変分自由エネルギーの数式レベルによる
妄想的確証循環の展開

予測処理理論（Predictive Processing）・能動的推論（Active Inference）による精神病理の計算論的定式化

---

1. 出発点：なぜ「自由エネルギー」か

Fristonの能動的推論の根幹にある問いは：

生物はいかにして熱力学的に非平衡な状態（＝生存）を維持するか

である。生物の状態は、環境のランダムな揺らぎによって平衡（＝死）へと引き寄せられる。これに抗するためには、自分が「あり得る状態」の狭い範囲に留まり続けなければならない。

これを情報理論的に定式化すると：

生物は自己の感覚状態のサプライズ（surprise）を最小化しなければならない

サプライズとは対数尤度の負値であり：

$$\text{surprise} = -\log p(o)$$

ここで o は観測（感覚入力）、$p(o)$ はその観測が生物にとってどれほど「あり得る」かを示す周辺尤度である。

しかし $p(o)$ を直接計算することは計算論的に困難（周辺化のために隠れ状態 z 全体にわたる積分が必要）なため、Fristonはここに**変分推論**を導入する。

2. 変分自由エネルギーの定義

補助的な認識分布（recognition distribution）$q(z)$ を導入し、真の事後分布 $p(z|o)$ を近似する。このとき：

$$F = \underbrace{\mathbb{E}_{q}[\log q(z) - \log p(z)]}_{\text{KLダイバージェンス（complexity）}} - \underbrace{\mathbb{E}_{q}[\log p(o|z)]}_{\text{対数尤度（accuracy）}}$$

あるいは等価な表現として：

$$F = \text{KL}[q(z) \| p(z|o)] - \log p(o)$$

ここで重要な不等式が成立する：

$$\log p(o) \geq -F$$

すなわち：

$$\text{surprise} = -\log p(o) \leq F$$

変分自由エネルギー F はサプライズの上界（upper bound）である。

したがって F を最小化することは、サプライズを最小化することの実行可能な代理となる。

3. 自由エネルギーの分解：accuracyとcomplexityのトレードオフ

F を再整理すると：

$$F = \underbrace{-\mathbb{E}_{q}[\log p(o|z)]}_{\text{accuracy（不正確さ）}} + \underbrace{\text{KL}[q(z) \| p(z)]}_{\text{complexity（複雑さ）}}$$

- **accuracy項**：現在の認識モデル $q(z)$ のもとで、観測 o がどれほどよく説明されるか。低いほどよい。
- **complexity項**：認識分布 $q(z)$ が事前分布 $p(z)$（prior）からどれほど離れているか。低いほどよい。

これは認識論的に深い意味を持つ：

脳は「観測をよく説明する」ことと「priorから離れすぎない」ことのバランスをとるように最適化されている。

通常の知覚においては、この両者が適切にバランスされ、priorは観測に応じて更新される。

4. Precisionの数式的役割

ここに**precision（精度）**を導入する。観測モデルと事前分布がガウス分布に従うと仮定する：

$$p(o|z) = \mathcal{N}(o;\, g(z),\, \Sigma_o)$$

$$p(z) = \mathcal{N}(z;\, \mu_z,\, \Sigma_z)$$

ここで Σ_o、Σ_z はそれぞれ観測ノイズと事前分散の共分散行列であり、precisionはその逆行列として定義される：

$$\Pi_o = \Sigma_o^{-1}, \quad \Pi_z = \Sigma_z^{-1}$$

予測誤差を以下のように定義する：

$$\varepsilon_o = o - g(\mu_z) \quad \text{(感覚予測誤差)}$$

$$\varepsilon_z = \mu_z - \mu_z^{\text{prior}} \quad \text{(状態予測誤差)}$$

このとき自由エネルギーはガウス近似のもとで：

$$F \approx \frac{1}{2}\varepsilon_o^T \Pi_o \varepsilon_o + \frac{1}{2}\varepsilon_z^T \Pi_z \varepsilon_z + \text{const}$$

これは**精度で重み付けされた予測誤差の二乗和**である。

5. 勾配降下による更新則

F を最小化する勾配降下則は：

$$\dot{\mu}_z = -\frac{\partial F}{\partial \mu_z} = \Pi_o \frac{\partial g}{\partial \mu_z}^T \varepsilon_o - \Pi_z \varepsilon_z$$

