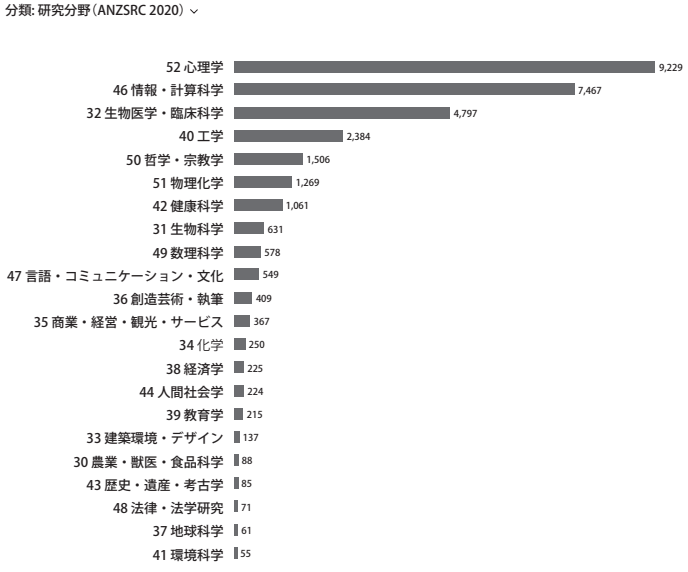


図２　図１と同じくDimensionsで集計された、“free energy principle” OR “active inference”を含む論文が出版された分野とそれぞれの分野における論文数（2025/12/19に取得）。

た。たとえば米国のテクノロジー／カルチャー誌ＷＩＲＥＤの長編記事では、自由エネルギー原理の野心的な側面と同時に、その難解さゆえに研究者間でも理解や評価が割れていることを伝えており、「すべてを説明する理論は結局何も説明できない危険性があるのでは？」といった疑問が取り上げられている（Raviv, 2018）。同様に自由エネルギー原理について包括的に紹介した英国発の一般向け科学誌New Scientistのより最近の記事でも「すべてを説明しようとするあまり、結局は何も説明できなくなるのではないか、と批判されている」という憂慮が紹介されている（Cutts, 2024）。

　また、自由エネルギー原理としばしば関連づけられるのが「ベイズ脳」と

いう考え方である。ベイズ脳仮説とは、脳は「世界で今何が起きているか?」という事前の推測と「いま入ってきた感覚情報」とをつき合わせながら、そのつど推測を更新している、という考え方だ。ここでいう更新の仕方が、統計学でいうベイズ推論に似ている、という意味で「ベイズ脳」と呼ばれる。たとえば知覚のベイズ理論については第2章で詳しく見ていくが、これは自由エネルギー原理と親和的ではあるものの、知覚という領域だけに適用することが念頭に置かれたものであり、十分な実験的な支持も受けている。そのような個別の理論を超えて、脳のはたらきすべてに適用できる理論にコミットするべき理由はなんなのだろうか。何でも説明する理論でも個別の現象を説明するには結局のところ個別の理論がいるのではないだろうか。そういった疑問も湧き上がってくるのだ。

また、そもそもこの原理は数学的に難解であるし、概念的にもこんがらがっている。自由エネルギー原理、能動推論、予測誤差最小化、予測符号化、予測処理、ベイズ脳仮説、仮説検証機としての脳などと言った、関連する枠組み、理論、仮説、用語が数多くあり、相互の関係性も不明瞭だ。一体この原理は、本当にさまざまな見解をまとめ、すべてを説明する統一理論なのか? 何をどこまで説明するのか? 実は何も説明しないのだろうか? 意識の問題も解決するのだろうか? 本書の大きな狙いの一つは、自由エネルギー原理を根本から捉え直すことで、こうした疑問に答えることである。難解でさまざまな期待と憂慮が向けられている自由エネルギーの実像を読者にわかりやすく提示し、その本当のインパクトを伝えたいのだ。

こうした取り組みを通じて本書で明らかにされる自由エネルギー原理の最大の魅力は、その理論の説明力というよりも、その理論がつなぐ生物の二つの側面にある。すなわち、自由エネルギー原理は

心と（生きている）身体との密接なつながりを明らかにするということだ。私たちは心をもつ存在であり、かつ生身の身体を持っている。この事実は当たり前のことのように思われるが、この二つの側面の関係は必ずしも明らかではなかった。

