

From Instinct to Reflective Agency: An Interdisciplinary Framework for Consciousness Integrating Evolutionary Biology and the Bhagavad Gītā

Dr. K. V. Sarma

Independent Researcher, Hyderabad, Telangana, India
ORCID: 0009-0001-1961-5377

Abstract

Background: Contemporary discussions of consciousness often treat biological instinct and deliberate moral action as opposed categories, without a shared vocabulary for the transitional process between them. Classical Indian philosophy, particularly the Bhagavad Gītā, offers a detailed phenomenology of this transition through the crisis and resolution of the warrior Arjuna, but this account is rarely brought into direct dialogue with contemporary evolutionary and cognitive science.

Objective: This paper proposes the **Reflective Agency Model (RAM)** of consciousness — a theoretical framework describing consciousness as the progressive expansion of self-governance over biologically conditioned impulses through discernment, deliberation, and intentional action — and examines its convergence with the Gītā's account of Arjuna's movement from instinctual paralysis to reasoned duty.

Method: This is a theoretical and comparative-philosophical study. It synthesizes literature from evolutionary biology, cognitive neuroscience, and psychology on impulse regulation and dual-process cognition with a close textual reading of relevant passages from the Bhagavad Gītā and allied Vedāntic sources, using comparative hermeneutics as its primary method.

Findings/Contribution: RAM offers a testable, cross-culturally coherent vocabulary for the instinct-to-agency transition; the Gītā's Arjuna narrative shows a substantial structural correspondence with mechanisms described independently in dual-process and self-regulation research; and the Gītā's near mirror-image vocabulary of collapse (2.62–63) and recovery (18.73) supplies internal textual confirmation of the model's two-pathway structure.

Implications: RAM offers a bridge for future empirical and contemplative research on discernment, moral development, and self-regulation.

Keywords

Consciousness; reflective agency; Bhagavad Gītā; dual-process cognition; self-regulation; comparative philosophy

1. Introduction

Human beings share the basic biological rhythm of every organism on Earth: securing nourishment, rest, and continuation. Sanskritic and vernacular Indian traditions capture this baseline economically; a common Telugu formulation — తిన్నామా, పడుకున్నామా, తెల్లారిందా (“did we eat, did we sleep, has the sun risen?”) — names the unadorned cycle of biological maintenance that underlies all higher activity. Evolutionary biology describes the same baseline in different terms: organisms are shaped by selection pressures oriented toward survival and reproduction, without any inherent reference to moral categories (Robert M. Sapolsky, 2017).

And yet human beings do not experience themselves as governed solely by this baseline. Across cultures, humans construct systems of ethics, law, and religious obligation that claim to orient action toward ends beyond immediate survival. A long-standing question — addressed independently by evolutionary theorists, cognitive scientists, and classical Indian philosophers — is whether these systems represent a genuine expansion of agency beyond biological conditioning, or whether they function as more elaborate expressions of the same underlying survival logic.

This paper takes up that question directly, using two literatures that are rarely read together: (a) contemporary research on dual-process cognition, self-regulation, and impulse control, which describes at a mechanistic level how organisms move from automatic to deliberate response; and (b) the Bhagavad Gītā’s account of Arjuna’s crisis on the battlefield of Kurukṣetra, which offers an unusually detailed pre-modern narrative treatment of exactly this transition — from instinct-driven paralysis, through dialogue and reframing, to autonomous, deliberate action.

The present inquiry did not begin purely as a literature synthesis. It extends and formalizes a framework already applied, though not previously named as such, in the author’s earlier self-published practitioner monograph (Sarma. K. V, 2016), which drew on several decades of one-on-one counseling practice. That work used the *buddhi*-hierarchy of Bhagavad Gītā 3.42 — body, senses, mind, intellect, and Self — alongside the Gītā’s account of how attachment to sense-objects gives rise to desire, desire to anger, anger to delusion, delusion to loss of memory, and finally to the collapse of discrimination itself (Bhagavad Gītā 2.62–63), as a working map for helping individuals interrupt exactly that cascade — settling a wandering mind and an agitated body toward a more stable baseline of wellbeing (Sarma K. V, 2016). The present paper takes that same textual pairing — Gītā 3.42’s hierarchy of governance and 2.62–63’s account of its breakdown — and reformulates it as a general theoretical model (RAM), extending a decade-old applied framework into a form suited to comparison with the contemporary cognitive-scientific literature. Because the counseling practice informing this work was individualized and non-standardized — consistent with the Gītā’s own account of each person’s unique constitution under the three *guṇas* (Chapters 14, 17–18) — it is offered here as originating context for RAM, not as clinical evidence; a more

systematic, ethically governed investigation of these observations is identified as a direction for future research in Section 9.

The paper's central theoretical contribution is the **Reflective Agency Model (RAM)**, introduced in Section 5, which formalizes this transition as a describable psychological process rather than a purely narrative or devotional one. RAM is offered as a framework capable of being tested, refined, or falsified by future empirical and contemplative research, not merely as a restatement of either tradition alone.