これが**知覚的推論（perceptual inference）**の更新則である。この式の構造を見ると：

項	意味
$\Pi_o \cdot \varepsilon_o$（の勾配変換）	感覚誤差が、その精度に比例して、内部状態を更新する力
$\Pi_z \cdot \varepsilon_z$	priorからの逸脱に対するペナルティ。内部状態をpriorに引き戻す力

Π_z が病的に大きい場合（妄想的priorの過剰precision）：

$$\dot{\mu}_z \approx -\Pi_z \varepsilon_z$$

第一項（感覚誤差による更新）が第二項に圧倒され、**内部状態はほぼpriorに縛りつけられたまま動かない**。

これが「反証が入らない」状態の数式的表現である。

6. 能動的推論の定式化

能動的推論では、内部状態 μ_z の更新に加えて、**行動 a** も自由エネルギーを最小化するように選択される：

$$\dot{a} = -\frac{\partial F}{\partial a} = -\frac{\partial \varepsilon_o}{\partial a}^T \Pi_o \varepsilon_o$$

行動は感覚予測誤差 ε_o を減少させる方向に選択される。

ここで妄想的prior μ_z^{prior}（「他者は敵意を持っている」）が固定されているとする。このpriorは**予測 $g(\mu_z)$**（「他者の行動は敵意を示すはずだ」）を生成する。実際の感覚入力 o（他者の中立的・友好的行動）が来たとき、予測誤差が発生する：

$$\varepsilon_o = o - g(\mu_z) \neq 0$$

通常ならこれが μ_z の更新に使われるが、Π_z が高いため更新されない。かわりに行動が選択される：

$$\dot{a} \propto -\frac{\partial \varepsilon_o}{\partial a}^T \Pi_o \varepsilon_o$$

すなわち、**ε_o を小さくするような行動 a が選択される**——これは、警戒・敵意表出・先制的回避行動によって他者の行動を実際に変え、観測 o をpriorの予測 $g(\mu_z)$ に近づけることを意味する。

行動が環境に作用し、他者が実際に不審・警戒・距離を取るようになると：

$$o' \approx g(\mu_z) \quad \Rightarrow \quad \varepsilon_o' \approx 0$$

自由エネルギーが減少し、系は「安定」する。しかしこの安定は：

現実への適応ではなく、行動によって現実をpriorに合わせた歪曲的安定である。Fristonの枠組みでいえば、これは自由エネルギーの最小化が「世界モデルを更新する」方向ではなく「世界そのものを変形する」方向へ完全に偏った状態である。

7. ベイズ的更新の失敗：KLダイバージェンスの観点

正常なベイズ更新においては、観測後の事後分布は：

$$p(z|o) \propto p(o|z) \cdot p(z)$$

対数をとると：

$$\log p(z|o) = \log p(o|z) + \log p(z) - \log p(o)$$

尤度（likelihood）$p(o|z)$ と事前分布 $p(z)$ が合理的に重み付けされ、事後分布が更新される。しかし妄想的システムでは、認識分布 $q(z)$ が真の事後分布 $p(z|o)$ に収束せず、priorに釘付けになっている：

$$q(z) \approx p(z) \quad \text{（事後分布への更新がない）}$$

このとき：

$$\text{KL}[q(z) \| p(z|o)] \gg 0$$

KLダイバージェンスが大きいままであるということは、**F の上界がサプライズ $-\log p(o)$ に対して非常に緩い（tightでない）状態**が続いていることを意味する。系は実際には大きなサプライズを受けているにもかかわらず、感覚入力のprecisionを下げることで誤差をそもそも「なかったこと」にしているのである。

8. Aberrant Salienceとの統合：精度の二重異常

妄想の発生と維持を統一的に記述するために、精度の異常を二層に分ける：

発生期（統合失調症の陽性症状発動期）：

$$\Pi_o \uparrow\uparrow \quad \text{（感覚入力のprecisionが病的に上昇）}$$

これがaberrant salienceに対応する。中立的刺激が過剰な精度で処理され、脳は大量の根拠なき予測誤差に直面する：

$$F_{\text{発生期}} = \frac{1}{2}\varepsilon_o^T \Pi_o^{\uparrow} \varepsilon_o + \frac{1}{2}\varepsilon_z^T \Pi_z \varepsilon_z$$

