A Minimal Process Model of Affective Metacognition: Gating (Threat/Safety $\times$ Defense), Contemplative Control, and Externalized Other-Model Updating (AI Outsourcing)

(anonymous draft)

Abstract

We propose a minimal process model of affective metacognition that bridges "brain–mind–body" dynamics with practical protocols for self-monitoring and AI-assisted cognition. The model combines (i) a gating mechanism driven by threat/safety appraisal (attachment) and defense processes, (ii) contemplative control that modulates prediction–error–salience via attention and interoception, and (iii) a label-as-hypothesis procedure: affect labels can down-regulate reactivity yet become harmful when reified. Finally, we treat intersubjectivity (shared reality) and AI outsourcing as the same class of externalized other-model updating, yielding testable trade-offs between cognitive relief and costs to memory, learning, and authorship/ownership.

1 Introduction

Figure 1 illustrates the proposed minimal process model and its operational protocol.

Affective metacognition—monitoring and steering one's emotional processes—often feels like a "between" space connecting brain, mind, and body. Existing literatures offer partial lenses: affect labeling reduces reactivity [9, 17], attachment organizes safety vs. threat regulation strategies [10, 14], mindfulness operationalizes attention and acceptance [2, 16], and predictive processing provides a unifying control language [6, 13]. Meanwhile, cognition is increasingly externalized to devices and LLMs; this resembles transactive memory and cognitive offloading [20, 12, 15], and may affect ownership of writing [8].

We unify these into a minimal model with four claims (C1–C4 in the accompanying state file). The goal is not a comprehensive theory, but a compact scaffold that (a) maps to measurable constructs, (b) yields practical self-protocols, and (c) supports design decisions for AI outsourcing without over-identifying with "tags."

2 Related Work (Very Brief)

Labeling and differentiation. Affect labeling can attenuate limbic responses [9] and is framed as implicit emotion regulation [17]. Yet verbal analysis can degrade preference quality in some contexts [21], motivating boundary conditions and careful label use. Emotion differentiation relates to regulation outcomes [1] and sits within broader process models of regulation [7].

Threat/safety and defenses. Attachment orientations are linked to hyperactivating vs. deactivating regulation strategies [10, 14]. Defense mechanisms admit hierarchical organization and measurement approaches [19, 18, 4]. Polyvagal theory provides a physiological lens on safety/threat states and behavioral repertoires [11].

Predictive processing and contemplative control. The free-energy principle frames perception/action as prediction error minimization [6]. Interoceptive inference connects bodily signals to felt emotion and selfhood [13]. Mindfulness is operationalized as attention regulation plus an accepting orientation [2], and has surveyed neural mechanisms [16].

Externalization, others, and AI. Shared reality describes motivation and mechanisms for aligning inner states [5]. Transactive memory shows distributed encoding in close relationships [20], and the Internet can behave like an external memory system [15]. Cognitive offloading formalizes when and why we outsource cognition [12]. The extended mind argues that external resources can be constitutive of cognition [3].

