Paris B. Obdan

Layers of Agency: A Three-Level Architecture of Human Action

Abstract

Contemporary theories of action implicitly assume that agency is governed from within reflection—by deliberation, endorsement, or planning. This paper argues that this assumption is structurally incomplete. It introduces a three-layer architecture of agency comprising reflective governance, arational–procedural processes, and an intermediate layer of identity-level motivation termed Subconscious Practical Identity (SPI). SPI consists of stable, affectively encoded motivational structures that organize action teleologically across time while remaining largely inaccessible to reflection. Recognizing this layer explains how agents can sustain coherent life trajectories while systematically misidentifying what organizes them, and why some failures of agency arise from architectural misalignment rather than deliberative breakdown. The model clarifies classical theories' scope conditions, distinguishes multiple failure modes of agency, and shows how integration can occur without reflective sovereignty.

1. Introduction

Contemporary philosophy of action is unified less by shared conclusions than by a structural assumption about where agency resides. Across otherwise divergent frameworks, the central engines of agency are located within the reflective point of view. Davidson explains intentional action by appeal to primary reasons the agent can avow. Mele analyzes failures of agency as conflicts within deliberation. Smith identifies an agent's real reasons with those they would endorse under ideal reflection. Bratman grounds diachronic agency in reflectively accessible plans and future-directed intentions. Despite their differences, these theories converge on a common picture: agency is governed from within reflection, endorsement, and planning (Davidson 1971; Mele 1987; Smith 1994; Bratman 1987).

This picture is powerful, but incomplete. Ordinary human lives exhibit a striking phenomenon that reflection-centered models struggle to explain: long-term, coherent, purposive trajectories that are not organized by reflectively accessible intentions, values, or plans. Agents often pursue stable careers, relationships, and styles of life over decades while sincerely misidentifying what is doing the organizing. Their self-explanations are intelligible and coherent, yet systematically incomplete. The problem is not simply irrationality, akrasia, or self-deception in the traditional sense. Rather, the organizing forces lie outside the deliberative economy itself. Some failures of agency arise not from distorted reflection, but from breakdowns in the control architecture that couples intention, identity, and action.

This paper argues that the difficulty arises from a missing layer in the architecture of agency. Standard theories capture important aspects of reflective governance and of arational, procedural behavior. What they lack is a principled account of identity-level motivational organization that operates beneath reflection while still exhibiting teleological structure. Many agents display patterns of striving, exhaustion, self-sabotage, or effortless coherence that are too stable and purposive to be reduced to habit or impulse, yet too inaccessible to be treated as ordinary plans. To account for these patterns, we must take seriously a domain of non-reflective but systematic practical organization.

The central claim of the paper is that between reflective governance and arational processes there is a structurally distinct motivational layer. I call this layer *Subconscious Practical Identity (SPI)*. SPI consists of stable, identity-like motivational structures that organize action across long temporal horizons while remaining largely inaccessible to reflection. These structures are neither episodic impulses nor explicit intentions. They are enduring patterns of practical orientation—often formed through early attachment, reinforcement histories, and culturally mediated norms—that shape what an agent does, sustains, and repeatedly reconstructs over time, without being chosen or endorsed as such within deliberation.

Introducing SPI fills the gap between two well-understood regions of agency. At one end lies Layer 1: reflective governance, comprising explicit intentions, articulated reasons, deliberation, and narrative self-understanding. At the other lies Layer 3: arational–procedural processes, including automatic, reactive, and biologically mediated behaviors that fall outside the space of reasons. SPI occupies Layer 2, a middle domain of identity-level motivation without reflective authorship. It is structured, teleological, and cross-temporally stable, but it does not present itself to reflection in propositional or deliberative form. Much of its influence is therefore visible only indirectly, through systematic patterns in what agents find salient, exhausting, sustainable, or strangely easy.

Once SPI is in view, a range of otherwise puzzling phenomena come into focus. It becomes possible to explain why reflective self-explanations often function as sincere rationalizations; why agents with similar outward behavior differ radically in experienced effort and flexibility; why some forms of self-sabotage are coherent and high-functioning rather than chaotic; and why certain failures of agency arise not from weakness of will, but from misalignment between motivational layers. Crucially, SPI is not a pathological addition to agency. Integrated forms of SPI scaffold reflective governance and make long-term coherence possible. It is compensatory or misintegrated SPI that produces identity drag, architectural tension, and more severe forms of agency breakdown.

The ambition of the paper is architectural rather than metaphysical. It does not attempt to locate personal identity or to resolve questions about the metaphysical subject of agency. Instead, it maps functional divisions within the cognitive control system that underwrites human action. Classical theories illuminate important aspects of this system, but they presuppose that agency either appears within reflection or collapses into arationality. The layered model rejects this dichotomy. Agency can be purposive, coherent, and life-organizing without being reflectively authored.

The paper proceeds as follows. Section 2 develops the three-layer architecture of agency in detail, distinguishing reflective governance (Layer 1), arational–procedural processes (Layer 3), and Subconscious Practical Identity (Layer 2). A contrastive set of case families is introduced at the end of Section 2 to illustrate how SPI can organize long-term action in markedly different ways under otherwise similar external conditions. Section 3 introduces the notions of alignment, identity drag, and cognitive bandwidth, explaining how relations between layers shape the phenomenology and cost of agency. Section 4 situates classical theories of action within the layered architecture, showing both their explanatory power and their shared limitations, and engages situationist and deep-self approaches as partial but incomplete interlocutors. Section 5 examines Bratman's planning theory as the strongest reflective account of diachronic agency, arguing that SPI reproduces many of its functional roles while violating its psychological assumptions. Section 6 extends the model to socially situated agency, introducing epistemic capture, runaway identity-level dominance, and non-epistemic failures of action such as motor-gating breakdown in Parkinsonism, culminating in a provisional taxonomy of architectural failure modes. Section 7 draws on clinical psychology, particularly the work of Carl Rogers, to illuminate the dynamics of integration and incongruence within this layered framework. Section 8 concludes.

Trilogy Note: This paper is the first in a coordinated three-paper sequence on the architecture of human agency under conditions of limited reflective authority. The present paper develops the three-layer model of agency—reflective governance (Layer 1), Subconscious Practical Identity (Layer 2), and arational–procedural processes (Layer 3)—and uses it to diagnose the limits of reflection-centered theories of action. A companion paper, *Fragility of Reflection: Agency Without Supervisory Authority*, focuses on the phenomenology and mechanics of reflective failure, arguing that reflection does not occupy a supervisory role in action initiation or control (Obdan 2026a). A third paper, *Reintegrated Agency: Self-Governance Without Transparency*, develops a positive account of self-governance that is compatible with this non-sovereign conception of reflection (Obdan 2026b). Each paper is intended to stand alone, but they are designed to be read together as addressing distinct stages of a single explanatory project.

2. Architecture of Agency

2.1 Layer 1: Reflective Governance

Layer 1 comprises the domain of reflective governance. It includes explicit intentions, articulated reasons, deliberative choice, evaluative judgment, and narrative self-understanding. When agents explain what they are doing, justify their actions, revise plans, or assess whether their behavior aligns with their values, they are operating within this layer.

Most philosophical theories of action implicitly identify agency with Layer 1. Davidsonian primary reasons, Melean deliberation and conflict, Smithian endorsement under ideal reflection, and Bratmanian planning all treat reflectively accessible states as the primary engines of action. Within its proper scope, this focus is well motivated. Reflective governance enables agents to coordinate means and ends, respond to reasons, regulate impulses, and construct temporally extended projects.

However, reflective governance has a distinctive limitation: it is epistemically bounded. It can operate only on motivational material that is available to reflection. When sources of motivation lie outside reflective awareness, Layer 1 does not go silent. Instead, it interprets. It generates sincere narratives, values, and reasons that render behavior intelligible from within the reflective standpoint, even when those narratives fail to identify the deeper organizing forces of action.

This interpretive function is not a defect. Reflective governance is not designed to excavate an agent's full motivational architecture. Its role is regulatory rather than archeological: to stabilize action given what is reflectively available, not to uncover the full motivational architecture.. When motivation originates elsewhere in the system, reflection supplies the best available explanation rather than the true source.

This point is crucial for what follows. Classical theories often treat reflective access as a condition on agency itself. When agents misidentify their motivations, the explanation is typically framed in terms of error, irrationality, or self-deception. The layered model rejects this inference. Reflection can function exactly as designed while nonetheless mislocating motivational sources, because its epistemic horizon is limited by architecture rather than by pathology.

Once reflective governance is understood as one layer within a broader control system rather than the sovereign locus of agency, it becomes possible to explain how agents can act coherently, purposively, and intelligibly even when reflection fails to track what is guiding them. Reflective narratives may be sincere, stable, and normatively structured while still being downstream interpretations of non-reflective motivational organization. This is not a failure of reflection. It is a consequence of its place in the architecture.

2.2 Layer 3: Arational–Procedural Processes

At the opposite end of the agency architecture lies Layer 3: arational–procedural processes. This layer includes automatic, reactive, and biologically mediated behaviors such as flinching, freezing, startle responses, affective outbursts, habitual motor routines, and other forms of behavior that bypass deliberation and evaluative judgment. These actions are often meaningful and expressive, but they are not guided by reasons and do not belong to the deliberative economy.

Rosalind Hursthouse's analysis of arational action captures this domain precisely (Hursthouse 1991). Such actions are neither rational nor irrational; they fall outside the space of reasons altogether. They are triggered directly by perceptual, affective, or bodily mechanisms rather than by evaluative assessment or intention formation. An agent who recoils, cries, or freezes does not fail to act for reasons; they act without reasons in the relevant sense.

Layer 3 processes are characteristically episodic rather than diachronic. They respond to local stimuli and immediate contexts rather than organizing behavior across extended temporal horizons. While some procedural routines can be trained and stabilized—such as skilled motor sequences or conditioned responses—they do not by themselves generate life-level projects, commitments, or identity-defining trajectories. They execute rather than govern.

