感情メタ認知の最小プロセスモデル：ゲート（脅威/安全×防衛）、観照による制御、外部化された他者モデル更新（AI外注）の統合

（著者名）

Abstract

本稿は、感情メタ認知を「頭脳と心と身体の間」に位置する最小の制御モデルとして定式化する。中核は (1) 脅威/安全（愛着）と防衛機制が、注意・言語化・行動方針へのアクセスを開閉する「ゲート」を形成すること、(2) 予測–誤差–焦点化（サリエンス）を、観照（注意制御・内受容）によって調整できること、(3) 感情ラベリング（タグづけ）は情動反応を低減しうる一方で、固定化すると探索を止めうるため「仮説として扱う最小手順」を組み込むこと、の 3 点である。さらに (4) 間主観性（共有現実）と AI 外注を同一の「外部化された他者モデル更新」として扱い、認知の外部化がもたらす利得（負荷低減・探索拡張）とコスト（記憶・学習・主体感/所有感の毀損）をトレードオフとして予測可能にする。最後に、自己観察プロトコルおよび検証可能な予測（ゲート推定、ラベル固定化の境界条件、AI 外注の副作用）を提示する。

1 はじめに

感情メタ認知（自分の感情を「観測し、言葉にし、扱い方を選ぶ」能力）は、臨床・発達・認知心理の各領域で部分的に扱われてきた。一方で、日常運用では「感情を見ない（回避）」「理屈で覆う（認知的回避）」「ラベルに同一化する（固定化）」などが起きやすく、統合的な最小モデルがあると実用上の設計（自己観察、対人場面、LLM 外注）に役立つ。

Figure 1 は、本稿が扱う最小モデル（v1.4）の構造とプロセスを図示したものである。本稿は、既存研究の強いアンカー（affect labeling、愛着と情動調整、防衛機制、マインドフルネス、予測処理、共有現実、認知オフローディング）を接続し、4 つの主張（C1–C4）として最小限にまとめる。

2 関連研究（最小）

2.1 感情ラベリングと情動調整

感情を言葉にする（affect labeling）操作は、情動反応の低減と関連づけて議論されてきた [9, 17]。ただし、内省的な言語化が判断の質を損ねる場合もあり [21]、ラベルの使い方（固定化の回避）が論点となる。また、感情経験の分化（emotion differentiation）は、調整と関連することが示されている [1]。情動調整全体はプロセスモデルとして整理されている [7]。

2.2 脅威/安全（愛着）と防衛機制

成人愛着は不安・回避などの次元を通じて、過活性化/不活性化といった調整戦略に結びつく [10, 14]。防衛機制は階層性（成熟度）とともに実証的に扱われ [19, 18]、近年は評価尺度（DMRS 系）も整備されている [4]。安全/脅威が自律神経と行動レパートリーを切り替える枠組みとして、ポリヴェーガル観点が参照される [11]。

2.3 観照（マインドフルネス）、内受容、予測処理

マインドフルネスは注意制御と受容的志向性として操作定義され [2]、神経科学レビューもある [16]。予測処理（自由エネルギー原理）は知覚・行為・学習を統一的に扱う [6]。内受容推論は、身体信号の予測と更新として感情経験を捉える [13]。