2. Research Question

Primary research question:

To what extent can the transition from instinct-driven reaction to deliberate, value-aligned action — as described independently in contemporary cognitive science and in the Bhagavad Gītā's account of Arjuna — be integrated into a single, cross-traditional theoretical model of consciousness?

Subsidiary questions:

1. What structural features do evolutionary-biological and cognitive accounts of impulse regulation share with the Gītā's depiction of Arjuna's crisis and resolution?
2. Can these shared features be formalized into a stage-based or dimensional model (RAM) with clear, distinguishable criteria for “instinctual” versus “reflective” action?
3. What are the limits of this comparative synthesis — where do the traditions diverge in ways that resist integration?

3. Methodology

This is a **theoretical and comparative-philosophical paper**; it does not involve human subjects, empirical data collection, or statistical analysis.

Design. The paper uses a two-part comparative method: (1) conceptual synthesis of foundational and contemporary work in evolutionary biology, cognitive neuroscience, and psychology addressing impulse control, dual-process cognition, and self-regulation, selected for direct relevance to the instinct/reflection distinction and prioritized by foundational status, direct relevance, and citation currency; and (2) close comparative textual analysis of the relevant portions of the Bhagavad Gītā (primarily Chapters 2–3, and selectively 14, 17–18), read alongside classical commentary (Śaṅkarācārya, 1984) and modern scholarly interpretation (Sarvepalli Radhakrishnan, 1948; Bimal Krishna Matilal, 2002; Jitendra Nath Mohanty, 2000).

Comparative framework. The paper uses a **structural-parallel method**: it identifies functionally analogous *stages* (trigger, appraisal, pause, reframing, chosen action — and their inverse, in the collapse pathway) independently described in both literatures, and examines where those stages converge, diverge, or illuminate one another, rather than treating either tradition as illustrative decoration for the other.

Scope and exclusions. The paper does not attempt an exhaustive review of Vedāntic soteriology or a full survey of consciousness studies. Empirical validation of RAM is identified as future work (Section 9) rather than undertaken here.

Limitations of method are addressed in full in Section 8, but are flagged here for transparency: comparative work of this kind carries an inherent risk of over-harmonizing traditions with different aims (liberation/*mokṣa* vs. behavioral prediction); Section 6 states explicitly where this paper declines to harmonize (notably, Śāṅkara's full non-dual conclusion regarding agency).

4. Literature Review

4.1 Evolutionary Biology

Human beings share a biological baseline with the rest of the animal kingdom, and evolutionary biology's central contribution here is explaining how cooperative, rule-governed behavior can emerge from selection pressures that are themselves indifferent to morality.

Robert M. Sapolsky's (2017) synthesis of neurobiology, endocrinology, and primatology traces behavior backward from the immediate neural event through hormonal state, developmental history, and evolutionary legacy, undercutting any simple instinct/culture dichotomy and supporting RAM's premise that "instinct" is a stack of biological influences operating on different timescales rather than a single mechanism. Michael Tomasello's (2016) account of the evolutionary origins of morality argues that human moral psychology evolved from the mutual obligations created by joint, cooperative activity, suggesting that self-governance is continuous with, not a late override of, older cooperative instincts. Frans de Waal's (2013) work on primate reconciliation and fairness-sensitivity reinforces this continuity thesis, arguing that moral behavior emerges "bottom-up" from social instinct rather than being imposed "top-down" by doctrine. Edward O. Wilson's (1975) foundational sociobiological framework remains relevant chiefly as the originating point for treating social and ethical behavior as a legitimate object of evolutionary explanation, a move Tomasello and de Waal both extend and refine.

Together, this literature supports one half of RAM's premise: biologically conditioned impulses are the evolved substrate from which reflective capacities are built, not a lower category separable from social cognition.

4.2 Neuroscience

Where evolutionary biology explains *why* reflective capacity might exist, neuroscience offers the closest available account of *how* it is implemented moment to moment.

Antonio R. Damasio's (1994) somatic marker hypothesis, drawn from case studies of prefrontal-damaged patients, demonstrated that emotion and bodily feeling are necessary inputs to rational decision-making rather than obstacles to it — foreclosing a naive version of "reflective agency" as the suppression of feeling by reason, and requiring instead that reflection be understood as a process that incorporates affective signals. Stanislas Dehaene's (2014) Global Neuronal Workspace model, developed with Jean-Pierre Changeux (Stanislas Dehaene & Jean-Pierre Changeux, 2011), offers a mechanistic account of the transition RAM calls "the pause": non-conscious information can, under conditions of attention and limited capacity, "ignite" a large-

scale prefronto-parietal network, becoming globally available for reasoning and voluntary control. This may provide a useful neuroscientific analogue to the phenomenological distinction between instinctual reaction and reflective choice, though the mapping between a neural ignition threshold and a first-person “pause” remains an interpretive bridge rather than a demonstrated identity. Giulio Tononi’s (2004) Integrated Information Theory approaches the same problem differently, proposing that consciousness corresponds to a system’s capacity for integrated information (Φ). RAM does not adjudicate between these competing theories of consciousness’s neural correlates; it takes their shared premise — that consciousness admits of empirically tractable degrees and transitions — as support for treating reflective agency as a graded, measurable capacity. Michael S. Gazzaniga’s (2011) argument that responsibility and agency emerge at the level of social interaction, not brain mechanics alone, supports RAM’s claim that reflective agency is completed in action rather than remaining a purely internal event.