Distinguishing Layer 3 from higher layers is essential for avoiding two common confusions. The first is the tendency to treat all non-reflective behavior as irrational or deficient. Arational actions are not failures of agency; they are part of the normal control repertoire of embodied agents. The second confusion runs in the opposite direction: collapsing identity-level motivational organization into mere habit or impulse. As later sections will show, Subconscious Practical Identity (SPI) is neither episodic nor reactive in this way. It is structured, teleological, and cross-temporally stable in a manner Layer 3 processes are not.

Layer 3 therefore sets the lower boundary of the agency architecture. It supplies the reactive and procedural substrate on which higher-order control operates, but it does not itself organize lives. Recognizing this boundary prevents identity-level motivational structures from being misclassified as arational behavior and clarifies what must be added—rather than reduced—when introducing a middle layer of agency.

2.3 Layer 2: Subconscious Practical Identity

Between reflective governance and arational–procedural processes lies a motivational domain that has received relatively little direct attention in contemporary philosophy of action. I call this domain Subconscious Practical Identity (SPI). SPI consists of stable, identity-like motivational structures that organize action across long temporal horizons while remaining largely inaccessible to reflection. These structures are neither episodic impulses nor explicit plans. They are enduring patterns of practical orientation that shape what agents do and sustain over time without reflective authorship.

SPI is introduced to explain a familiar but theoretically underdescribed phenomenon: coherent, purposive life trajectories whose organizing motivations are not ordinarily available to reflective governance. Agents often pursue stable careers, relationship patterns, and styles of life over decades while sincerely misidentifying what is driving them. Their self-explanations are intelligible and coherent, yet systematically incomplete. The difficulty here is not best described as irrationality, akrasia, or self-deception in the traditional sense. Rather, the organizing forces lie outside the deliberative economy itself.

SPI is not a pathological add-on to agency. It is a normal feature of human motivational architecture. Much of adult life is scaffolded by tacit identity-level structures formed through early attachment, reinforcement histories, culturally mediated norms, and emotionally salient experiences. These structures operate beneath reflection but exert a persistent gravitational pull on practical reasoning, option salience, and long-term commitment.

For present purposes, five features characterize SPI.

First, SPI is identity-like. It encodes implicit self-conceptions, emotional needs, attachment orientations, and internalized standards that function as lived answers to the question "what kind of person am I?"—even when the agent cannot articulate those answers. These structures are not momentary states or local evaluations. They form part of the background of agency, shaping what feels natural, threatening, dignified, or worthwhile.

Second, SPI is cross-temporally stable. Unlike impulses or moods, SPI persists across years or decades. It does not merely bias isolated decisions; it organizes extended trajectories. Career choices, recurring relational patterns, characteristic ways of striving, and tolerances for risk or dependency often exhibit a coherence that exceeds anything present in the agent's reflective planning. SPI supplies that diachronic organization.

Third, SPI is emotionally structured. Its contents are not primarily inferential or propositional. They are encoded through affective learning, attachment dynamics, and emotionally charged reinforcement. As a result, SPI does not present itself to reflection in the form of explicit beliefs or intentions. It exerts pressure through felt salience, attraction, aversion, shame, comfort, or urgency rather than through deliberative endorsement.

Fourth, SPI is largely inaccessible to reflection. Agents typically cannot retrieve SPI directly through introspection. It may surface in therapy, crisis, or rare moments of insight, but most of the time it operates silently. Reflective governance therefore tends to misidentify SPI-driven action as the product of explicit values, ideals, or intentions. This misidentification is usually sincere rather than deceptive. Reflection is doing its best given what it can see.

Fifth, and most importantly, SPI is teleological. It generates goal-directed, trajectory-level organization. SPI structures do not merely push behavior reactively; they pull it toward certain forms of life. They stabilize pursuits, constrain deliberation, and make some options feel viable while others never seriously arise. In this respect, SPI behaves functionally like long-term intention, even though it is not reflectively accessible, endorsable, or directly revisable.

This combination of features places SPI in a structurally distinct motivational category. It is not part of Layer 1, because it does not depend on deliberation, endorsement, or narrative self-governance. It is not part of Layer 3, because it is not episodic, reactive, or stimulus-bound. SPI occupies a middle layer: identity-level motivation without reflective authorship.

Recognizing this middle layer allows us to distinguish failures of agency that arise from deliberative breakdown from those that arise from architectural misalignment. It also explains how agents can act coherently, purposively, and intelligibly while systematically mislocating the sources of their motivation. Reflective narratives can be sincere, stable, and normatively structured while nonetheless functioning as downstream interpretations of deeper motivational organization.

The next subsection introduces a set of contrastive case families that make these abstract features concrete. By holding external structure fixed while varying identity-level motivational organization, those cases will illustrate how SPI can scaffold agency, impose chronic drag, or generate self-stabilizing but destructive equilibria—without any appeal to irrationality or reflective failure.

2.4 Contrastive Case Families: Identity-Level Organization in Practice

The abstract features of Subconscious Practical Identity become clearer when we examine cases that hold external structure fixed while varying identity-level motivational organization. The following four cases are deliberately drawn from a single professional setting—medical practice—so that differences in agency cannot be attributed to role, competence, intelligence, or environmental demand. All four agents are highly trained physicians working in demanding institutional contexts. What differs is not what they do, but how their agency is organized and sustained.

These cases are not intended as diagnostic profiles or moral exemplars. They are architectural case families: stylized patterns that isolate the functional role of identity-level motivation in shaping agency across time.

2.4.1 The Insecure Doctor: Compensatory Identity Maintenance

The first physician exhibits a stable, outwardly successful professional trajectory. He completes medical training, performs competently under pressure, and maintains a respected position within his institution. His career is marked by diligence, persistence, and high standards. From the outside, his agency appears robust.

At the identity level, however, his motivational architecture is compensatory. He was raised by a harsh, status-oriented parent who strategically withheld approval and affection, offering recognition only when the child's achievements reflected well on the parent. Over time, the physician internalized a conditional self-valuation: worth is secured through status, performance, and external validation. This structure is not reflectively endorsed or even recognized as such. It functions as background identity.

Medicine becomes the obvious site for this compensation. It offers prestige, evaluative clarity, and socially sanctioned admiration. The physician experiences his career as chosen and meaningful, yet the work itself is not self-rewarding. Helping patients does not replenish motivation; it merely justifies continued striving. Success temporarily stabilizes the identity, but never resolves it.

As a result, agency here is high-functioning but effortful. Reflective governance must continually regulate anxiety, sensitivity to evaluation, and fear of failure. Setbacks impose disproportionate psychological costs. The physician is disciplined, but depleted. His life is organized coherently, yet sustained through chronic identity drag.

This is not a failure of will, planning, or reflection. It is a case of compensatory SPI generating a viable but costly equilibrium.

2.4.2 The Secure Doctor: Autocatalytic Identity Support

The second physician occupies a nearly identical external role. She works comparable hours, faces similar institutional pressures, and carries equivalent responsibility. Yet her experience of agency is markedly different.

Her identity-level motivational architecture is non-compensatory. She was raised in a stable environment with consistent approval and emotional attunement. As a result, her sense of worth is not contingent on performance or status. Medicine is not a proving ground; it is an extension of existing values and capacities.

Here, professional activity is autocatalytic. The work itself generates motivational return. Caring for patients, mastering complex cases, and collaborating with colleagues reinforce rather than drain identity-level motivation. Reflective governance is used sparingly—not to manage internal threats, but to coordinate logistics and respond to genuine novelty.

Setbacks are metabolized without destabilization. Fatigue occurs, but not existential depletion. The physician does not need to continually justify her trajectory to herself. Agency is sustained with relatively low regulatory cost because SPI and reflective governance are aligned.

This case illustrates that coherence does not require reflective authorship of motivation. It requires integration. Identity-level structure can scaffold agency quietly, without drama, friction, or heroics.

2.4.3 The Addict Doctor: Self-Stabilizing Destructive Equilibrium

The third physician is also highly competent and outwardly successful. She works long shifts in a demanding specialty and is widely regarded as capable and reliable. Unlike the previous cases, however, her identity-level motivational architecture has evolved into a self-stabilizing but destructive equilibrium.

At the center of this structure is a substance addiction that remains largely invisible to reflective governance. The addiction is not experienced as a discrete problem to be solved. Instead, it is embedded within a broader identity-level arrangement that continually generates justification for itself.

The physician unconsciously maintains chronic overwork, financial stress, and secondary compulsive behaviors in order to sustain a narrative of deserved relief. Exhaustion becomes evidence of virtue. Stress licenses chemical escape. Attempts to reduce workload, address financial instability, or question substance use are deflected or minimized—not through explicit denial, but through identity-level rationalization.

Crucially, this equilibrium is self-protective. Any intervention that threatens the structure is experienced as illegitimate or hostile, even when it risks harm to close relationships or dependents. Reflective governance is not absent; it has been co-opted. It functions to stabilize the equilibrium rather than interrogate it.

This is not mere impulsivity or lack of self-control. It is a case of runaway SPI, where identity-level motivation dominates both action and interpretation. Agency remains coherent and goal-directed, but no longer self-correcting.

2.4.4 The Vibes Doctor: Generative Alignment and Low-Friction Agency

The final physician also practices medicine competently, often in lower-mortality or relationally focused specialties such as pediatrics. His defining feature is not charisma or ambition, but a pervasive ease of engagement.

His identity-level motivational architecture is deeply aligned with reflective governance. He is dispositionally optimistic, non-competitive, and emotionally secure. He does not seek validation through dominance or recognition, and he does not experience psychological reward from retaliation or comparison.

As a result, his presence subtly reshapes social environments. Patients relax. Colleagues lower their guard. Potential conflicts dissolve before forming. None of this is strategic. It is an emergent consequence of low-threat identity organization.

