Relegated Structure:
Practice, Attention, and the Dual-Process Partition

Flyxion
Independent Researcher

November 2026

Abstract

The distinction between fast, automatic, intuitive cognition and slow, effortful, deliberative cognition is among the most widely used ideas in psychology, behavioral economics, and current commentary on artificial intelligence. This monograph argues that the distinction is a useful description of how performance varies and a poor description of how minds, or machines, are built. It develops Aspect Relegation Theory, on which learned automatic performance that was acquired through explicit coordination is the residue of structure that was once explicit and has been compressed out of attention through practice, continues to run while its distinctions go untracked, and can be promoted back into view, selectively, when conditions demand. It supplements that account with a cost model of amortization, on which the apparent difference between easy and hard tasks reflects how much deliberative work has already been paid for and reused. It then separates a limited thesis, which holds that speed, effort, and reportability do not identify cognitive kinds, from a stronger thesis, which holds that all fast cognition is relegated deliberation, and defends only the first in full. It shows formally that difficulty is a relation between a task and a practice history, so that a partition of tasks by cost can be reversed without any change in the tasks, and that the pattern of fast and slow performance can be produced by routing among stored and freshly computed results without two systems. Hardwired automaticity is acknowledged as a genuine second source, and so are inherited compression, in which a system exhibits fluent output without a trajectory of its own, and conditioning, in which the exposure history is authored by someone else. The remaining problem is the regulation of resolution: how a system detects that its current level of detail is inadequate and what it does next. The monograph applies this framework to a case of model transfer, in which a taxonomy of performance profiles migrates into discussion of language models together with an interpretation that the taxonomy's structure does not license, and it proposes a behavioral test, based on how task performance varies with inference-time computation, together with a controlled-mismatch experiment, that replaces the question of whether a system has a second faculty with questions that can be investigated.

Contents

1 Introduction **4**

2 From Operational Contrasts to Architecture **7**
2.1 Antecedents 7
2.2 Ontological drift 8
2.3 Single-process and dual-process positions 8
2.4 The working-memory criterion 9
2.5 What a partition claim asserts 10

3 Aspect Relegation **12**
3.1 The core definition 12
3.2 Learning as relegation 13
3.3 Reversibility 13
3.4 What persists and what returns 14
3.5 Intuition and effort 14
3.6 Graded and local automaticity 15
3.7 A formal sketch of relegation and promotion 15
3.8 Ease is not simplicity 17

4 Difficulty Is a Relation **18**
4.1 Setting 18
4.2 Reversal 19
4.3 Practice curves and the absence of a threshold 20
4.4 A memoization model 21

5 Amortized Deliberation **23**
5.1 Reuse and the cost of relegation 23
5.2 A cost model 23
5.3 Depth, missed failures, and false alarms 25
5.4 Reuse and the preservation of verification conditions 26

5.5 A cached architecture as an illustration 26

6 Routing and the Appearance of Two Systems 28

6.1 The routing argument 28

6.2 Why the partition looks real 29

7 Routes to Automaticity 31

7.1 The objection from innate automaticity 31

7.2 Where the weight falls 32

7.3 Compression and its dependence on exploration 32

7.4 Inherited compression 33

7.5 Conditioning 34

7.6 Three routes stated on common ground 34

7.7 Relegation as one route among several 35

8 Resolution and Its Regulation 37

8.1 The reframed question 37

8.2 Selective return to scrutiny 37

8.3 A human study 38

8.4 The switch problem 39

8.5 A rational counterpoint 40

9 Model Transfer and Borrowed Interpretations 41

9.1 The notion of model transfer 41

9.2 Taxonomies as interpretive shortcuts 42

9.3 Systems vocabulary and language models 42

9.4 Intelligence as a borrowed concept 43

10 Inference-Time Computation Reconsidered 44

10.1 What the practice is 44

10.2 A behavioral test 45

10.3 A discriminating experiment 45

10.4 Matching requirements 46

10.5 Evidential roles of the human and artificial studies 47

10.6 Practice that teaches a procedure 47

10.7 Cautions 48

10.8 Deficits relocated 49

11 Objections and Limits **50**
11.1 The objection from reflexes 50
11.2 The objection from metacognition 50
11.3 Evidential status 50
11.4 What would count against the account 51