4.3 Cognitive Psychology

Cognitive and social psychology offer the most direct empirical challenge to RAM, and this section engages that challenge rather than setting it aside.

Daniel Kahneman’s (2011) dual-process account distinguishes System 1 (fast, automatic) from System 2 (slow, effortful), showing that System 1 governs most judgment and choice, with System 2 typically endorsing rather than independently generating its output. Jonathan Haidt’s (2012) social intuitionist model sharpens this: moral judgment arises primarily from automatic intuition, with reasoning functioning largely as post-hoc justification — his rider-and-elephant metaphor casts reasoning as a rider who believes himself in control while intuition determines the direction of travel. RAM’s response, developed fully in Section 6, is that reflective governance is not claimed to be the *typical* case — Haidt’s finding is not in dispute — but rather an *achievable, trainable capacity*, empirically rare in an untrained population. Walter Mischel’s (2014) delay-of-gratification research supports the achievability claim directly: children’s capacity to delay gratification improved substantially through specific attentional and reframing strategies, and this capacity predicted long-term outcomes, converging with the Gītā’s own treatment of *guṇa* balance as cultivable (14.11).

Baumeister’s influential ego-depletion model proposed that self-control draws upon a limited resource, such that successive acts of self-regulation become harder to sustain (Roy F. Baumeister et al., 1998). Although subsequent large-scale replication studies have produced mixed findings regarding the strength and generality of the ego-depletion effect, resource-based accounts of self-regulation continue to influence theoretical discussions of self-control (Roy F. Baumeister & John Tierney, 2011). RAM draws on this literature only as one possible analogue to the Gītā’s collapse pathway (2.62–63; Section 5) — a discriminative faculty narrowing under repeated demand — rather than as settled empirical fact, and notes a genuine difference in proposed mechanism: Baumeister’s account is resource-based, while the Gītā’s is better read as attentional/appraisal-based. These are treated as complementary hypotheses, not established, competing explanations of the same phenomenon.

Viktor E. Frankl’s (1946/2006) account of meaning-making under extreme duress

adds a different kind of evidence: even under near-total loss of external control, a space for chosen response remained available to at least some individuals — his account of the freedom to choose one's attitude in any given circumstance, which he treats as the one liberty that cannot be taken away, echoes the Gītā's claim that discernment can, in principle, govern reaction even under severe conditions. Taken together, Mischel, Baumeister, and Frankl do not refute Haidt and Kahneman; they specify conditions under which the pause can nonetheless be cultivated — the empirical space RAM occupies.

4.4 Indian Philosophy

The scientific and Western-philosophical literature above supplies RAM's mechanism and empirical constraints; the Indian philosophical literature supplies its textual architecture and its claim to rigorous, methodologically defensible scholarship rather than devotional restatement. Bimal Krishna Matilal (2002) and Jitendra Nath Mohanty (2000) both treat classical Indian philosophy as a rational, argumentative tradition answerable to the same standards as analytic philosophy — organizing concepts like *buddhi* and *citta-vṛtti* as candidates for the same conceptual analysis applied to comparable Western constructs, rather than reporting them as doctrine. RAM follows this precedent throughout.

Sarvepalli Radhakrishnan's (1948) translation and commentary is the paper's primary textual source, chosen because Radhakrishnan reads the Gītā as a synthesis of Sāṃkhya, Yoga, and Vedānta — matching this paper's own use of its Sāṃkhya-Yoga psychological layer while treating strict non-dual metaphysics more cautiously (Section 6) — and because he situates the Gītā in explicit dialogue with Western philosophy, methodologically consonant with this paper's cross-traditional project. Śaṅkarācārya's commentary (Śaṅkarācārya, 1984) provides the interpretive authority behind this paper's reading of 3.42 — that *buddhi* is capable of directing the lower faculties — while this paper deliberately does not adopt his full Advaitic conclusion that the Self never truly acts, for the reasons given in Section 6: a strong non-dual metaphysic, taken as RAM's operative psychology rather than its horizon, would undercut the paper's claim of agency. This is a stated divergence, not an oversight.