Agency here is not merely autocatalytic but generative. The physician does not need to manage impressions or protect self-worth. He simply acts, and the world responds cooperatively. Minor setbacks are framed as information. Errors are metabolized as learning. Reflection functions as curiosity rather than defense.

This case illustrates the upper bound of integrated SPI. Identity-level coherence does not merely sustain agency; it improves the local world in which agency is exercised.

2.4.5 Structural Lessons

These four cases show that differences in agency do not reduce to intelligence, effort, values, or reflective capacity. All four physicians deliberate, plan, and articulate reasons. What differs is the identity-level motivational architecture organizing those processes.

Compensatory SPI sustains agency at high cost. Integrated SPI sustains it efficiently. Runaway SPI sustains it destructively. Generative SPI expands it outward.

These distinctions cannot be captured by reflection-centered theories alone. They require recognizing identity-level motivation as a distinct organizing layer within the architecture of agency.

3. Alignment, Identity Drag, and Cognitive Bandwidth

The contrastive cases in Section 2.4 show that agents can exhibit equally coherent, long-term patterns of action while differing radically in the subjective cost, flexibility, and sustainability of their agency. These differences do not track intelligence, effort, values, or deliberative competence. They track the relation between reflective governance (Layer 1) and Subconscious Practical Identity (Layer 2).

This section develops three connected ideas. First, it distinguishes alignment from misalignment as architectural relations rather than intradeliberative conflicts. Second, it introduces identity drag as the functional cost imposed by persistent misalignment. Third, it explains how these costs are realized through cognitive bandwidth, understood as the finite regulatory resources shared across layers of agency.

3.1 Alignment and Misalignment Across Layers

An agent's motivational architecture can be aligned or misaligned depending on how SPI relates to reflective governance. Alignment occurs when identity-level motivational structures operating beneath reflection support, rather than undermine or bypass, the agent's explicit commitments, values, and self-understanding. Misalignment occurs when SPI organizes action in ways that conflict with, distort, or silently override reflective governance.

Crucially, alignment does not require reflective authorship of SPI. In many ordinary cases, identity-level motivation precedes reflection developmentally and remains largely inaccessible to it. What matters is not origin but fit. When SPI pulls the agent toward outcomes that reflective governance can endorse—or at least does not experience as alien—agency feels coherent even in the absence of explicit planning. Reflection does not generate the trajectory; it stabilizes and interprets it.

Misalignment, by contrast, arises when SPI exerts teleological pressure toward outcomes that reflective governance cannot accurately recognize, articulate, or evaluate. Importantly, this conflict need not appear within deliberation itself. The agent may feel unified, disciplined, and motivated, as in the Insecure Doctor or Addict Doctor cases. The tension lies between layers rather than within reflection.

This point corrects a common diagnostic error. Classical theories often treat agency failure as arising from conflicts among reflectively accessible states—competing desires, evaluative inconsistency, or weakness of will. But in many cases of misalignment, reflective governance is functioning exactly as designed. What fails is not deliberation, but the assumption that deliberation has access to the full set of motivational determinants it is meant to regulate.

3.2 Identity Drag

Persistent misalignment between reflective governance (Layer 1) and Subconscious Practical Identity (Layer 2) imposes a characteristic functional cost: *identity drag*. Identity drag is the chronic expenditure of regulatory resources required to sustain action when identity-level motivation and reflective self-understanding are out of sync.

When layers are aligned, agency benefits from an efficient control economy. Commitments sustain themselves. Deliberation is selective rather than constant. Reflective intervention is required primarily in response to genuine novelty or conflict. This pattern is evident in the Secure Doctor and Vibes Doctor cases, where agency feels fluent rather than effortful.

When layers are misaligned, reflective governance must continually compensate. Because Layer 1 lacks representational access to the true source of motivational pressure, it works harder to justify, stabilize, or narratively repair a trajectory whose organizing force lies elsewhere. This compensation manifests as chronic rumination, evaluative anxiety, decision fatigue, and a persistent sense of effortfulness even in domains where competence and commitment are otherwise high.

Identity drag does not require conscious conflict. Agents may experience themselves as motivated and coherent while nonetheless paying a continual regulatory cost. This explains why two agents can exhibit similar outward behavior—long hours, discipline, achievement—while differing radically in exhaustion, brittleness, and resilience. The difference lies not in what they do, but in how much control effort is required to keep doing it.

3.3 The Asymmetry of Alignment

The foregoing considerations reveal a crucial asymmetry in the architecture of agency: Integrated forms of Subconscious Practical Identity scaffold reflective governance without systematically distorting it, whereas misintegrated or compensatory SPI structures impose chronic regulatory costs and distort reflective self-understanding. I call this the *asymmetric alignment principle*.

This asymmetry explains why identity-level motivation is often invisible in cases of smooth agency. When SPI and reflective governance are aligned, there is little phenomenological pressure to notice the identity-level structure at all. Agency "just works." By contrast, misalignment produces drag, distortion, and compensatory narrative activity, drawing attention to itself through effort, depletion, or instability.

The asymmetry also explains why negative cases are diagnostically salient while positive cases often go unnoticed (Kahneman and Tversky 1979). Integrated identity-level motivation leaves fewer diagnostic traces because it does not interfere with reflection. Misaligned SPI, by contrast, generates symptoms—overregulation, exhaustion, rationalization—that invite explanation.

3.4 Cognitive Bandwidth and the Economy of Agency

Identity drag can be further clarified by appeal to cognitive bandwidth. Agency operates under finite regulatory resources: attention, working memory, executive monitoring, and affective regulation. These resources are shared across layers.

Aligned SPI structures offload regulatory work from reflective governance. They pre-filter options, stabilize priorities, and reduce the need for constant self-monitoring. Reflection is free to operate opportunistically rather than defensively.

Unintegrated SPI has the opposite effect. Because reflective governance cannot represent the source of motivational pressure, it must manage its effects indirectly. Control becomes reactive rather than anticipatory. Over time, bandwidth is diverted away from learning, adaptation, and flexible planning toward mere maintenance.

This framework helps explain why identity misalignment is often misdiagnosed as weakness of will or poor self-control. The issue is not insufficient executive capacity, but inefficient allocation of control resources driven by architectural misfit.

3.5 Summary

Alignment and misalignment between reflective governance and Subconscious Practical Identity determine not only what agents do, but how agency feels and how costly it is to sustain. Identity drag explains why misalignment produces exhaustion without overt conflict and why integration produces fluency without explicit planning. These phenomena cannot be captured by deliberation-centered models alone, because they arise from relations between motivational layers rather than from failures within reflection.

The next section situates classical theories of action within this layered framework, showing both what they illuminate and what they systematically overlook.

4. Classical Theories Across the Layers

The three-layer model does not aim to displace dominant theories of action. Its ambition is diagnostic rather than eliminative: to show which strata of agency classical frameworks successfully illuminate, and where their explanatory reach ends. Davidsonian, Melean, Smithian, and related accounts capture important aspects of reflective governance and deliberative control, but they do so against a shared background assumption—namely, that the organizing forces of agency are transparent, or at least accessible in principle, to reflection. Subconscious Practical Identity (SPI) exposes the limits of that assumption.

This section situates several canonical theories of action within the layered architecture developed in Section 2. Each framework is shown to map cleanly onto a particular layer of agency while systematically overlooking identity-level motivational organization operating beneath reflective governance. The point is not that these theories are mistaken, but that they are partial. They explain the reflective surface of agency while leaving its deeper load-bearing structures untheorized.

4.1 Davidson: Reasons, Rationalization, and Reflective Authority

Davidson's causal theory of action explains intentional action in terms of primary reasons—belief–desire pairs that both cause and rationalize an agent's behavior from their own point of view (Davidson 1971). For an action to count as genuinely agential, it must be intelligible as something the agent saw reason to do. This requirement places reflective avowal at the center of agency explanation.

Within the layered model, Davidson's account is best understood as a theory of Layer 1 reflective governance. It captures how agents understand, justify, and narrate their actions when asked to explain themselves. In that respect, it remains one of the most powerful tools for analyzing the intelligibility of action.

However, SPI-driven agency exposes a structural limitation of Davidson’s framework. When long-term behavior is organized by identity-level motivational architecture that is not reflectively accessible, the actual source of teleological organization does not function as a rationalizing reason from the agent’s own perspective. Reflective explanations remain sincere and coherent, but they mislocate the organizing force of the trajectory.

This generates a dilemma for Davidsonian explanation. If reasons must be reflectively avowable to count as explanatory, then large classes of coherent, purposive behavior fall outside the scope of agency. If unavowable identity-level structures are permitted to count as reasons, then the rationalization requirement loses its distinctive force. Either way, Davidson’s theory captures the reflective interface of agency while remaining blind to deeper motivational organization.

4.2 Mele: Motivational Conflict, Akrasia, and the Limits of Deliberative Diagnosis

Mele’s work focuses on failures of rational self-control: akrasia, self-deception, evaluative conflict, and weakness of will (Mele 1987)[1]. His analyses presuppose that the relevant motivational states are available—at least in principle—to deliberation, and that agency fails when reflective processes malfunction, conflict, or are overridden.

SPI-driven misalignment does not fit this model. In many cases, there is no felt conflict within deliberation at all. The agent experiences themselves as coherent, motivated, and even disciplined. The tension lies beneath reflection, between reflective governance and identity-level motivational structures that never enter the deliberative arena.

[1] Aristotle’s treatment of akrasia already resists a purely deliberative diagnosis, treating it as a structural conflict between reason and desire rather than a simple failure of reflective judgment (Nicomachean Ethics VII). The present paper does not engage in Aristotelian exegesis, but its architectural approach is more continuous with that tradition than with contemporary intention-centered accounts of weakness of will.