12 Conclusion **52**
12.1 What remains of the distinction 52
12.2 Summary of the argument 53

Chapter 1

Introduction

Almost everyone who has read about human judgment in the last two decades knows the picture. One system is fast, automatic, associative, and prone to bias. The other is slow, effortful, rule-governed, and capable of correcting the first. The picture is intuitive because it names a contrast that everyone has experienced: the difference between recognizing a friend's face and working out a long division, between steering a familiar route and navigating an unfamiliar city with a paper map. It is also portable. It has been carried from experimental psychology into economics, medicine, education, public policy, management advice, and, more recently, into discussion of what large language models can and cannot do. In each setting it supplies a short vocabulary for explaining why a judgment went wrong (the fast system answered when the slow system should have) and for specifying a remedy (engage the slow system).

The portability is also the problem. A vocabulary that fits every case can come to explain none of them, and a description of how performance varies can quietly harden into a claim about what kinds of things a mind contains. The thesis of this monograph is that the second step is unwarranted, and that the history of the framework, the structure of the evidence, and the way the framework is now used to assess artificial systems all show the cost of taking it. What the framework correctly describes is a family of contrasts in how performance is carried out: the same competence can run with little attention or with a great deal, with little effort or with much, and with little ability to report its steps or with a detailed ability to do so. What it incorrectly suggests is that these contrasts mark the boundary between two cognitive engines, each with its own character, its own failure modes, and its own moral valence.

The alternative developed here has two parts that are best kept distinct. The first is Aspect Relegation Theory. It holds that when a task is first encountered, many of its constraints, distinctions, and checks have to be held explicitly and coordinated by attention. With repetition, strategies stabilize, errors decline, and the distinctions no longer need supervision; the system economizes by letting them run without tracking them. Intermediate representations, conditional branches, evaluative criteria, and error checks are relegated, which means that they continue to

operate while no longer being attended. Relegation is reversible. When expectations fail, errors accumulate, or the environment becomes unpredictable, relegated structure is promoted back into view. The transition that the dual-process vocabulary describes as a shift from one system to the other is, on this account, a shift in the resolution at which a capacity is being exercised.

The second part is a claim about cost. The fast and slow performances that the framework contrasts often differ in how much deliberative work has already been done and how many times its results have been reused. A task that feels hard is frequently a task whose costs have not been amortized, and a task that feels easy is frequently one whose costs have been spread over many prior occasions. The relevant difference is then one of practice history, and practice history is a property that the same agent can have with respect to one task and lack with respect to another. This explains why the same person can be fluent with one keypad layout, one microwave, or one style of door handle and clumsy with another that differs only slightly, without any change in the person's cognitive architecture. It also explains why the contrast between fast and slow cannot be used to sort tasks, or the cognitive operations required by them, into two kinds.

The empirical weight of the account rests on a claim about selective return to scrutiny. Practice-dependent speed is compatible with several accounts of cognitive architecture, including accounts on which practice recruits a different mechanism, so it cannot carry the argument. A more discriminating prediction is that when a practiced routine breaks down, the distinctions that return to explicit control are those relevant to the breakdown, and not the whole task at once. Chapter 8 states this claim, Chapter 10 turns it into an experiment for artificial systems, and the human examples and the cost model are organized around the question it answers: how a performer detects which part of a routine needs renewed attention.

The monograph separates a limited thesis from a stronger one, and defends the limited thesis in full. The limited thesis holds that speed, effort, and reportability do not identify cognitive kinds: that a process's being fast, or effortless, or hard to report does not show that it is a different sort of process from one that is slow, effortful, and reportable. The stronger thesis holds that all fast cognition is relegated deliberation, that every automatic process began as an explicit one. The stronger thesis is false as stated, because some automaticity is hardwired, and the framework developed here acknowledges that without embarrassment. What it claims is that the competencies that matter for debates about rationality, bias, expertise, and artificial intelligence, such as route choice, linguistic interpretation, social inference, probabilistic expectation, and strategic choice, are overwhelmingly learned, and that for these the relegation account is the appropriate one. The distinction between hardwired and relegated automaticity is taken up in Chapter 7, and the evidential status of the claims at each level is taken up in Chapter 11.