The Yogasūtras, in Edwin F. Bryant's (2009) edition, supply RAM's most technical formulation of the pause: *yogaś citta-vṛtti-nirodhaḥ* (1.2), the restraint of the mind's fluctuations, treated as a specific, describable operation rather than a figure of speech. The Kaṭha Upaniṣad (Sarvepalli Radhakrishnan, 1953) supplies the oldest textual precedent for the governance hierarchy formalized at Bhagavad Gītā 3.42 — its chariot analogy is widely read, including by Radhakrishnan, as the likely source for the Gītā's own hierarchy — establishing that RAM compares modern science to a motif with independent, multi-textual attestation across the Vedāntic corpus, not an isolated convenient verse.

5. The Reflective Agency Model (RAM): Positive and Collapse Pathways

Definition. Reflective Agency is the progressive expansion of conscious self-governance over biologically conditioned impulses through discernment, deliberation, and intentional action.

RAM formalizes the instinct-to-agency transition surveyed across Sections 4.1–4.4 as a single describable process rather than a metaphor borrowed loosely from either science or scripture. It rests on three claims defended across this paper: (1) biologically conditioned impulses are not a lower, separable category of behavior but the evolved and neurally implemented substrate from which reflective capacity itself is built (Section 4.1–4.2); (2) the “pause” between stimulus and response — variously described as System 2 engagement (Daniel Kahneman, 2011), global workspace ignition (Stanislas Dehaene & Jean-Pierre Changeux, 2011), or *citta-vṛtti-nirodha* (Yoga Sūtra 1.2) — may be understood as a graded and potentially empirically tractable capacity, rather than an illusion generated after the fact (Section 4.3); and (3) this capacity is trainable and individually variable rather than fixed (Section 4.3–4.4). RAM’s claim is deliberately narrower than “reason governs instinct” as a general description of human cognition — a claim the psychological literature does not support (Section 4.3) — and is instead that reflective governance is an *achievable* state whose conditions of success and failure can be specified.

Figure 1 presents both pathways of the model; the collapse pathway is analyzed later in this section.

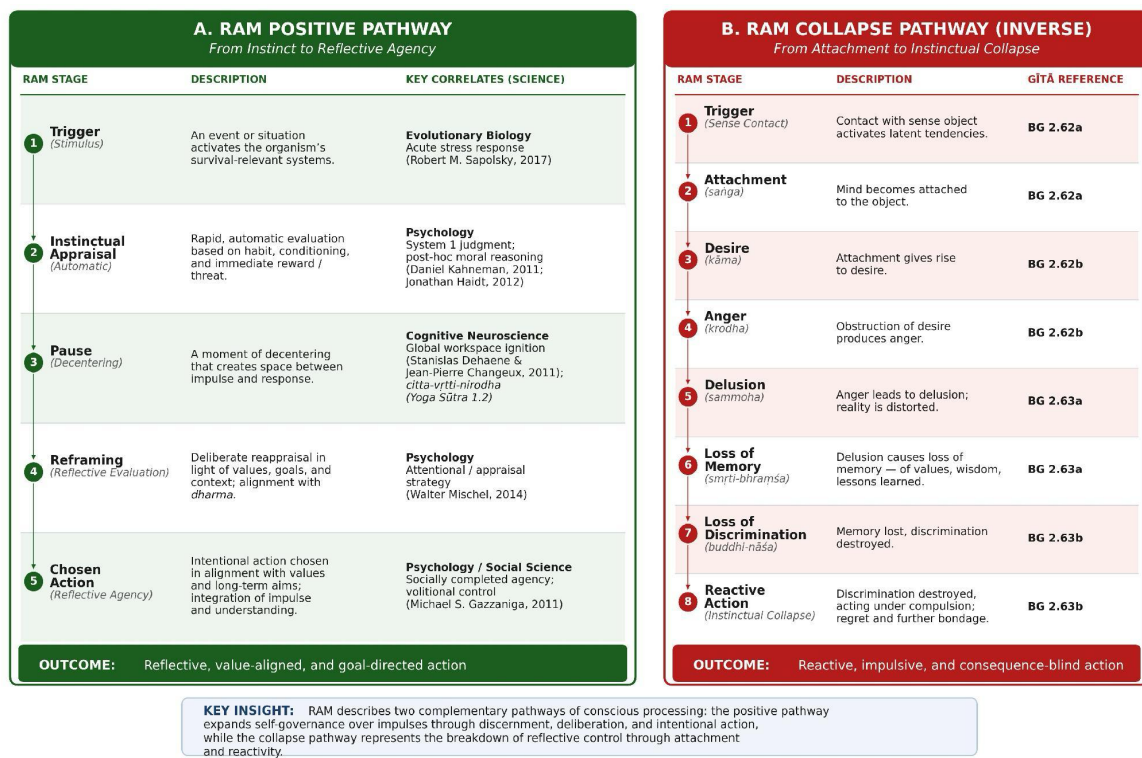


Figure 1. Reflective Agency Model (RAM): Two Pathways of Conscious Processing. Panel A presents the RAM Positive Pathway, illustrating the progressive expansion of reflective agency from stimulus through instinctual appraisal, pause, reframing, and intentional action. Panel B presents the RAM Collapse Pathway (Inverse), illustrating the progressive breakdown of reflective agency—from attachment through desire, anger, delusion, loss of memory, and loss of discrimination to reactive

action—as described in Bhagavad Gītā 2.62–63. Together, the two pathways depict complementary trajectories of conscious processing within the Reflective Agency Model.