As a result, SPI-induced agency costs are often misdiagnosed as failures of self-control. But the problem is not insufficient executive regulation; it is architectural. Reflective governance expends continual regulatory effort to manage effects whose causes it cannot represent. Mele's framework accurately diagnoses failures *within* Layer 1, but it lacks the resources to explain failures that arise from misalignment *between* layers.

4.3 Smith: Ideal Reflection and the Inheritance of Blind Spots

Smith's ideal advisor theory identifies an agent's real reasons with those they would endorse under conditions of full information and rationality (Smith 1994). The aim is to preserve the authority of the agent's values while filtering out distortion, ignorance, and error.

This model presupposes that deeper motivational structures are, in principle, accessible to idealized reflection. SPI challenges that presupposition. Identity-level motivational architecture may be structurally unavailable to reflection—not merely hidden by ignorance, but encoded affectively and developmentally rather than propositionally.

In such cases, the ideal advisor inherits the blind spots of the reflective standpoint it idealizes. Compensatory values generated by unintegrated SPI appear as authentic commitments rather than as artifacts of identity-level misalignment. The framework therefore stabilizes distorted trajectories instead of diagnosing them.

Smith's account remains compelling as a theory of reflective endorsement. What it does not explain is how reflective endorsement itself can be systematically miscalibrated by motivational structures operating beneath it.

4.4 Doris: From Situationism to Reflective Narrativization

Doris’s early situationist work emphasizes the extent to which local behavior is shaped by situational factors rather than stable character traits (Doris 2002). On its own, this emphasis risks flattening agency into a sequence of context-sensitive responses, underestimating the role of long-term motivational organization.

However, Doris’s later work marks an important shift. In developing an account of reflection as narrativizing rather than governing, Doris argues that reflective self-understanding often functions to make sense of behavior after the fact, rather than to generate or control it (Doris 2015). Reflection, on this view, constructs intelligible self-narratives without occupying a supervisory role in action production.

This later position aligns closely with the layered model. Treating reflection as narrativizing rather than sovereign allows for the possibility that agency is organized elsewhere—by non-reflective structures that reflection interprets rather than commands. What Doris does not provide, however, is a positive account of what those organizing structures are, or how they generate cross-temporal coherence.

SPI fills that gap. It explains how long-term purposive organization can arise beneath reflection while still producing the kind of narrative intelligibility Doris describes. In this sense, Doris’s later work offers partial support for the layered view, even if it stops short of articulating a full architectural alternative.

4.5 Interim Summary: Local Illumination, Global Incompleteness

Each of the classical theories examined here captures a genuine stratum of agency. Davidson explains reflective rationalization. Mele explains deliberative conflict and breakdown. Smith explains idealized endorsement. Doris explains the narrativizing function of reflection and the limits of character-based explanation.

What none of these frameworks explains is identity-level motivational architecture operating beneath reflective governance. They presuppose that agency either appears within reflection or collapses into arationality. The layered model shows that this is a false dichotomy.

Agency can be purposive, coherent, and life-structuring without being reflectively authored. Recognizing this middle layer does not undermine classical theories; it situates them within a more complete architecture of human action.

5. Bratman and the Limits of Planning-Centered Agency

Among contemporary theories of action, Bratman's planning theory offers the most sophisticated account of long-term agency grounded in reflective structure. Unlike Davidsonian models that focus on momentary reason–action explanations, Bratman emphasizes the role of future-directed intentions and plans in organizing conduct across time. Intentions, on this view, are not merely commitments to act; they are elements of a planning system that stabilizes deliberation, coordinates future behavior, and supports diachronic self-governance (Bratman 1987).

For this reason, Bratman's framework represents the strongest reflection-centered account of extended agency. If any theory can explain coherent life trajectories without appeal to non-reflective motivational structure, it is this one. The pressure posed by Subconscious Practical

Identity (SPI) is therefore most acute here. The question is not whether Bratman captures something real—he plainly does—but whether planning intentions exhaust the sources of long-term practical organization.

5.1 What SPI Shares with Planning Intentions

SPI structures replicate several of the functional roles Bratman assigns to planning intentions.

First, SPI supports cross-temporal stability. Agents governed by identity-level motivational architecture often sustain careers, relationships, and styles of life over decades with remarkable consistency. This stability is not episodic or accidental; it exhibits the same kind of diachronic coherence Bratman treats as distinctive of planning agency.

Second, SPI constrains deliberation. Bratman emphasizes that intentions function as filters on future reasoning: once an intention is in place, agents do not continually reopen deliberation over settled matters (Bratman 1987, 29–31). SPI performs an analogous role, often more pervasively. Identity-level motivation determines which options ever appear as live possibilities, shaping deliberation before reflective choice begins.

Third, SPI enables diachronic coordination. Bratman highlights how plans coordinate an agent's actions with one another and with the actions of others over time. SPI likewise stabilizes patterns of prioritization, effort, and responsiveness, allowing behavior to remain coherent without continuous reflective re-endorsement.

In these respects, SPI behaves functionally like long-term intention. It explains how agents maintain organized trajectories without relying on constant planning or deliberative supervision.

5.2 Where SPI Violates Bratman's Conditions for Intention

Despite these similarities, SPI cannot be assimilated to Bratmanian intention without distorting the planning framework. Bratman's account imposes psychological and normative conditions on intentionhood that SPI systematically violates.

First, reflective accessibility. Bratmanian intentions are states the agent can cite, revise, and reason from. SPI structures are largely inaccessible to reflection. Agents typically cannot represent them propositionally or bring them under direct deliberative control.

Second, endorsement and revisability. Intentions, for Bratman, are subject to norms of consistency and means–end coherence enforced through reflective monitoring. SPI structures persist even when reflective governance would reject them if made explicit, and they are not directly revisable through deliberation.

Third, representational format. Bratman's intentions are propositional attitudes embedded in a planning system. SPI is affectively and developmentally encoded. It exerts pressure through salience, attraction, aversion, and emotional regulation rather than through explicit content.

SPI is therefore intention-like without being an intention. It organizes action across time while violating the psychological assumptions that make planning intentions suitable objects of reflective governance.

5.3 The Bratman Dilemma

SPI poses a dilemma for planning-centered theories of agency.

One option is to restrict the category of intention to reflectively accessible planning states, as Bratman does. On this view, SPI is excluded by definition. But this response concedes the substantive point: the planning framework then fails to explain a large class of long-term, coherent, purposive behavior that is clearly agentive but not reflectively planned.

The alternative is to broaden the category of intention to include non-reflective, identity-level motivational structures. But this move collapses the distinction between planning systems and deeper motivational architecture, diluting the role of reflective governance and eroding the normative pressures Bratman treats as constitutive of intention.

Either way, SPI exposes the limits of a planning-centered account. Bratman's theory accurately characterizes reflective long-term agency. It does not capture identity-driven long-term agency—the form most responsible for shaping human lives.

5.4 Planning as a Local, Not Global, Account of Agency

The lesson is not that Bratman's theory is mistaken. It is that it is locally correct but globally incomplete. Planning intentions are one layer of agency, not its foundation. They presuppose a background motivational architecture that determines which plans are formed, sustained, or abandoned in the first place.

SPI supplies that background. It explains why some agents rely heavily on explicit planning while others exhibit coherent trajectories with minimal deliberation, and why planning often fails to restore agency in cases of deep misalignment. When identity-level motivation and reflective governance are out of sync, adding plans does not resolve the underlying architectural tension.

Recognizing SPI therefore situates Bratman's theory within a layered model rather than opposing it. Planning is a powerful instrument of reflective governance, but it is not the source of long-term agency itself.

6. Situated Agency, Epistemic Capture, and the Fragility of Reflective Governance

So far, the layered model has treated failures of agency primarily as intra-agent phenomena, arising from misalignment between reflective governance (Layer 1) and Subconscious Practical Identity (Layer 2). But agency is not exercised in isolation. It is embedded in social, linguistic, and normative environments that can systematically shape how agents interpret their own actions, reasons, and commitments.

Once this broader context is taken seriously, further vulnerabilities emerge. Reflective governance can be undermined not only by internal motivational dynamics, but also by external epistemic pressures, by identity-level dominance over interpretation, and by breakdowns in the control mechanisms linking motivation to action. This section examines these forms of architectural fragility, moving from socially situated epistemic capture, through intra-agent failures of reflective authority, to non-epistemic disruptions of action execution. It concludes by situating these failure modes within a provisional taxonomy and by clarifying the limits of value-based accounts of agency.

6.1 Situated Agency and Epistemic Capture

Agency is exercised in social and normative environments, not in isolation.
Reflection—understood as the agent's capacity to articulate reasons, assess commitments, and regulate action—depends on epistemic scaffolding supplied by those environments. This

dependence introduces a distinctive architectural vulnerability: even when reflective governance (Layer 1) is intact in principle, it can be systematically compromised by conditions that distort the agent's access to their own reasons, experiences, and motivational structure.

This phenomenon can be described as *epistemic capture*. Epistemic capture occurs when the informational and interpretive environment an agent inhabits progressively displaces their own reflective authority. Rather than merely influencing what the agent believes, the environment reshapes the conditions under which beliefs about oneself and one's reasons are formed. The agent does not simply acquire false beliefs; they lose reliable access to the standpoint from which such beliefs could be evaluated as their own.

The clearest illustrations of epistemic capture appear in the literature on gaslighting. In canonical cases, an interlocutor persistently denies, reframes, or pathologizes the agent's perceptions, memories, or emotional responses. Over time, the agent's confidence in their own interpretive capacities erodes, and external narratives come to function as epistemic substitutes for first-person judgment. Importantly, this process does not require irrationality or cognitive deficit. The captured agent may remain articulate, reflective, and logically competent, yet no longer occupy a position of epistemic authority with respect to their own experience (Abramson 2014).