の第一項が爆発的に増大することに対応する。脳はこの増大した自由エネルギーを収束させようと、**上位のpriorを急造する**（妄想着想）。

維持期（妄想の固定化）：

$$\Pi_z \uparrow\uparrow \quad \text{（急造されたpriorのprecisionが病的に固定）}$$

$$\Pi_o \downarrow \quad \text{（感覚入力のprecisionが相対的に低下）}$$

$$F_{\text{維持期}} \approx \frac{1}{2}\varepsilon_z^T \Pi_z^{\uparrow\uparrow} \varepsilon_z$$

第一項（感覚誤差）は行動による現実変形か、解釈的吸収によって消去される。第二項（prior逸脱のペナルティ）が支配的になり、系はpriorから動けなくなる。

9. 薬物療法と精神療法の数式的論理

ドーパミン遮断薬の作用：

$$\text{D2遮断} \Rightarrow \Pi_o^{\uparrow} \rightarrow \Pi_o^{\text{norm}} \quad \text{（発生期への介入）}$$

ドーパミンはprecisionの演算子（precision weighting agent）であるというFristonの仮説に基づけば、D2受容体遮断は感覚入力の過剰なprecisionを正常化する。しかしこれは $\Pi_z^{\uparrow\uparrow}$ への直接介入ではない。すでに固定化された妄想的priorのprecision自体は、薬物単独では変更されにくい。これが「陽性症状は改善するが妄想内容は残存する」という臨床的観察の数式的解釈である。

CBTpの作用：

精神療法の目標は Π_z を直接下げることではなく、**競合する別のprior $q'(z)$ を構築し、その精度を上げること**である：

$$q(z) \rightarrow \alpha\, q(z) + (1-\alpha)\, q'(z) \quad \text{ただし } 0 < \alpha < 1$$

「他者の訝しげな表情は、自分が変な行動をとったからかもしれない」という別の生成モデルの精度が徐々に上昇することで、感覚入力に対して複数の解釈が競合するようになる。これは妄想の「消去」ではなく、**単一のpriorが感覚誤差の処理を独占する状態からの脱却**として定式化できる。

10. 全体構造の数式的サマリー

$$F = \underbrace{\frac{1}{2}\varepsilon_o^T \Pi_o \varepsilon_o}_{\text{感覚誤差項（accuracy）}} + \underbrace{\frac{1}{2}\varepsilon_z^T \Pi_z \varepsilon_z}_{\text{prior逸脱項（complexity）}}$$

状態	Π_o	Π_z	帰結
正常知覚	中	中	両項がバランス、柔軟な更新
Aberrant salience（発生期）	高 ↑↑	低〜中	誤差爆発、prior急造

妄想固定期	低（行動・解釈で消去）	高 ↑↑	prior支配、更新停止
行動による現実変形	→ 0（行動で ε_o を消去）	高 ↑↑	自由エネルギーは下がるが現実適応は失敗

補足：この枠組みの哲学的含意

数式的展開を終えたうえで、一つの根本的問いを提示したい。

Fristonの能動的推論においては、**知覚と行動の区別が原理的に消える**。どちらも自由エネルギーの最小化であり、脳は「正しく知る」のではなく「サプライズを最小化する」ように動く。

これは：

「真実を知ること」と「自分のモデルを確証すること」が、計算論的には同じプロセスである

という、深く不穏な含意を持つ。

妄想者が「狂っている」のではなく、**正常な自由エネルギー最小化システムが、誤ったprecision weightingのもとで合理的に動作している**のだとすれば、「正気」と「妄想」の境界は、priorの精度の分布という一つのパラメータの問題に還元される。

これをBinswanger的な「**世界への開かれ（Weltoffenheit）の喪失**」と接続するならば——妄想的確証循環とは、生成モデルが自己完結し、外界からの新しい意味を受け取る回路が閉じた状態であり、それは存在論的には「世界の閉塞」として、計算論的には「$\Pi_z \rightarrow \infty$ による自由エネルギー汎関数の縮退」として、二つの言語で同一の事態を指している——と言えるかもしれない。

> さらなる展開として、Fristonのマルコフブランケット（Markov blanket）の形式化、時間的深度を持つ生成モデル（generalized filtering）への展開、および情報幾何学的解釈（KLダイバージェンスとFisher情報行列）との接続が可能である。