Guṇa proportion as a source of individual variation in reflective capacity. RAM, as defined above, describes reflective agency as a progressive *expansion* of self-governance — which implies that the starting capacity for that expansion is not uniform across individuals. The Gītā’s account of the three *guṇas* offers a principled explanation for this variation rather than treating it as unexplained noise. Chapter 14 characterizes *sattva* as associated with clarity and illumination, *rajas* with restless activity and craving, and *tamas* with inertia and delusion (Bhagavad Gītā 14.5–14.13), while Chapters 17 and 18 extend this tripartite analysis to faith, knowledge, action, the agent, and even happiness, treating *guṇa* proportion as a pervasive, individually variable feature of personality and conduct (Bhagavad Gītā 17.1–17.22; 18.19–18.39). Mapped onto RAM, a *tamas*-predominant state may impede a person from noticing an impulse at all; a *rajas*-predominant state may permit noticing but drive premature action before discernment (*buddhi*) can complete its work; a *sattva*-predominant state most readily supports the pause and the deliberate choice that follows it. The Gītā explicitly treats the balance among the *guṇas* as cultivable (14.11, 14.22–14.26), giving RAM a textually grounded, non-deterministic account of why reflective capacity varies between individuals and within the same individual over time.

The collapse pathway: a failure-mode account of reflective agency. RAM’s account of the pause describes how reflective agency is achieved; the Gītā also offers a detailed account of how it is lost, strengthening the model with an equally specific account of its breakdown. Bhagavad Gītā 2.62–63 traces a five-step cascade: dwelling on a sense-object produces attachment; attachment produces desire; frustrated desire produces anger; anger produces delusion (*sammoha*); delusion produces loss of memory; and loss of memory produces the destruction of *buddhi* itself. Read against RAM, this cascade describes the progressive narrowing of the space between stimulus and response until it closes entirely, with action reverting to the most reactive, instinct-driven mode RAM describes. This pairing — Gītā 3.42’s governance hierarchy and 2.62–63’s account of its collapse — has an applied precedent in the counseling framework described in K. V. Sarma (2016), where interrupting the cascade at an early stage (attachment or desire) was treated as more tractable than restoring *buddhi* after its “destruction.” RAM’s contribution is to reformulate this two-part structure as a general model open to comparison with contemporary accounts of self-regulation failure, including resource-based theories of self-control such as the ego-depletion literature (Roy F. Baumeister & John Tierney, 2011; Section 4.3), rather than as a therapeutic technique alone.

RAM is proposed as a conceptual framework intended to generate empirically testable hypotheses rather than as a completed theory of consciousness. Its value lies in organizing independently developed scientific and philosophical observations into a common structural model open to future refinement or falsification.

6. Comparative Analysis with the Bhagavad Gītā

6.1 The Arjuna Narrative as a Structural Case of RAM

Section 5 formalized RAM as a five-stage process — trigger, appraisal, pause, reframing, and chosen action — abstracted from the scientific literature reviewed in Section 4. This section argues that the Bhagavad Gītā’s account of Arjuna does not merely illustrate this process loosely; it closely parallels it, stage for stage, with textual markers at each transition. The mapping is set out in full, citation-ready detail in Table 1 (Section 6.3).

Trigger and instinctual appraisal (BG 1.28–1.47). Arjuna’s crisis begins with a specific perceptual trigger — seeing his own teachers and kinsmen arrayed for battle — followed by described physiological symptoms of distress and a collapse of resolve. Read through Daniel Kahneman’s (2011) framework, this is System 1 operating under acute threat; read through Antonio R. Damasio’s (1994) somatic marker framework, it is the bodily, affect-laden signal RAM treats as necessary input rather than noise. Arjuna immediately generates a moral argument to justify the instinctual retreat, consistent with Jonathan Haidt’s (2012) claim that reasoning frequently follows and rationalizes intuition — his Chapter 1 speech is best read, on RAM’s account, as post-hoc moral argument in service of instinctual withdrawal, not yet reflective agency.

The pause (BG 2.7). Arjuna states that he does not know what is right, that his judgment is compromised, and asks Krishna directly to instruct him, naming himself a student who has taken refuge. This is the moment RAM calls the pause — the interruption of automatic translation from impulse to action, and the explicit acknowledgment that the current appraisal cannot be trusted. Arjuna does not resolve his own crisis through willpower alone; he opens himself to reframing from outside his own immediate appraisal, consistent with Walter Mischel’s (2014) finding that successful delay of gratification typically depends on a specific attentional or reframing strategy, not raw self-restraint.