From the perspective of the layered architecture developed earlier, epistemic capture operates by severing the normal coupling between reflective governance (Layer 1) and the motivational architecture it is meant to interpret and regulate. Reflective governance relies on memory, narrative coherence, and trust in one's own evaluative responses. When environmental pressures systematically destabilize these capacities—by denying emotional signals, rewriting shared histories, or enforcing authoritative interpretations—reflection loses its grip on the motivational structures that guide action. It continues to function, but as an interpretive surface increasingly governed by external inputs.

This point is crucial. Epistemic capture is not primarily a failure of reasoning or deliberation. Classical theories of action tend to locate agency failure in distorted belief formation, motivational conflict, or weakness of will. Epistemic capture involves none of these in the first instance. The agent may deliberate competently and endorse intelligible reasons. What fails is the architecture that allows those reasons to be *the agent's*, rather than imposed interpretive artifacts.

Situational epistemic capture therefore reveals a limitation of reflection-centered models of agency. Such models implicitly assume that reflective access, once present, is self-stabilizing. But reflection is not self-grounding. Its authority depends on a surrounding epistemic environment that preserves the agent's ability to treat their own experiences and evaluations as authoritative inputs. When that environment becomes hostile or systematically distorting, reflective governance can be hollowed out without collapsing into irrationality or arational behavior.

The significance of this failure mode extends beyond cases of overt manipulation. Institutional settings, social roles, and normative cultures can exert similar pressures without any identifiable manipulator. When agents are embedded in environments that reward conformity, suppress dissent, or reinterpret self-trust as pathology, reflective agency becomes fragile. The agent continues to act coherently and purposefully, yet their self-understanding is increasingly authored elsewhere.

Epistemic capture thus marks a first point at which agency can persist without self-governance. Action remains intelligible and coordinated, but reflective authority is no longer internal. This vulnerability is not accidental; it follows directly from the layered architecture of agency. Once reflection is no longer treated as a sovereign controller but as an interface dependent on upstream motivational structure and downstream social scaffolding, its susceptibility to capture becomes intelligible.

The next subsection extends this insight inward. If reflective governance can be displaced by external environments, it can also be displaced by the agent's own identity-level motivational architecture. The resulting phenomenon—runaway Subconscious Practical Identity—represents a deeper and more self-sealing form of epistemic capture.

6.2 Runaway Subconscious Practical Identity and Intra-Agent Epistemic Capture

Situational epistemic capture shows how reflective governance can be undermined by external environments. A deeper and more troubling vulnerability arises when a similar displacement occurs from within the agent themselves. In such cases, Subconscious Practical Identity (SPI) does not merely guide action beneath reflection; it begins to dominate the interpretive function of reflection itself. This phenomenon can be described as *runaway Subconscious Practical Identity*—a form of intra-agent epistemic capture in which identity-level motivation annexes reflective interpretation (Obdan 2025).

Runaway SPI occurs when identity-level motivational architecture progressively annexes the mechanisms through which reflective governance interprets reasons, experiences, and self-conceptions. Rather than standing in a regulative relation to SPI, reflection becomes its expressive instrument. The agent continues to deliberate, explain, and justify their actions, but those activities no longer function as checks on identity-level motivation. They function instead as narrative elaborations that stabilize and protect it.

This is not ordinary misalignment. In cases of misalignment, reflective governance retains the capacity to register tension, discomfort, or motivational drag. The agent experiences effort, ambivalence, or unease, even if they cannot fully articulate its source. Runaway SPI marks a different structural condition. Here, the interpretive channel itself is compromised. Reflection no longer has access to signals that would indicate misalignment, because those signals are filtered, reinterpreted, or excluded before they can register.

The resulting phenomenology is often one of clarity rather than confusion—a sense of coherence achieved through epistemic closure rather than integration. Agents subject to runaway SPI typically experience themselves as coherent, justified, and self-knowing. Their reasons make sense to them. Their narratives are fluent. What is lost is not intelligibility but epistemic independence. Reflection ceases to function as an autonomous standpoint from which identity-level motivation could be evaluated.

This pattern closely mirrors the structure of interpersonal gaslighting, but with a crucial difference. In canonical gaslighting cases, an external interlocutor supplies the narrative pressure that destabilizes self-trust. In runaway SPI, the pressure originates internally. Identity-level motivational architecture generates emotionally stabilizing narratives, and reflective governance is recruited to maintain them. The roles of speaker and hearer collapse into a single system.

Seen this way, gaslighting is not exclusively interpersonal. Its most structurally significant form may be intra-agent. Agents construct self-interpretations that regulate affect, preserve identity coherence, and minimize threat, and then subject their own reflective judgments to those interpretations. Memory, intention attribution, and evaluative assessment are retroactively reorganized to fit the demands of identity-level stability.

This dynamic helps explain why some agents remain impervious to counterevidence, feedback, or self-reflection even in the absence of social manipulation. Classical accounts of self-deception typically presuppose conflict within reflection—competing beliefs, selective attention, or motivated reasoning. Runaway SPI involves no such conflict. Reflective governance has not been *overridden*; it has been *absorbed*.

The layered architecture clarifies how this absorption is possible. Reflective governance does not generate identity-level motivation; it interprets and regulates it. When reflective capacities are weak, underdeveloped, or bypassed—whether due to developmental history, temperament, or prolonged stress—SPI can expand unchecked. Over time, reflection loses its status as an independent epistemic interface and becomes a vehicle for identity-preserving narration.

This condition is especially likely when reflective skills such as metacognition, perspective-taking, and affective awareness are poorly developed. In such agents, the mechanisms required to interrogate identity-level motivation are never fully online. Under threat, SPI does not encounter resistance. It becomes self-authorizing, generating narratives that insulate it from correction while maintaining subjective coherence.

Runaway SPI therefore represents a terminal failure mode of misalignment. It is not merely that reflective governance and SPI pull in different directions. It is that the very conditions under which such divergence could be recognized have collapsed. Agency persists, often with impressive coherence and effectiveness, but it is no longer self-correcting.

This diagnosis also clarifies why such agents can appear decisive, confident, or even admirable from the outside. Runaway SPI can support disciplined action, long-term projects, and apparent integrity. What distinguishes it from healthy integration is not behavioral chaos but epistemic closure. The architecture has achieved stability by sacrificing reflective independence.

In this respect, runaway SPI marks a boundary condition of agency rather than its negation. The agent continues to act purposively and intelligibly, but the layered structure that allows agency to revise itself from within has been compromised. Understanding this failure mode requires abandoning the assumption—shared by much of action theory—that reflection is always available as a supervisory authority. Sometimes reflection remains articulate and fluent, but no longer free.

The next subsection isolates a different kind of vulnerability altogether. Parkinsonian motor gating failure shows that agency can break even when reflective authority and identity-level motivation remain intact, revealing a non-epistemic and non-identity-based breakdown within the architecture of action.

6.3 Parkinsonism and Motor-Gating Failure: A Non-Epistemic Breakdown of Agency

Parkinsonism provides a philosophically underexploited dissociation within the architecture of agency. Unlike cases of epistemic capture or runaway Subconscious Practical Identity (SPI), Parkinsonian agency failure does not primarily involve distorted self-interpretation, identity dominance, or reflective annexation. Instead, it reveals a breakdown in the coupling between higher-order motivational states and action initiation. Intentions remain intelligible, endorsed, and often motivationally sincere, yet fail to issue in movement.

Clinically, Parkinson’s disease is characterized by bradykinesia, akinesia, rigidity, and resting tremor, with cognitive function often preserved in early and mid stages. Patients frequently report wanting to act, intending to move, or endorsing reasons for action, while being unable to initiate the corresponding behavior. Crucially, this failure is not experienced as akrasia, indecision, or motivational ambivalence. The agent does not feel torn or conflicted. They feel blocked.

From the perspective of the layered model, this pattern is neither a failure of reflective governance (Layer 1) nor a distortion of identity-level motivation (Layer 2). Reflective agency remains articulate and normatively intact. SPI often remains stable: patients continue to care about projects, relationships, and self-conceptions that predate the onset of motor symptoms. What fails is the gating function that normally allows Layer-1 and Layer-2 states to recruit motor execution systems.

This diagnosis aligns with contemporary neurobiological accounts of Parkinsonism, which locate the core deficit in dopaminergic disruption of cortico–basal ganglia–thalamic loops responsible for action initiation and motor selection. These circuits do not merely execute motor commands; they regulate which potential actions are released for execution. When this gating mechanism is compromised, intentions can be formed and sustained without being able to trigger bodily movement (Jankovic 2008; Mink 1996; Redgrave, Prescott, and Gurney 1999).

Philosophically, this matters because it exposes a failure mode orthogonal to those emphasized in action theory. The agent’s reasons are intact. Their evaluative judgments are stable. Their motivational identity remains largely unchanged. Yet agency falters. This shows that the architecture of action includes a control interface downstream from intention and identity but upstream from motor output—an interface not captured by deliberative, epistemic, or identity-based accounts alone.

The significance of Parkinsonism becomes even clearer when one attends to Layer-3 leakage. Despite profound difficulty initiating voluntary movement, Parkinsonian patients often retain a range of arational and procedural responses. Startle reactions, affective expressions, reflexive movements, and externally cued actions can remain partially preserved. In some cases, patients who cannot voluntarily initiate walking are able to step over visual cues or respond automatically to sudden stimuli (Nieuwboer et al. 2007).

This dissociation is philosophically revealing. It shows that arational actions are not intention-based by default. Layer-3 processes can bypass impaired gating mechanisms and issue directly in behavior. The persistence of such responses confirms the autonomy of arational–procedural agency and reinforces the layered distinction developed earlier. Action execution is not a single pipeline flowing from intention to movement; it is a plural system with multiple access routes to behavior.