生成ＡＩが驚異的な進化を遂げる現代では、人間のように言語を操る生成ＡＩがいつか心を持つのではないかという疑問がしばしば検討される。しかしこの問いを考えるとき、これはすでに心を持つのに私たちのような身体は必要ではないかもしれないと考えていることになる。実際、心の哲学の有力な立場の一つである計算論的機能主義は、「心とはある種の計算に過ぎない」と考えてきた。そうすると、私たちの身体は（あるいは脳すら）たまたま心というソフトがインストールされたハードウェアにすぎず、それを機械の体にインストールすることも可能なように思われるのだ。

果たして、いつかコンピュータのシリコンでできた回路にも私たちと同じ「痛み」や「喜び」が宿るのだろうか。この問いに答えうるのが「自由エネルギー原理」なのだ。というのも自由エネルギー原理は、このもろく、常に栄養供給が必要で、だが長く持続する私たちの身体からなぜ心が生まれてくるのかを説明してくれるからだ。それゆえ本書のもう一つの目標は、自由エネルギー原理の無機質な数学的言語の中に潜む「生命の躍動」を、掘り出していくことにある。

以下に簡単に本書の旅路を紹介しておこう。第1章では自由エネルギー原理が、「生きている」とはどういうことかを数学の言葉で解き明かす理論だということが明らかにされる。「生きている」とは変化し続ける環境の中でいわば「自分らしさ」を守り抜くことであることが示される。第2章では、知覚を中心的に扱う。私たちの脳は、受動的に情報を受け取り外の世界の様子を映し出す鏡ではない。

私たちの脳はむしろ過去の経験から「次はこうなるはずだ」と予測し、その答え合わせを繰り返しながら、世界像を「作り上げる」器官なのだ。第3章では「イライラ」や「喜び」のような感情は、なぜ生まれるのかについて、自由エネルギー原理の観点から考察する。おおざっぱに言えば感情とは、脳が身体内部（心拍や体温）の異常を察知し、その異常を解消して生き延びるために下した「判断」のようなものである。感情はまさに心と体をつなぐ重要な心のはたらきであることが論じられる。その次の第4章では私たちのふるまいの諸相について検討する。私たちはなぜ、反射的に動くこともあれば、じっくり計画を立てることもあるのか。単純な「行動」から高度な「行為」への段階的な発展が統一的な枠組みのもとで説明される。また、意思決定とは未来をシミュレーションしてもっとも生存確率の高い選択肢を選ぶことだという説明が、脳の予測能力の観点から試みられる。続くインターミッションでは前半の導入編を総括する。私たちの心の働きが生きている身体の維持に究極的には動機づけられていることをみたあとは、この理論をより応用的な観点から検討していく。第5章では先ほど紹介した「この理論は、何でも説明できてしまう反証不可能な理論なのではないか?」といった哲学的な疑念などに応える。これらの疑問に答えることを通じて、自由エネルギー原理とそれに関係する理論には複数の層があることが明らかになる。自由エネルギー原理そのものは「生存概念の数理的分析」を提供し、それ自身は反証不可能だが、経験科学の具体的な理論と結びつくことで実質的な予測と反証が可能になるという構造をしていることを明らかにする。第6章では心の病や神経発達症（いわゆる発達障害）を、脳の「予測の失敗」として捉え直す。自閉スペクトラム症や妄想のような例だけではなく、現代社会の「病理」である陰謀論まであつかい、それらが単に「異常」だというより

は、「脳の適応の結果」としても理解できることを示す。

第7章、第8章では心の科学と哲学の最後のフロンティアである「意識」の領域に踏み込んでいく。第7章では私たちが意識しているものと、無意識のうちに処理されているものの境目はどこにあるのかについて、自由エネルギー原理の含意を探っていく。本章では有力な意識の科学的理論の一つであるグローバルワークスペース説との接続を模索し、脳内で勝ち残ったもっとも有力な予測が、私たちの「意識」として立ち上がると論じる。最終章である第8章では、先ほど簡単に触れたＡＩはいつか意識を持つのか、という問いについて正面から取り組む。この問いに対する本書の結論は「ＮＯ」だ。「死」から逃れる必要のないＡＩには、意識的な心が宿る可能性はないことを示し、生命と機械の間に（重要だが一時的な）境界線を引く。