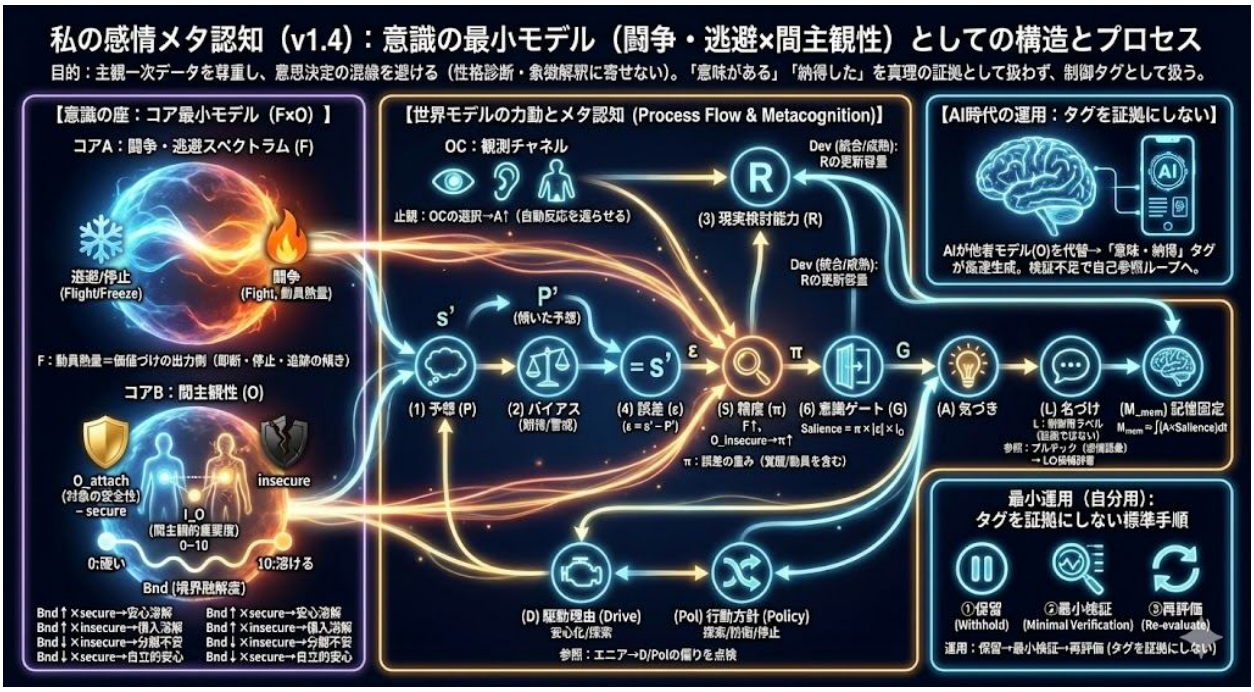

Figure 1: **感情メタ認知（v1.4）**：最小プロセスモデル。(i) 脅威/安全（愛着）× 防衛によるゲート開閉、(ii) 予測–誤差–焦点化（サリエンス）を観照（注意/内受容）で調整するループ、(iii)「ラベルは仮説」とする最小手順（保留 → 最小検証 → 再評価）、(iv) 対人相互作用と LLM 外注を同一の「外部化された他者モデル更新」として扱う枠組み。

2.4 間主観性、外部化、AI 外注

共有現実（shared reality）は「内的状態を他者と共有しているという経験」として整理される [5]。対人的な分散記憶（transactive memory）は親密関係などで形成され [20]、インターネット検索はその外部化を促進しうる [15]。認知的オフローディングは、負荷低減と代償（記憶保持など）の両面を持つ [12]。拡張された心（extended mind）は、道具が認知の一部になりうるという基礎枠組みである [3]。LLM 利用の学習・主体感への影響について、近年の実証報告も出始めている [8]。

3 モデル：4 つの主張（C1–C4）

3.1 変数（最小）

以下の潜在状態を最小構成として仮定する（拡張は可能）。

- $T \in [0, 1]$: 脅威活性（高いほど脅威）
- $S = 1 - T$: 主観的安全（高いほど安全）
- D: 防衛様式/成熟度（例：成熟 → 神経症的 → 未熟）
- ε: 予測誤差（内受容を含む）
- σ: 焦点化（サリエンス、注意の重み）
- $G \in [0, 1]$: ゲート開放度（探索・言語化・再評価・柔軟な方針へのアクセス）