Parkinsonism therefore undermines a widespread but often implicit assumption in philosophy of action: that failures to act on one's intentions must reflect either motivational weakness or deliberative breakdown. In Parkinsonism, neither diagnosis applies. The agent's practical reasoning is intact, their self-understanding is undistorted, and their motivational commitments remain in place. What is missing is the capacity to translate endorsed intentions into bodily action through the normal control architecture.

While philosophers have occasionally gestured at Parkinsonism in discussions of weakness of will or motor incapacity, it has rarely been integrated into a systematic architecture of agency. By situating Parkinsonism within a layered model, we can see it not as an anomaly but as a structurally illuminating case: one that isolates a specific control interface whose failure leaves other layers intact.

This insight also points beyond the present paper. Parkinsonism suggests that agency can fragment not only along epistemic or identity lines, but along control-mechanical ones. Understanding how such failures arise, and how compensatory pathways such as external cueing temporarily restore action, opens a path toward a more genuinely interdisciplinary philosophy of agency—one that takes neurobiology seriously without collapsing agency into mechanism.

6.4 Architectural Failure Modes: A Provisional Taxonomy of Agency Breakdown

The preceding sections identified two distinct but interacting vulnerabilities in human agency: situational epistemic capture, in which external environments distort reflective governance, and intra-agent epistemic capture, in which Subconscious Practical Identity (SPI) annexes the interpretive function of reflection itself. They also highlighted an orthogonal failure mode—motor gating breakdown in Parkinsonism—in which reflective authority and identity-level motivational organization can remain largely intact while action initiation fails. These phenomena suggest that failures of agency cannot be adequately described as isolated lapses of rationality, motivation, or willpower. Instead, they reflect structural failure modes within a layered architecture of control.

This section offers a provisional taxonomy of such failure modes. The aim is diagnostic rather than exhaustive: to distinguish architecturally different ways in which agency can succeed, strain, or fail, depending on how reflective governance (Layer 1), identity-level motivation (Layer 2), and arational–procedural processes (Layer 3) interact, misalign, or collapse. The categories below are unified not by surface behavior or moral diagnosis, but by patterns of inter-layer relation.

6.4.1 Alignment and Control Economy

At one end of the spectrum of agency architecture lies alignment. In aligned agents, Subconscious Practical Identity (SPI) and reflective governance exert compatible pressures on action. Identity-level motivational structures scaffold deliberation rather than distort it, and reflective governance can interpret, endorse, and regulate those motivations without persistent friction.

Alignment does not require transparency. SPI may remain largely inaccessible to reflection while still supporting coherent agency. What matters is not that identity-level motivation is reflectively articulated, but that it does not systematically undermine reflective self-understanding or regulation. When SPI pulls agents toward forms of life that reflection can recognize as intelligible or acceptable, agency remains stable even in the absence of explicit planning or deep self-analysis.

Architecturally aligned systems exhibit an efficient control economy. Reflective governance intervenes selectively rather than chronically. Deliberation is reserved for genuinely novel, conflicting, or high-stakes decisions, rather than being continuously mobilized to stabilize motivation. Action unfolds with relatively low regulatory cost, and effort is directed outward rather than inward.

Phenomenologically, alignment is experienced as fluency. Agents report a sense that their lives “make sense,” not because they are constantly monitoring themselves, but because fewer corrective operations are required. Commitments sustain themselves, priorities remain stable, and setbacks are integrated without threatening identity coherence.

Importantly, alignment is not equivalent to moral virtue, psychological insight, or explicit self-authorship. An agent may be deeply unreflective yet well aligned. What distinguishes alignment is not self-knowledge but structural fit: identity-level motivation and reflective governance are pulling in roughly the same direction.

6.4.2 Misalignment and Identity Drag

A more common and diagnostically important condition is misalignment. In misaligned agents, SPI and reflective governance exert competing pressures while remaining mutually legible. Reflective governance retains the capacity to register discomfort, tension, or motivational resistance, but lacks the resources to resolve it directly.

Misalignment does not necessarily manifest as akrasia or deliberative conflict. Agents may endorse their projects, articulate intelligible reasons for their actions, and experience themselves as motivated and disciplined. What distinguishes misalignment is the cost of sustaining agency. Action requires disproportionate regulatory effort. Deliberation becomes repetitive.
Self-correction is frequent and exhausting.

This cost can be described as identity drag: the persistent expenditure of cognitive and motivational resources required to counteract identity-level motivational structures that do not fully align with reflective commitments. Reflective governance must continually compensate for pressures whose sources it cannot represent, generating chronic rumination, vigilance, or affective volatility even in the absence of overt conflict.

Crucially, misalignment presupposes a functioning interpretive channel between layers. The agent can feel that something is off, even if they cannot articulate what. This discomfort is not a defect of agency but a sign of its integrity. Misalignment indicates that reflective governance is still operative and still capable of registering tension.

For this reason, misalignment is often a precondition for change. Identity drag brings architectural strain into awareness, creating the possibility of insight, integration, or reorganization. Classical theories of action often misdiagnose such cases as weakness of will or motivational deficiency. On the layered model, the problem lies not in insufficient control but in excessive compensatory control.

Misalignment therefore occupies a middle position in the taxonomy. It is neither efficient alignment nor terminal breakdown. Agency remains self-correcting, but at a cost. The discomfort it produces is a structural signal, not merely a psychological symptom.

6.4.3 Runaway SPI and Centrifugal Identity Dominance

Beyond misalignment lies a more severe architectural failure mode: runaway Subconscious Practical Identity (SPI). In these cases, identity-level motivational architecture does not merely compete with reflective governance; it progressively displaces it. SPI becomes the dominant organizing force not only of action but of interpretation, annexing the mechanisms through which reflective governance would normally register tension or evaluate reasons.

In runaway SPI, reflection does not oppose identity-level motivation. It serves it. Reflective deliberation, self-explanation, and narrative construction are recruited to stabilize and protect SPI rather than to regulate it. The agent continues to reason, justify, and articulate values, but these activities function as downstream rationalizations of identity-level imperatives rather than as independent checks on them.

This condition differs structurally from misalignment. In misalignment, reflective governance remains epistemically independent and can register discomfort or motivational drag. In runaway SPI, that interpretive independence collapses. Signals that would normally indicate tension—fatigue, inconsistency, affective disturbance—are filtered, reinterpreted, or excluded before they can gain traction. The result is not confusion but apparent clarity.

Runaway SPI therefore tends to produce agents who experience themselves as confident, decisive, and internally coherent. Their reasons make sense to them. Their narratives are fluent. What is lost is not intelligibility but epistemic openness. The architecture has achieved stability by eliminating internal friction rather than resolving it.

This centrifugal expansion of identity-level motivation is especially likely when reflective capacities such as metacognition, perspective-taking, and affective awareness are weak, underdeveloped, or developmentally compromised. In such agents, reflective governance lacks the structural resources required to interrogate SPI. Under pressure, identity-level motivation becomes self-authorizing, generating narratives that insulate it from corrective feedback.

Importantly, runaway SPI is not inherently pathological in its behavioral profile. Agents in this condition can be highly functional, disciplined, and socially successful. What distinguishes runaway SPI from healthy integration is not outward disorder but epistemic closure. Reflection no longer functions as a site of genuine evaluation. Agency persists, but it is no longer self-correcting.

6.4.4 Collapse of Reflective Governance

A related but distinct failure mode involves the collapse or atrophy of reflective governance itself. Whereas runaway SPI involves the domination of reflection by identity-level motivation, collapse of reflective governance involves the erosion of reflection as an operative control interface.

In such cases, Layer 1 does not merely lose authority; it loses functional integrity. Deliberation becomes episodic, reactive, or entirely absent. The agent may continue to act coherently, guided by SPI and procedural routines, but lacks the capacity for higher-order regulation, self-assessment, or revision.

This collapse can arise through multiple pathways. Prolonged epistemic capture, developmental deprivation, or sustained environments that systematically bypass reflective agency can all erode the conditions under which reflection operates. Over time, reflective governance ceases to function as an independent epistemic standpoint and becomes either vestigial or purely expressive.

The behavioral consequences of reflective collapse are heterogeneous. Some agents appear impulsive or erratic; others appear rigid and habitual. What unifies these cases is not surface behavior but architectural structure. Agency is no longer regulated through reflective endorsement or deliberative control. Identity-level motivation and procedural mechanisms dominate by default.

It is important to distinguish collapse of reflective governance from arational action. In collapse cases, behavior remains teleological and identity-guided rather than purely reactive. The agent's life may exhibit long-term coherence, but it is closed to revision from within. There is no longer an operative layer capable of assessing or renegotiating commitments.

This failure mode underscores a central claim of the layered model: reflective governance is neither the source nor the guarantor of agency. It is a fragile interface whose operation depends on developmental, environmental, and architectural support. When that support erodes, agency does not necessarily disappear. It reconfigures around deeper motivational and procedural structures.

6.4.5 Self-Stabilizing Delusion and Epistemic Closure

In its terminal form, architectural breakdown yields *self-stabilizing delusion*. This condition represents not confusion or fragmentation, but excessive internal coherence achieved through epistemic closure. Identity-level motivational architecture (SPI) not only governs action and interpretation, but actively suppresses the conditions under which alternative interpretations could arise.

Unlike ordinary self-deception, self-stabilizing delusion does not involve episodic bias, motivated reasoning, or selective attention operating within an otherwise intact reflective framework. Instead, the reflective layer itself has been structurally repurposed. Reflection remains articulate, fluent, and internally consistent—but it no longer functions as an epistemic checkpoint. It functions as an enforcement mechanism.

In this state, discrepancies are not experienced as problems to be resolved. They are reclassified as noise, hostility, misunderstanding, or irrelevance. Counterevidence does not generate tension;

it is absorbed, reframed, or excluded before it can register as a challenge. From within the system, the world makes sense—often with striking confidence.

This form of delusion is self-stabilizing because it eliminates friction rather than managing it. The agent's narratives, memories, and value judgments are organized to preserve identity coherence at all costs. The system reaches a pathological equilibrium: internally ordered, externally impermeable, and resistant to disruption.