Reframing (BG Chapters 2–17, condensed). Krishna’s teaching performs the reframing stage: relocating Arjuna’s understanding of death and duty within a wider frame, rather than issuing a direct behavioral command. This is consistent with Michael S. Gazzaniga’s (2011) view that reflective agency is completed by the agent’s own uptake of a wider frame — Krishna provides *why*, in RAM’s terms, rather than *what*.

Chosen action (BG 18.63, 18.73). At 18.63, Krishna explicitly returns the decision to Arjuna, instructing him to deliberate fully and then act as he wishes. Arjuna’s reply at 18.73 is the structurally decisive verse for this paper’s argument: he states that his delusion (*moha*) is destroyed and his memory (*smṛti*) regained, and that he will act accordingly. This is the near-exact terminological inverse of the collapse cascade at 2.62–63 (Section 5), in which delusion produces loss of memory, which produces the destruction of discrimination. Arjuna’s resolution reverses that sequence point for point, giving RAM’s two-part structure direct textual confirmation: the Gītā appears to treat recovery from instinctual collapse as the systematic reversal of the same terms in which it describes the collapse itself.

6.2 Where RAM Stops Short of Śaṅkara

This structural mapping is deliberately partial. Śaṅkara's commentary (Śaṅkarācārya, 1984) reads the Gītā's ultimate teaching as Advaitic: the Self (Ātman) is the unchanging witness (*sākṣin*) of all mental and physical activity, including *buddhi* itself. On a strict Advaitic reading, Arjuna's "choice" at 18.63 and his transformation at 18.73 occur entirely within the domain of *prakṛti*; the Self, properly understood, never acts.

RAM does not adopt this conclusion. A theory whose central claim concerns *reflective agency* would be self-undermining if its own philosophical foundation held that agency does not, strictly speaking, exist. RAM instead treats *buddhi*-governance as its operative psychological claim, while treating Śaṅkara's witness-consciousness (*sākṣin*) as a *horizon concept* — the metaphysical condition that makes a pause possible at all — without claiming its comparative project can adjudicate whether that witness is, at the deepest level, an agent or a non-acting substrate. This divergence is stated explicitly rather than smoothed over, per the paper's methodological commitment (Section 3) to mark points where the traditions resist full harmonization.

6.3 Summary of the Comparative Mapping

Table 1

Structural Mapping of RAM Stages to Scientific Correlates and Bhagavad Gītā Textual Markers

RAM Stage	Scientific Correlate (Section 4)	Gītā Textual Marker
Trigger	Acute stress response (Robert M. Sapolsky, 2017)	BG 1.28–29
Instinctual appraisal	System 1 judgment; post-hoc moral reasoning (Daniel Kahneman, 2011; Jonathan Haidt, 2012)	BG 1.31–1.47
Pause	Global workspace ignition (Stanislas Dehaene & Jean-Pierre Changeux, 2011); <i>citta-vṛtti-nirodha</i> (Yoga Sūtra 1.2)	BG 2.7
Reframing	Attentional/appraisal strategy (Walter Mischel, 2014)	BG 2–17
Chosen action	Socially completed agency (Michael S. Gazzaniga, 2011)	BG 18.63, 18.73
Collapse (inverse case)	Resource-based self-control / ego depletion (Roy F. Baumeister & John Tierney, 2011)	BG 2.62–63

This mapping is the paper's central comparative claim: not that science and the Gītā say the same thing in different vocabularies, but that they show substantial structural convergence — trigger, appraisal, pause, reframing, action, and its inverse, collapse — while retaining genuine, stated differences in what that structure ultimately means.

7. Discussion

7.1 Answering the Research Questions

Primary question — can the instinct-to-agency transition be integrated into a single cross-traditional model? The comparative mapping in Section 6.3 suggests yes, with an important qualification: integration succeeds at the level of *structure* but not at the level of *ultimate metaphysical interpretation*, where Section 6.2 identified a genuine, stated limit. RAM is therefore a structural, not a metaphysical, integration.

Subsidiary question 1 — shared structural features. The first is between global workspace ‘ignition’ (Stanislas Dehaene & Jean-Pierre Changeux, 2011) and *citta-vṛtti-nirodha*/BG 2.7: both may be interpreted as describing a threshold between automatic and deliberately available processing. The second is between resource-based self-control theory and the Gītā’s collapse cascade (2.62–63) — independently developed accounts of self-regulatory capacity narrowing under sustained demand (Section 4.3).

Subsidiary question 2 — can this be formalized with distinguishing criteria? RAM offers three: (a) presence or absence of a measurable pause between stimulus and response; (b) whether action is oriented toward short-term relief/dominance or a self-endorsed, longer-horizon value; and (c) whether *guṇa*-linked variability (Section 5) is a stable trait or a state shifting within the same individual across contexts — RAM predicts the latter.

Subsidiary question 3 — where do the traditions resist integration? Most sharply at whether “agency” is, at the deepest level, real (Section 6.2). RAM should be read as a working model adequate to the empirical and practical questions it addresses, not a resolution of the metaphysical question of free will.