A further aim is to identify the problem that remains once the partition has been set aside. If System 1 is relegated structure and System 2 is the same capacity exercised at higher

resolution, then the interesting question is not whether a system has a second faculty but how it decides that its current resolution is inadequate and what resources it then deploys. Humans have rich triggers for this decision: surprise, error, conflict, novelty, affective signals, and social cues. Artificial systems typically rely on externally imposed criteria or narrow objectives. The difference is functional and can be stated without categorical exclusion of one kind of system by the other. Chapter 8 develops the question, and Chapter 10 shows how it bears on the current practice of giving language models extra computation at inference time.

Finally, the monograph treats the migration of the dual-process vocabulary into discussion of artificial systems as an instance of a wider phenomenon in the philosophy of science, the carrying of a model, or of a taxonomy, from one domain to another together with an interpretation that the structure of the model or taxonomy does not license. Chapter 9 uses a discussion by Mel Andrews of model transfer, hermeneutic shortcuts, and the interpretive baggage that comes with imported formalisms as a worked case. The point of the case is not that the dual-process vocabulary has no legitimate use; it is that its use in a new setting requires fresh evidence for the interpretive claims it carries, and that the labels attached to current systems, such as reasoning models and expert-level performance, borrow standing that the benchmarks cannot supply.

A note on scope and method. The account offered here is a functional one. It makes no commitment to a neural architecture or computational formalism. Its central notions, relegation, resolution, and re-promotion, are introduced in Chapter 3 in a form that supports a bounded-attention update rule, and three results are stated formally so that the claims can be examined: a reversal result showing that cost-based classifications of tasks depend on practice history (Chapter 4), a cost model of amortization (Chapter 5), a routing argument showing that the appearance of two systems can arise from selection among stored and computed results (Chapter 6). A fourth claim, that relegation is one route among several to automaticity, is a statement of scope supported by the literature and not a formal result (Chapter 7). The formal claims rest on a smaller base of citations than the philosophical ones. The formal treatment is a proposal and not an empirical result; it is offered so that the claims can be examined and, where the evidence requires, refuted.

A roadmap may help. Chapters 2 and 3 are conceptual and set out the positions and the account of relegation. Chapters 4, 5, and 6 are formal, and state the reversal result, the cost models, and the routing argument. Chapter 7 separates routes to automaticity and fixes the scope of the claims. Chapter 8 and the experimental sections of Chapter 10 are empirical proposals, one for humans and one for artificial systems. Chapter 9 and the two illustrations in Chapters 5 and 10 are case studies, and Chapter 11 and the conclusion assess the evidential standing of the whole.

Chapter 2

From Operational Contrasts to Architecture

2.1 Antecedents

The contrast between habitual and effortful mental life long predates the vocabulary of Systems. William James described habit as the great conserver of effort, a matter of sequences of actions becoming automatic with repetition so that attention is freed for other things, and he contrasted this with deliberate reasoning as a matter of subjective experience and not as a division into separate mechanisms [11]. The distinction was never at first an architectural one. It marked a difference in how mental life presents itself, and it was explained, where it was explained at all, by what repetition does to a process.

The mid-twentieth-century experimental literature gave the contrast operational definitions. Processes were called automatic when they were fast, placed little load on limited resources, and were difficult to interrupt or control consciously, and controlled when they were slow, resource-demanding, and dependent on attention. Posner and Snyder's treatment of attention and cognitive control is a standard reference for this operationalization [20], and Bargh's work on the conditions under which automatic processing effects appear showed how the features that define automaticity can come apart and be separately manipulated [3]. These definitions describe how processes behave under experimental conditions. They do not say what the processes fundamentally are, and the experimental literature itself recorded that the defining features need not travel together.

Wason and Evans introduced the dual-process idea into the study of reasoning, proposing that performance on selection tasks reflects a process that is not the one participants report [28]. Over the following decades the reasoning literature developed a large number of dual-process and dual-system proposals. Evans and Stanovich later argued that the diverse proposals should be organized in terms of two types of process, distinguished at the core by whether a process

requires working memory, and that the many correlated attributes found in the literature should be treated as typical and not as defining [9]. That move was an explicit attempt to prevent exactly the reification that this monograph is concerned with, and it is a useful marker of where the framework's careful practitioners stand: the label "Type" was introduced because the label "System" invites the reading in which the mind contains two machines.