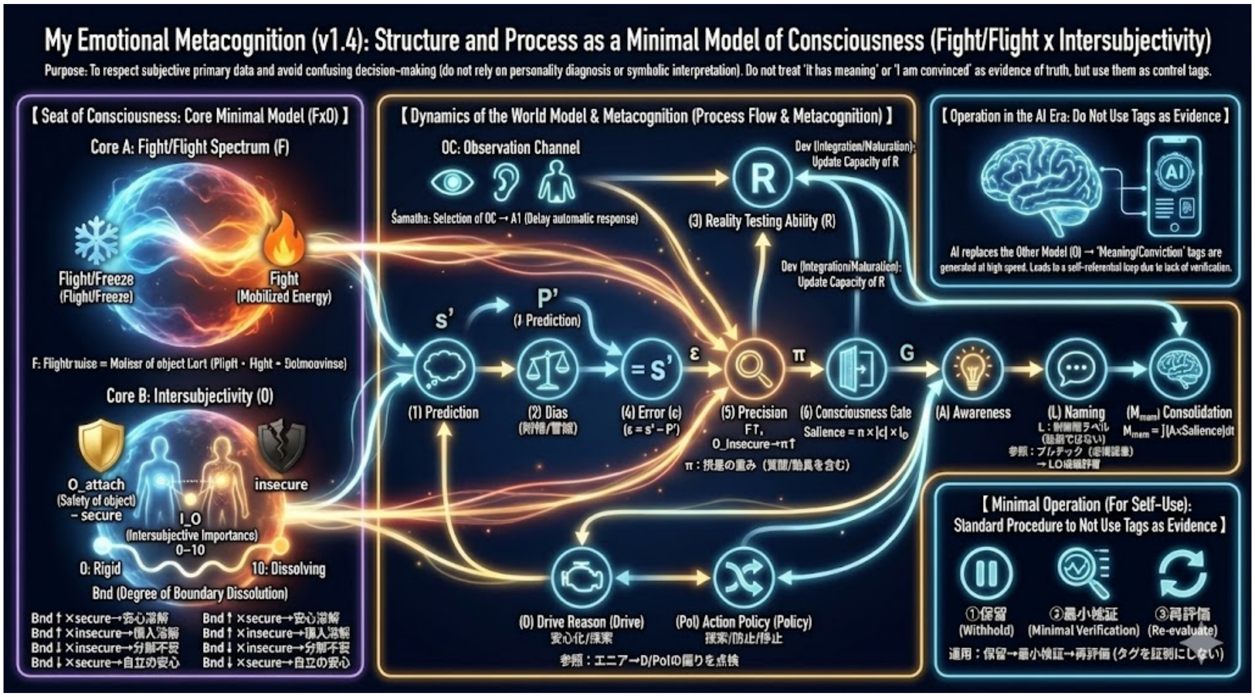
My Emotional Metacognition (v1.4): Structure and Process as a Minimal Model of Consciousness (Fight/Flight x Intersubjectivity)
Purpose: To respect subjective primary data and avoid confusing decision-making (do not rely on personality diagnosis or symbolic interpretation). Do not treat 'it has meaning' or 'I am convinced' as evidence of truth, but use them as contrel tags.
【Seat of Consciousness: Core Minimal Model (FxO)】
Core A: Fight/Flight Spectrum (F)
Flight/Freeze
(Flight/Freeze)
Fight
(Mobilized Energy)
F: Flightrιuise = Molisr of object Lort (Plight · Hght · Dolmoovinse)
Core B: Intersubjectivity (O)
O_attach
(Safety of object)
– secure
insecure
I_O
(Intersubjective Importance)
0–10
0: Rigid
10: Dissolving
Bnd (Degree of Boundary Dissolution)
Bnd↑×secure→安心溶解
Bnd↑×insecure→侵入溶解
Bnd↓×insecure→分離不安
Bnd↓×secure→自立の安心
Bnd↑×secure→安心溶解
Bnd↑×insecure→侵入溶解
Bnd↓×insecure→分離不安
Bnd↓×secure→自立の安心
【Dynamics of the World Model & Metacognition (Process Flow & Metacognition)】
OC: Observation Channel
Śamatha: Selection of OC → A1 (Delay automatic response)
R
(3) Reality Testing Ability (R)
Dev (Integration/Maturation): Update Capacity of R
Dev (Integration/Maturation): Update Capacity of R
s'
P'
(↓ Prediction)
= s'
ε
π
G
(1) Prediction
(2) Dias
(4) Error (c)
(ε = s' – P')
(5) Precision
FT,
O_Insecure→π↑
(6) Consciousness Gate
Salience = π × |c| × I_O
(A) Awareness
(L) Naming
L：制御用ラベル
(証拠ではない)
(M_mem) Consolidation
M_mem = ∫(A×Salience)dt
(D) Drive Reason (Drive)
安心化/探索
(Pol) Action Policy (Policy)
探索/防止/停止
参照：エニアー→D/Polの偏りを点検
【Operation in the AI Era: Do Not Use Tags as Evidence】
AI
AI replaces the Other Model (O) → 'Meaning/Conviction' tags are generated at high speed. Leads to a self-referential loop due to lack of verification.
【Minimal Operation (For Self-Use): Standard Procedure to Not Use Tags as Evidence】
①保留
(Withhold)
②最小検証
(Minimal Verification)
③再評価
(Re-evaluate)
適用：保留→最小検証→再評価（タグを証明にしない）

Figure 1: **Affective metacognition (v1.4):** a minimal process model integrating (i) threat/safety-driven gating (attachment × defenses), (ii) a prediction–error–salience loop modulated by contemplative observation (attention/interoception), (iii) a minimal "label-as-hypothesis" procedure (withhold → check → re-evaluate), and (iv) externalized other-model updating via social interaction and LLM outsourcing.