7.2 What RAM Adds Beyond Existing Frameworks

First, the collapse pathway is underused in dual-process literature as an organizing structure: Daniel Kahneman (2011) and Jonathan Haidt (2012) describe System 1’s dominance as a baseline condition rather than a *process of collapse* from a prior, more governed state, while the Gītā’s cascade — and its terminological reversal at 18.73 — offers a more specific, stage-based account of how self-regulatory capacity narrows.

Second, the *guṇa* framework gives RAM a principled account of individual variation largely absent from dual-process literature, which tends to treat System 1/System 2 balance as a general cognitive property rather than a substantially person-specific one — addressing a standard objection to dual-process models: that they underspecify why some individuals reliably exercise more reflective control than others under similar conditions.

Third, RAM formalizes a framework already used, prior to this paper, in a counseling context (K. V. Sarma, 2016; Section 1) — though, as stated throughout, that practice-based origin is motivating context, not validating evidence.

7.3 Implications for Consciousness Studies

RAM does not adjudicate between competing neuroscientific theories of consciousness (Section 4.2). Its more modest claim is that both are compatible with,

and arguably illuminated by, a cross-traditional account of what a transition into deliberate availability looks like from the inside — which the Gītā's Arjuna narrative supplies in first-person, dramatized form. This suggests treating well-developed contemplative and narrative traditions as sources of phenomenologically detailed hypotheses rather than mere illustration of scientific findings (Section 9).

7.4 Implications for Practice

If RAM's structural claims hold, the practical implication — consistent with, though not proven by, K. V. Sarma (2016) — is that interventions aimed at strengthening reflective agency should target the earliest stages of the collapse cascade rather than attempting to restore discrimination after anger and delusion have compounded, and should be calibrated to an individual's predominant *guṇa* state rather than a uniform protocol.

8. Limitations

Comparative over-harmonization. As flagged in Section 3, comparative work spanning traditions with different ultimate aims risks producing an artificially tidy synthesis. This paper has tried to guard against that risk by stating divergences explicitly, most importantly in Section 6.2, where RAM's operative claim of agency is distinguished from Śāṅkara's full Advaitic conclusion. Readers should treat RAM's structural mapping (Section 6.3) as a mapping of process architecture, not a claim that the traditions surveyed agree about the deeper nature of consciousness or the self.

Practitioner-background material is not clinical evidence. The counseling framework described in K. V. Sarma (2016) motivated this paper's theoretical development but is not presented as data. The work was individualized and non-standardized, consistent with the Gītā's account of person-specific *guṇa* constitution; no outcome measures, controls, or systematic case documentation exist for it in a form suitable for evaluation. RAM's *guṇa*-variation claim and the practical implications drawn in Section 7.4 should be read as hypotheses this background suggested, not findings it demonstrates.

Absence of empirical testing of RAM itself. RAM's three central claims — that the pause is graded and measurable, that *guṇa*-linked variation predicts individual differences in reflective capacity, and that the collapse cascade parallels resource-based self-regulation failure — are testable hypotheses rather than demonstrated results. Nothing in this paper constitutes empirical support for RAM beyond the convergence of independently developed literatures.

Single-translation and single-commentary reliance. The Gītā material draws primarily on Sarvepalli Radhakrishnan's (1948) translation and Śāṅkara's commentary. Other major commentarial traditions — notably Rāmānuja's qualified non-dualism and Madhva's dualism — read the same verses differently, particularly regarding the relationship between the individual self and ultimate reality, and might yield a different account of where RAM's mapping does or does not hold. This paper's conclusions are specific to the Advaita-inflected, Sāṃkhya-Yoga-psychological reading it has adopted.

Author positionality. The author's own decades of counseling practice, grounded

in this same textual material, is both the origin of this paper's central insight and a potential source of confirmation bias. Section 3's structural-parallel method and the explicit divergence stated in Section 6.2 are offered as partial safeguards, but readers should weigh this positionality when evaluating the paper's comparative claims.

9. Future Research

Systematic, ethically governed follow-up to the practitioner observations. The counseling framework described in K. V. Sarma (2016) suggests testable hypotheses — for instance, that intervention at the attachment/desire stage of the collapse cascade produces better outcomes than intervention after anger and delusion have compounded (Section 5) — that could be investigated through a prospective, consent-governed case series or small controlled study, with standard outcome measures rather than retrospective recollection.

Empirical operationalization of the *guṇa*-variation claim. Section 5's proposal that *sattva/rajas/tamas* proportion predicts individual variation in reflective capacity is, at present, textually motivated rather than measured. A natural next step would be to examine how validated personality instruments — for example, the Five-Factor Model (Paul T. Costa Jr. & Robert R. McCrae, 1992) — correlate with self-report or behavioral measures of delay-of-gratification and self-regulatory capacity (Walter Mischel, 2014), without requiring acceptance of the *guṇa* framework's metaphysical claims, only testing its behavioral predictions.