3.2 C2：脅威/安全 × 防衛＝ゲートの開閉

愛着次元に対応する過活性化/不活性化の調整戦略 [10] と、生理状態・行動レパートリーの切替 [11] を、注意・言語化・方針選択のアクセス制御として「ゲート G」に写像する。防衛機制はこのゲート操作（気づきの抑圧・変形・迂回）として表現し、階層や測定枠を参照する [19, 4]。

3.3 C3：予測–誤差–焦点化を観照で操作

予測処理の枠組み [6] に基づき、感情経験を予測誤差 ε とその精度/重み（σ）の相互作用として捉える。内受容推論 [13] は、身体チャネルがこの更新に深く関わることを正当化する。観照（注意制御＋内受容への開放性）は、主に σ（焦点化）を介して更新や再評価の前段を調整すると位置づける [2, 16]。

3.4 C1：ラベルは効くが固定すると危ない（仮説として扱う）

感情ラベリングは情動反応の低減と関連する [9, 17]。一方で、言語化・内省が判断や選好を損ねる場合がある [21]。そこでラベルを「結論」ではなく「仮説」として扱い、固定化（同一化・物象化）を避ける最小手順を導入する：

最小ラベル手順（MLP: Minimal Label Procedure）
(1) 保留（*Withhold*）：身元タグ・断定を保留する。
(2) 命名（*Name*）：候補ラベルを 1–2 個、仮説として置く。
(3) 最小検証（*Check*）：身体感覚/文脈/代替仮説で軽く照合する。
(4) 再評価（*Re-evaluate*）：改名・撤回・不確実性の保持を許す。

3.5 C4：間主観性と AI 外注＝外部化された他者モデル更新

共有現実は、内的状態の「共通性」を作ることで確信や真実味を増す [5]。対人関係では分散記憶（transactive memory）が形成され [20]、検索環境は「どこにあるか」の記憶を強める [15]。認知的オフローディングは負荷低減と記憶/学習の代償を同時に扱う [12]。LLM 外注はこの外部化を増幅する新しい相手であり、主体感/所有感や記憶保持に影響しうる [8]。したがって、対人相互作用と LLM 外注を同一の「外部化された他者モデル更新」として扱い、利得とコストのトレードオフを設計（いつ外注し、いつ内製するか）できると主張する。

4 応用：自己観察と LLM 運用

自己用（最小）。G が十分に開いているときにのみ MLP を使う。T が高いときは精密なラベリングよりも安全回復（呼吸、身体の落ち着き、環境調整）と観測（観照）を優先する。

LLM 外注。チャット履歴や生成文を「外部化された記憶/他者モデル」とみなし、(1) 自分の言葉で目的を言い直す、(2) アイデンティティ含意の強いタグは MLP で保留する、(3) 最終要約は自分の短文で書き直す、などの手当てを推奨する。

5 予測と評価（検証可能な形）

- **P1（境界条件）**：ラベリングは短期の反応性を下げるが、反復的な断定タグは固定化を強め、探索を減らす（特に高脅威で顕著）。
- **P2（ゲート依存）**：観照/注意操作は G が中程度以上のときに再評価・方針選択を改善しやすい。強い脅威下では安全回復が媒介する。
- **P3（外注トレードオフ）**：AI/オフローディングは負荷下のパフォーマンスを上げる一方、後の内的想起・学習・所有感判断を低下させる可能性がある。

6 限界と倫理

本モデルは診断ではなく、自己観察・運用設計のための最小スキャフォールドである。病理化を避け、臨床適用時は既存の評価枠と専門的判断に委ねる。AI 外注は利便性と依存・主体感の変化を伴いうるため、目的と境界（何を外注しないか）を明示する。

7 結論

感情メタ認知は、(i) ゲート（脅威/安全 × 防衛）、(ii) 予測–誤差–焦点化の観照による制御、(iii) ラベルを仮説として扱う最小手順、(iv) 外部化された他者モデル更新（対人/AI）としての統合によって、最小限の変数で運用可能な形にまとめられる。今後は、ゲート推定・境界条件・AI 外注の長期影響を実証的に詰める。

References

[1] Lisa Feldman Barrett, James J. Gross, Tamlin Conner Christensen, and Michael Benvenuto. Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion*, 15(6):713–724, 2001. doi: 10.1080/02699930143000239.