3 Model

3.1 Core variables and notation

We define a minimal set of latent state variables (scalars; extensions are possible):

- $T \in [0, 1]$: threat activation (higher = more threatened).
- $S = 1 - T$: perceived safety (higher = safer).
- D: dominant defense mode/level (e.g., mature → neurotic → immature).
- ε: prediction error (including interoceptive prediction error).
- σ: salience (attention weight) assigned to a channel/content.
- $G \in [0, 1]$: gate openness (access to exploration, labeling, reappraisal, flexible policy).

3.2 Gating: threat/safety × defense

We model a gating function:

$$G = f(S, D) \quad \text{with } \frac{\partial G}{\partial S} > 0, \tag{1}$$

and defense-dependent modulation (e.g., deactivation can keep G low by narrowing channels; hyperactivation can keep G noisy/high but unstable). Empirically, attachment insecurity correlates with deactivating vs. hyperactivating strategies [10]. Physiologically, safety/threat states reconfigure available behaviors [11]. Defenses provide a clinical language and hierarchy that can be measured [19, 4].

3.3 Contemplative control: prediction–error–salience loop

We assume salience allocation σ shapes how prediction error ε is amplified into felt urgency and subsequent policy selection. "Contemplative control" denotes deliberate regulation of σ (attention) and interoceptive sampling, consistent with operational definitions of mindfulness [2]. Predictive processing provides the control intuition: changing attention/precision alters the influence of prediction error [6, 13].

A minimal loop:

$$\text{(Observe)} \Rightarrow \varepsilon \tag{2}$$

$$\text{(Set precision / salience)} \Rightarrow \sigma \tag{3}$$

$$\text{(Update / choose policy)} \Rightarrow \pi \tag{4}$$

where mindfulness-like operations act primarily on σ (and secondarily on the generative model that produces ε).

3.4 Label-as-hypothesis: a minimal safe procedure

Affect labels can reduce reactivity [9, 17], but reification can freeze exploration or distort preferences [21]. We propose a minimal procedure that treats labels as hypotheses:

Minimal Label Procedure (MLP)
(1) *Withhold*: pause; do not finalize identity-level tags.
(2) *Name*: propose 1–2 candidate labels as hypotheses.
(3) *Check*: verify against body/interoception and context; seek alternatives.
(4) *Re-evaluate*: rename or drop; keep uncertainty explicit.

This aims to preserve the benefits of labeling while reducing fixation.

3.5 Externalized other-model updating: human and AI as one class

Let O denote an "other model" used for calibration of beliefs/feelings (shared reality) [5]. Interaction with another person or an LLM both instantiate an external update channel:

$$\theta_{t+1} = \theta_t + \eta \cdot U(\text{dialogue with } O), \tag{5}$$

where θ are internal models (self, others, world). This channel overlaps with cognitive offloading and transactive memory: distributing storage/processing can reduce load [12, 20], but can shift what is remembered ("where" over "what") [15]. Recent work suggests LLM interaction can influence perceived ownership of writing [8], making it essential to specify when outsourcing is beneficial vs. costly.

4 Applications

Self-use (minimal). Use MLP only when G is sufficiently open; when T is high, prioritize safety restoration (downshift) and basic observation rather than fine-grained labeling.

Clinical/developmental connection (optional layer). Map recurring patterns of low G to attachment and defense profiles, and track shifts in D or T across time using validated instruments [4].

LLM outsourcing. Treat prompts and chat history as externalized memory/other-model. Design safeguards: (i) keep "authorial core" notes internal, (ii) periodically restate goals in one's own words, (iii) run MLP before committing to identity-laden labels suggested by the model.

5 Predictions (Testable)

- **P1 (Boundary condition):** labeling reduces immediate reactivity, but repeated identity-level labels increase fixation and reduce exploration, especially under high threat (low S).

- **P2 (Gate dependence):** mindfulness/attention manipulations improve reappraisal mainly when G is moderately open; under strong threat, effects are mediated by safety restoration.
- **P3 (Outsourcing trade-off):** AI/offloading improves task performance under load but reduces later internal recall and may alter authorship/ownership judgments.

6 Limitations

This is a minimal scaffold. It does not diagnose, and it abstracts complex constructs (attachment, defenses, predictive processing) into low-dimensional variables. Empirical work should specify measurement models and boundary conditions.

7 Conclusion

Affective metacognition can be made tractable by a compact control model: gating (threat/safety × defense), contemplative control (prediction–error–salience), and safe labeling (hypothesis procedure), extended to modern cognition via externalized other-model updating (human/AI).

References

[1] Lisa Feldman Barrett, James J. Gross, Todd C. 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 et al. 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.

[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 Pub, 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.