Testing the collapse-pathway parallel. RAM's claim that the *Gītā*'s collapse cascade (2.62–63) and resource-based theories of self-control, including the ego-depletion literature, describe convergent phenomena via different mechanisms (Section 4.3) is, in principle, testable using attentional-reframing manipulations modeled on Walter Mischel's (2014) strategies, potentially with neuroimaging designed to detect the global-workspace “ignition” threshold (Stanislas Dehaene & Jean-Pierre Changeux, 2011). Given ongoing replication debates around ego depletion specifically, such a design would need to treat the resource-based mechanism as one hypothesis among several rather than a settled baseline.

Extension to other contemplative traditions. Future work could test whether RAM's structural mapping holds against other traditions with detailed introspective psychologies — for instance, Buddhist Abhidharma accounts of mental factors, or Stoic accounts of the interval between impression (*phantasia*) and assent — clarifying whether RAM's structure reflects something general about human self-regulation or something more specific to the Vedāntic tradition it was developed from.

10. Conclusion

This paper set out to ask whether the human capacity to move from instinctual reaction to deliberate, value-aligned action could be described in a single framework, integrating evolutionary, neuroscientific, and psychological accounts of self-regulation with the Bhagavad *Gītā*'s account of Arjuna's transformation. The proposed answer, the **Reflective Agency Model**, does not claim that these traditions were ever saying the same thing in different languages. It claims something more specific: that they show substantial structural convergence — a measurable threshold between automatic and deliberate processing, a trainable rather than fixed capacity for pausing before action,

a principled account of individual variation, and a matching, specifiable account of how that capacity collapses under sustained demand — while continuing to disagree, in ways this paper has tried to state rather than obscure, about what that structure ultimately means for the nature of the self.

The clearest piece of evidence for treating this convergence as more than coincidental is textual: the Gītā's account of Arjuna's collapse (2.62–63) and his eventual resolution (18.73) use near-mirror-image vocabulary — delusion and lost memory in one direction, dispelled delusion and regained memory in the other — describing breakdown and recovery as structurally symmetric processes, independently of any modern psychological framework brought to bear on the text.

Where this leaves the pause itself is close to where Arjuna is left at Bhagavad Gītā 18.63: not with an instruction, but with the deliberate return of the choice. RAM does not claim to make that choice easier. It claims, more modestly, to have described what is actually happening in the moment a person makes it.

The principal contribution of this paper is not to demonstrate that ancient philosophy anticipated modern neuroscience, but to propose the Reflective Agency Model (RAM) as a common conceptual framework through which independently developed scientific and philosophical accounts of reflective action may be systematically compared.

References

- Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, 74(5), 1252–1265. <https://doi.org/10.1037/0022-3514.74.5.1252>
- Baumeister, R. F., & Tierney, J. (2011). *Willpower: Rediscovering the greatest human strength*. Penguin Press.
- Bryant, E. F. (Trans.). (2009). *The Yoga Sūtras of Patañjali: A new edition, translation, and commentary*. North Point Press.
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources.
- Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. Putnam.
- de Waal, F. (2013). *The bonobo and the atheist: In search of humanism among the primates*. W. W. Norton & Company.
- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking.
- Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227. <https://doi.org/10.1016/j.neuron.2011.03.018>
- Frankl, V. E. (2006). *Man's search for meaning* (I. Lasch, Trans.). Beacon Press. (Original work published 1946)
- Gazzaniga, M. S. (2011). *Who's in charge? Free will and the science of the brain*. Ecco.
- Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*.

Pantheon Books.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Matilal, B. K. (2002). *Ethics and epics: The collected essays of Matilal* (Vol. 2, J. Ganeri, Ed.). Oxford University Press.

Mischel, W. (2014). *The marshmallow test: Mastering self-control*. Little, Brown and Company.

Mohanty, J. N. (2000). *Classical Indian philosophy: An introductory text*. Rowman & Littlefield.

Radhakrishnan, S. (Trans.). (1948). *The Bhagavadgītā: With an introductory essay, Sanskrit text, English translation and notes*. George Allen & Unwin.

Radhakrishnan, S. (Trans.). (1953). *The principal Upaniṣads*. George Allen & Unwin.

Sapolsky, R. M. (2017). *Behave: The biology of humans at our best and worst*. Penguin Press.

Sarma, K. V. (2016). *Dr. Karra's notes on quality of life by design*. [Self-published].

Śāṅkarācārya. (1984). *Bhagavad Gītā: With the commentary of Śāṅkarācārya* (S. Gambhirananda, Trans.). Advaita Ashrama.

Tomasello, M. (2016). *A natural history of human morality*. Harvard University Press.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, Article 42. <https://doi.org/10.1186/1471-2202-5-42>

Wilson, E. O. (1975). *Sociobiology: The new synthesis*. Belknap Press of Harvard University Press.