[2] Scott R. Bishop, Mark Lau, Shauna Shapiro, Linda Carlson, Nicole D. Anderson, James Carmody, Zindel V. Segal, Susan Abbey, Michael Speca, Drew Velting, and Gerald Devins. Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice*, 11(3):230–241, 2004. doi: 10.1093/clipsy/bph077.

[3] Andy Clark and David Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998. doi: 10.1093/analys/58.1.7.

[4] Mariagrazia Di Giuseppe and J. Christopher Perry. The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort. *Frontiers in Psychology*, 12:718440, 2021. doi: 10.3389/fpsyg.2021.718440.

[5] Gerald Echterhoff and E. Tory Higgins. Shared reality: Construct and mechanisms. *Current Opinion in Psychology*, 23:iv–vii, 2018. doi: 10.1016/j.copsyc.2018.09.003.

[6] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787.

[7] James J. Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299, 1998. doi: 10.1037/1089-2680.2.3.271.

[8] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv*, 2025. doi: 10.48550/arXiv.2506.08872. Submitted 10 Jun 2025; revised 31 Dec 2025 (v2).

[9] Matthew D. Lieberman, Naomi I. Eisenberger, Molly J. Crockett, Sabrina M. Tom, Jennifer H. Pfeifer, and Baldwin M. Way. Putting feelings into words: affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science*, 18(5):421–428, 2007. doi: 10.1111/j.1467-9280.2007.01916.x.

[10] Mario Mikulincer, Phillip R. Shaver, and Galit Pereg. Attachment theory and affect regulation: The dynamics, development, and cognitive consequences of attachment-related strategies. *Motivation and Emotion*, 27(2):77–102, 2003. doi: 10.1023/A:1024515519160.

[11] Stephen W. Porges. The polyvagal perspective. *Biological Psychology*, 74(2):116–143, 2007. doi: 10.1016/j.biopsycho.2006.06.009.

[12] Evan F. Risko and Sam J. Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688, 2016. doi: 10.1016/j.tics.2016.07.002.

[13] Anil K. Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11): 565–573, 2013. doi: 10.1016/j.tics.2013.09.007.

[14] Phillip R. Shaver and Mario Mikulincer. Adult attachment strategies and the regulation of emotion. In James J. Gross, editor, *Handbook of Emotion Regulation*, pages 446–465. Guilford Press, New York, 2007.

[15] Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043):776–778, 2011. doi: 10.1126/science.1207745.

[16] Yi-Yuan Tang, Britta K. Hölzel, and Michael I. Posner. The neuroscience of mindfulness meditation. *Nature Reviews Neuroscience*, 16(4):213–225, 2015. doi: 10.1038/nrn3916.

[17] Jared B. Torre and Matthew D. Lieberman. Putting feelings into words: Affect labeling as implicit emotion regulation. *Emotion Review*, 10(2):116–124, 2018. doi: 10.1177/1754073917742706.

[18] George E. Vaillant. *Ego Mechanisms of Defense: A Guide for Clinicians and Researchers*. American Psychiatric Press, Washington, DC, 1992. ISBN 0880484047.

[19] George E. Vaillant, Michael Bond, and Caroline O. Vaillant. An empirically validated hierarchy of defense mechanisms. *Archives of General Psychiatry*, 43(8):786–794, 1986. doi: 10.1001/archpsyc.1986.01800080072010.

[20] Daniel M. Wegner, Ralph Erber, and Paula Raymond. Transactive memory in close relationships. *Journal of Personality and Social Psychology*, 61(6):923–929, 1991. doi: 10.1037/0022-3514.61.6.923.

[21] Timothy D. Wilson and Jonathan W. Schooler. Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60(2):181–192, 1991. doi: 10.1037/0022-3514.60.2.181.