Crucially, this is not a breakdown of agency in the sense of passivity or loss of control. Agency persists. The agent acts, plans, justifies, and coordinates. What is lost is epistemic openness—the capacity for the system to register that something might be wrong with its own organizing principles.

Self-stabilizing delusion therefore marks the endpoint of intra-agent epistemic capture. It is the condition in which agency becomes closed-loop: self-maintaining, self-justifying, and no longer corrigible from within.

6.4.6 Provisionality and Programmatic Scope

The taxonomy developed in this section is intentionally provisional. Its aim is not to exhaustively classify all failures of agency, but to distinguish architecturally distinct ways in which agency can succeed, strain, or break depending on how reflective governance (Layer 1), identity-level motivation (Layer 2), and arational–procedural processes (Layer 3) interact.

Taken together, the failure modes identified here—misalignment and identity drag, runaway SPI, collapse of reflective governance, self-stabilizing delusion, and non-epistemic motor-gating failure—demonstrate that agency failure is not monolithic. Different breakdowns reflect different structural distortions, and they cannot be adequately captured by appeals to akrasia, weakness of will, or deliberative malfunction alone.

The inclusion of Parkinsonism alongside epistemic and identity-based failures reinforces this point. Agency can fail without confusion, distortion, or domination. Sometimes the agent knows exactly what they are doing and why. The system simply cannot move. Any reflection-centered theory that treats all agency failures as variants of deliberative error will systematically misdiagnose such cases.

The broader lesson is architectural. Human agency is a layered control system with multiple points of vulnerability. Failures can arise from epistemic capture, motivational dominance, loss of reflective integrity, or breakdowns in execution interfaces. Diagnosing agency therefore requires attention to structure rather than surface behavior or normative deviation.

A fuller theory would extend this taxonomy in at least three directions: first, by tracing developmental pathways into each failure mode; second, by specifying conditions under which reflective governance can be restored or rebuilt; and third, by examining how social environments interact with internal architecture to stabilize or destabilize agency over time.

Some of these extensions are pursued in companion work within the trilogy, while others remain open directions for future research. What matters here is the recognition that reflective supervision was never the sole linchpin of agency. Once that assumption is abandoned, agency failure appears not as a single phenomenon but as a family of structurally distinct breakdowns—each requiring its own diagnosis.

6.5 Valuation Architecture and the Limits of Reflective Access

Recent work in moral psychology and philosophy of action has increasingly emphasized the role of values in structuring agency. Among the most developed accounts is Chandra Sripada's valuationist model of human agent architecture, which treats action as guided by a hierarchically organized system of values rather than by isolated desires or momentary intentions (Sripada 2016). On this view, agency is explained by the interaction between value representations, evaluative updating, and decision mechanisms that select actions in light of what the agent cares about over time.

Sripada's framework marks a significant advance over deliberation-centered models. By shifting explanatory focus from episodic choice to standing evaluative structure, it captures how agency can exhibit diachronic coherence without requiring constant deliberation or explicit planning. Valuations constrain choice, stabilize priorities, and generate systematic patterns of action across contexts. In this respect, the valuationist model converges with the layered account developed here in rejecting the idea that agency is governed exclusively by moment-to-moment reflective endorsement.

However, despite this structural sophistication, Sripada's account retains a crucial assumption shared by classical theories: that the values doing the explanatory work are, at least in principle, available to reflective access. Valuations are treated as elements of the agent's practical point of view—states the agent can identify, articulate, and potentially revise through reflection, even if they are not always explicitly entertained.

This assumption marks the limit of the valuationist framework. Subconscious Practical Identity (SPI) is not merely a set of deeply held values whose influence is underestimated or whose articulation is deferred. SPI structures are identity-level motivational architectures that may never be available to reflection in value form at all. They are not simply unarticulated valuations; they are affectively encoded, developmentally sedimented, and teleologically operative without being representable as objects of endorsement.

The distinction matters because valuationist models explain agency by appeal to what agents care about, whereas SPI explains agency by appeal to how agents are organized. Two agents may endorse the same values, articulate similar commitments, and deliberate in similar ways, yet differ radically in motivational cost, resilience, and susceptibility to failure. These differences cannot be captured by valuation alone, because they arise from architectural relations between reflective governance and identity-level motivation rather than from differences in value content.

From the perspective of the layered model, Sripada's framework therefore occupies an intermediate position. It improves on reflection-centered theories by acknowledging non-episodic motivational structure, but it stops short of recognizing a motivational layer that operates independently of reflective access. Valuations explain how agents choose among options they recognize; SPI explains why certain options are recognized, sustained, or never seriously considered in the first place.

This contrast helps clarify the distinctive contribution of SPI. The claim is not that values are unimportant, nor that valuation architecture is misguided. It is that value-based explanations presuppose a background motivational organization that they do not themselves explain. SPI names that background structure.

Accordingly, Sripada's work should be read not as a competitor to the present model, but as a boundary case. It shows how far one can go in explaining agency by appealing to structured motivation while still remaining within the orbit of reflective accessibility. SPI marks the point at which that orbit is left behind.

7. Non-Supervisory Integration: Carl Rogers and Identity-Level Realignment

The preceding sections diagnosed a range of vulnerabilities in human agency. Reflective governance can be undermined by epistemic capture, displaced by identity-level motivational dominance, or rendered inert by downstream control failures. Together, these analyses leave a residual question unanswered: how can agency regain integration once reflective supervision has been compromised or abandoned? If reflection is neither sovereign nor reliable, what mechanism—if any—allows human agency to recover coherence rather than collapse into rigidity or drift?

Carl Rogers offers an answer that is strikingly consonant with the layered architecture developed here, despite emerging from a radically different intellectual tradition. Writing decades before contemporary philosophy of action turned its attention to motivational architecture, Rogers articulated a model of psychological integration that does not rely on reflective command, deliberative control, or value endorsement. Instead, he proposed that agency possesses an internal capacity for reorganization that operates beneath reflection and can re-establish coherence when obstructive constraints are removed (Rogers 1951).

What makes Rogers philosophically significant in the present context is not his therapeutic method, but his underlying theory of motivation and integration. Read through the lens of Subconscious Practical Identity (SPI), Rogers's claims cease to look like optimistic clinical intuition and instead appear as a systematic—if pre-formal—account of identity-level regulation.

7.1 The Organismic Valuing Process as Identity-Level Directionality

At the center of Rogers's theory is the organismic valuing process (OVP): a pre-reflective tendency of the person toward greater integration, vitality, and coherence. The OVP is not a deliberative faculty, a moral sense, or a set of endorsed values. It is a directional property of the motivational system itself.

Rogers insists that this process operates independently of reflective awareness. Individuals do not consult the organismic valuing process; they *are guided by it*. When unobstructed, it organizes experience, motivation, and action toward patterns that feel internally coherent and externally adaptive.

From the perspective of the layered model, the organismic valuing process cannot plausibly be located in reflective governance (Layer 1). It does not consist in reasons the agent can articulate, endorse, or revise. Nor is it arational in the sense of Layer 3. It is not reactive, episodic, or stimulus-bound. Instead, it exhibits precisely the features that characterize Subconscious Practical Identity: cross-temporal stability, affective encoding, teleological organization, and relative inaccessibility to reflection.

The organismic valuing process is therefore best understood as a regulative tendency operating at the identity-motivational level. It supplies direction without deliberation, organization without planning, and integration without command. That Rogers could identify such a process through clinical observation is remarkable. That such a process should exist at all would be difficult to explain unless some form of identity-level motivational architecture were already in play.

7.2 Conditions of Worth and the Formation of Distorted Identity Architecture

Rogers’s most diagnostically powerful concept is that of conditions of worth. Conditions of worth arise when acceptance, care, or belonging are made contingent on the individual meeting externally imposed standards. Over time, certain experiences, needs, emotions, or desires become incompatible with being a “viable self” and are therefore excluded from awareness (Rogers 1951, esp. chs. 7–9).

Crucially, this exclusion is not primarily cognitive or deliberative. It is motivational. The system learns which forms of experience are admissible and which are not, not through explicit belief but through affective reinforcement. Rogers emphasizes that individuals subject to conditions of worth are typically sincere, morally motivated, and reflective. What is distorted is not their reasoning, but the architecture that determines what enters reasoning at all.

In the present framework, conditions of worth correspond to the formation of misintegrated SPI. Identity-level motivational structures become organized around compensatory demands—approval, status, safety, control—rather than around organismic integration. These structures then exert teleological pressure on action across time, while remaining largely inaccessible to reflective governance.

This explains why agents governed by conditions of worth can exhibit long-term coherence without fulfillment. Their lives are organized, but the organization is costly. Reflective narratives develop to rationalize the trajectory, but these narratives misidentify its source. The resulting pattern is indistinguishable, at the surface level, from principled commitment. Architecturally, however, it is a form of identity-level distortion.

Rogers's analysis anticipates, with striking precision, the failure modes described in Section 6. Misalignment, identity drag, and runaway SPI are not anomalies. They are predictable consequences of identity-level motivational architecture formed under conditions of conditional acceptance.

7.3 Incongruence and the Phenomenology of Identity Drag

Rogers's notion of incongruence further clarifies the phenomenology of misalignment. Incongruence does not typically present as explicit inner conflict. Individuals may feel anxious, depleted, rigid, or vaguely dissatisfied without being able to identify a clear source of tension. They may function at a high level while experiencing their agency as effortful or hollow.

This phenomenology aligns closely with identity drag. When reflective governance is forced to compensate for misintegrated SPI—without representational access to the source of motivational pressure—regulatory resources are consumed continuously. The result is not akrasia but exhaustion. The system works, but at a cost.

Rogers's insistence that incongruence can persist in the absence of conscious conflict is philosophically important. It undermines the assumption, shared by many action theories, that reflective endorsement tracks integration. An agent may sincerely endorse their life, values, and commitments while remaining structurally misaligned. What matters is not endorsement, but architecture.

7.4 Integration Without Reflective Sovereignty

The most radical aspect of Rogers's theory—and the one that directly addresses the residual question left by earlier sections—is his account of integration. Rogers claims that when certain environmental conditions are present—most notably empathic understanding and unconditional positive regard—the individual's motivational system reorganizes itself toward greater congruence.

This reorganization is not achieved through reflection, planning, or value revision. It occurs because reflective governance ceases to interfere with identity-level reorganization. As experiences previously excluded by conditions of worth become admissible, SPI realigns. The organismic valuing process is no longer deflected, and agency regains coherence.

From an architectural standpoint, this is not a triumph of reflection but a relinquishment of its supervisory pretensions. Reflection does not heal the system; it steps aside. Integration is achieved not by better control, but by the removal of constraints that prevented identity-level motivational structures from reorganizing themselves.

This is the sense in which Rogers provides an existence proof. He shows that agency can recover unity and generativity without reflective command. If reflective sovereignty were necessary for integration, Rogers's clinical observations would be inexplicable. That they are not miraculous, but systematic and repeatable, strongly suggests that identity-level motivational architecture plays the organizing role his theory presupposes.

7.5 Philosophical Payoff

Rogers's work does not compete with the layered model; it corroborates it. Without SPI—or something functionally equivalent—Rogers's central claims would amount to optimism unsupported by mechanism. With SPI in view, they become intelligible.

The philosophical payoff is twofold. First, Rogers explains how integration is possible in agents whose reflective authority has been compromised. Second, he demonstrates that agency need not collapse once reflection loses its supervisory role. Human agency is not saved by better self-command, but by architectural realignment at a level beneath command.

In this respect, Rogers completes the arc of the paper. Bratman shows how reflection can organize action when it works. Section 6 shows how reflection can fail without agency disappearing. Rogers shows how agency can re-integrate without reflection ruling.

What emerges is a conception of agency that is neither anarchic nor authoritarian. Reflection matters—but it is not sovereign. Identity-level motivational architecture does the real work of organizing lives. Reflection interprets, regulates, and sometimes obstructs that work. When it learns to stop obstructing, agency can heal itself.

8. Conclusion

This paper has argued that much of contemporary philosophy of action rests on a structural illusion: that the principal engines of agency are transparent to reflection. Davidson locates agency in avowable reasons, Mele in deliberative conflict and control, Smith in idealized endorsement, and Bratman in reflectively accessible plans and future-directed intentions. These frameworks disagree about the content and norms of agency, but converge on a shared architectural picture: if agency is genuinely one's own, it must ultimately be governed from within the reflective standpoint.

The three-layer model developed here challenges that assumption without discarding the insights that motivated it. Layer 1 captures the domain these theories describe best: reflective governance, where agents form intentions, deliberate about reasons, narrate their lives, and assess their own commitments. Layer 3 captures the arational–procedural substrate: automatic, reactive, and biologically mediated processes that fall outside the space of reasons. Between them lies a structurally distinct domain, Subconscious Practical Identity (SPI): stable, identity-like motivational architecture that organizes action teleologically across time while remaining largely inaccessible to reflection.

Introducing SPI explains a phenomenon that reflection-centered models leave obscure: long-term, coherent, purposive life trajectories whose organizing motivations are not the ones agents can articulate, endorse, or revise. Agents often live disciplined, intelligible, and apparently value-driven lives while systematically misidentifying what is steering them. Their reflective explanations are not cynical cover stories; they are sincere rationalizations generated by a Layer-1 interface that lacks representational access to the Layer-2 structures doing the real organizational work.

Once SPI is made visible, several puzzles fall into place. First, the familiar contrast between “impulsive” and “rational” action proves too crude. SPI is neither mere impulse nor explicit plan. It is cross-temporally stable, emotionally structured, and functionally intention-like without being reflectively authored. Second, failures of agency cannot all be traced to defective deliberation, insufficient willpower, or irrational belief. Misalignment between Layer 1 and Layer 2 generates identity drag: chronic, non-akratic effortfulness in which reflection works overtime to stabilize trajectories for whose underlying direction it is not responsible. Third, classical theories reveal their own scope conditions. Davidson, Mele, and Smith illuminate different patterns at the reflective surface; Hursthouse clarifies what lies below that surface;

Bratman offers the most sophisticated account of long-term reflective organization. None, however, explains identity-level motivation operating beneath reflection while still exhibiting purposive, life-structuring force.

The Bratman pressure point is especially revealing. SPI reproduces many of the functional roles Bratman attributes to planning states: cross-temporal stability, constraint on deliberation, and coordination of action over time. Yet SPI systematically violates his conditions on intention: it is not reflectively accessible, not straightforwardly endorsable, and not encoded in propositional form. We are left with a choice. Either we restrict "intention" to reflectively available planning states and concede that large regions of long-term human agency fall outside planning theory's remit; or we broaden "intention" to include identity-level architecture and thereby erode what was distinctive about planning in the first place. The present proposal is to respect the local success of Bratman's account while treating it as exactly that: a local success, situated within a broader architecture whose deepest load-bearing structures are not plans at all.

The analysis of failure modes in situated agency reinforces this reorientation. Epistemic capture shows that reflective governance can be distorted from without, as social and communicative environments progressively undermine an agent's capacity to interpret their own motivational landscape. Intra-agent epistemic capture shows that a similar annexation can occur from within, when runaway SPI co-opts reflective interpretation and converts it into a narrative instrument for preserving identity coherence. Parkinsonism, by contrast, isolates a non-epistemic failure mode in which reflective intentions and identity-level motivation remain intact while motor gating collapses. Together, these cases demonstrate that agency can fragment along distinct architectural fault lines—epistemic, identity-level, and control-mechanical—that do not reduce to standard categories of irrationality, akrasia, or weakness of will.

Carl Rogers's work provides an unexpected but powerful form of corroboration. Long before layered models of agency or contemporary debates about reflective authority, Rogers posited an organismic valuing process that operates beneath reflection and tends toward greater integration, and he traced the effects of "conditions of worth" on the formation of distorted identity-level structures. Read through the present architecture, his clinical observations amount to an existence proof: identity-level motivational organization is real; it can be misintegrated or integrated; and deep realignment is possible even when reflective sovereignty has been compromised or abandoned. Reflection does not repair the system by exerting stronger command; it helps by ceasing to interfere with identity-level reorganization. On that picture, agency is restored not by a more authoritative supervisory standpoint, but by architectural reconvergence between layers.

The ambition of this paper has been architectural rather than metaphysical. It has not attempted to locate the metaphysical subject of thought, resolve questions about personal identity over time, or adjudicate between competing ontologies of persons. Its claim is instead that any adequate theory of action must acknowledge three structurally distinct domains of control, must allow that identity-level motivational architecture can organize lives without being reflectively authored, and must recognize that reflective governance is a vulnerable, partial, and revisable interface within that broader system.

If this is right, then the explanatory task for philosophy of action shifts. The central question is no longer how a transparent, supervisory reflection governs agency, but how a layered, partially opaque architecture sustains, distorts, and sometimes recovers coherent lives. Classical theories retain their importance as local maps of the reflective surface. A complete cartography of agency, however, must include the submerged terrains of Subconscious Practical Identity and the fragile control interfaces through which reflection sometimes manages— and often fails—to keep up with what we are already, and have long been, doing.

References

Abramson, Kate. 2014. “Turning Up the Lights on Gaslighting.” *Philosophical Perspectives* 28 (1): 1–30.

Bratman, Michael E. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Davidson, Donald. 1971. “Agency.” In *Agent, Action, and Reason*, edited by Robert Binkley, Richard Bronaugh, and Austin Gill, 3–25. Toronto: University of Toronto Press.

Doris, John M. 2002. *Lack of Character: Personality and Moral Behavior.* Cambridge: Cambridge University Press.

Doris, John M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

Hursthouse, Rosalind. 1991. “Arational Actions.” *The Journal of Philosophy* 88 (2): 57–68.

Jankovic, Joseph. 2008. “Parkinson’s Disease: Clinical Features and Diagnosis.” *Journal of Neurology, Neurosurgery & Psychiatry* 79 (4): 368–376.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision Under Risk.” *Econometrica* 47 (2): 263–291.

Mele, Alfred R. 1987. *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York: Oxford University Press.

Mink, Jonathan W. 1996. “The Basal Ganglia: Focused Selection and Inhibition of Competing Motor Programs.” *Progress in Neurobiology* 50 (4): 381–425.

Nieuwboer, Alice, Wim De Weerdt, Simon Dom, Marc Lesaffre, and Eric Vandenberghe. 2007. “A Frequency and Correlates of Freezing of Gait in Parkinson Disease.” *Neurology* 68 (2): 146–151.

Obdan, Paris B. 2025. “Meaning as a Weapon: A Pragmatic Analysis of Narcissistic Communication.” PhilArchive preprint.

Obdan, Paris B. 2026a. “Fragility of Reflection: Agency Without Supervisory Authority.” PhilArchive preprint.

Obdan, Paris B. 2026b. Reintegrated Agency: Self-Governance Without Transparency. PhilArchive preprint.

Redgrave, Peter, Tony J. Prescott, and Kevin Gurney. 1999. “The Basal Ganglia: A Vertebrate Solution to the Selection Problem?” *Neuroscience* 89 (4): 1009–1023.

Rogers, Carl. 1951. *Client-Centered Therapy: Its Current Practice, Implications, and Theory*. Boston: Houghton Mifflin.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.

Sripada, Chandra Sekhar. 2016. “Self-Expression: A Deep Self Theory of Moral Responsibility.” *Philosophical and Phenomenological Research* 93 (2): 393–418.