2.2 Ontological drift

The drift from operational contrast to architecture happened largely outside the careful literature. In the heuristics-and-biases program, the systematic errors in judgment catalogued by Tversky and Kahneman were explained by the use of heuristics [24]. Stanovich and West's work on individual differences in reasoning gave the two-system vocabulary a prominent role in discussions of rationality [23]. Kahneman's popular synthesis made the vocabulary part of ordinary educated speech [12], and in behavioral economics and popular psychology the labels became explanatory agents: biases were attributed to System 1, rational correction to System 2, and the mind came to be pictured as containing two qualitatively distinct engines of cognition.

Three consequences of the drift matter for what follows. The first is that it moralizes cognition. System 1 is cast as impulsive, biased, and unreliable, System 2 as rational, corrective, and effortful, so that effort becomes a sign of rationality and automaticity a sign of suspicion. Skilled expertise, which consists to a large degree of automatic and effortless performance, is then awkwardly placed. The second is that it encourages the explanation of failure by assignment to a system: an error is explained by saying that the fast system produced it, which adds nothing to the observation that a fast answer was given and was wrong. The third is that it supplies a ready criterion for excluding artificial systems from "true" reasoning, since if System 2 is a distinct faculty then a system can be said to lack it, and a system that produces fluent output without displayed intermediate steps can be said to have only the first. The consequence is taken up in Chapter 10, where the deficits attributed to a missing System 2 are relocated.

2.3 Single-process and dual-process positions

The literature itself contains an extended debate about whether the contrast marks a difference in kind or in degree. De Neys has argued that the debate has been unproductive: that there is currently no good evidence allowing the dual-process and single-process positions to be decided against each other, and that even a resolution would not advance understanding of the processing mechanisms that underlie thinking [5]. In a later survey he identified two foundational assumptions of fast-and-slow models as poorly supported, namely exclusivity, the assumption that some responses are only available to one mode of processing, and the switch

feature, the assumption that a reasoner can decide to shift between modes, and he argued that leading accounts of the switch are conceptually problematic because they presuppose exclusivity [6]. Dewey has examined the question of metacognitive control in single-process and dual-process theories along related lines [7].

These positions bear on the present argument in two ways. They show that skepticism about the ontological reading of the partition is not a lone provocation, since researchers within the field who have examined the evidence regard the debate over kinds as undecided. De Neys's position is neutral on whether there are two kinds. The present argument is narrower in one respect and different in another: it holds that the inference from a performance profile to a kind is unsupported, and it does not claim that the evidence shows there to be a single process. De Neys is cited for the state of the evidence and is not enlisted for a conclusion that he does not draw. And they show what is at stake in the question of the switch. If the mechanism by which a reasoner shifts from fast to slow processing presupposes the very exclusivity that the evidence does not support, then the switch needs to be reconceived, and the reconception developed in Chapter 8 treats it as a regulation of resolution in a capacity that is exercised at varying resolution. This monograph does not claim to settle the single-process and dual-process debate as a question of cognitive architecture; it claims that the particular inference from fast, effortless, hard-to-report performance to a second kind of process is not supported, and that a practice-based account explains the same contrasts without it.

2.4 The working-memory criterion

Evans and Stanovich's criterion deserves a direct answer. On their account, Type 2 processing requires working memory and Type 1 does not [9]. A dual-process theorist can readily accept that the same task becomes easier with practice, so the dispute concerns what would distinguish the proposed mechanisms. Practice could change a task's working-memory demands in at least three ways: the same operation could make smaller demands as its distinctions are compressed, which is relegation as defined in the next chapter; the task could come to recruit a different operation, such as retrieval of a stored answer in place of computation, which is substitution; or both pathways could occur, in different tasks or at different stages of one task. The relegation account is committed to the first pathway for the competencies to which it is applied and does not exclude the others.

The pathways differ in what they predict about learning history and about repair, though the differences are less clean than a first statement suggests. Retrieval can become more frequent gradually, and a reasoner who has retrieved a stored answer can recompute a single component when it fails, so gradual change and local repair do not by themselves separate compression from substitution. What separates them is the history of the routine. If practice compresses an operation, then distinctions that were explicitly coordinated during training should later

require less explicit tracking, and perturbing one of those distinctions after practice should bring that distinction back under explicit control. If practice substitutes retrieval, the explicit distinctions of the earlier computation need not be preserved in the routine, and a perturbation is handled by whatever fallback procedure the performer has, which may or may not correspond to them. Section 8.3 describes a study that uses this contrast. These predictions are offered as a way of stating what would separate the mechanisms, and their evidential status is that of hypotheses to be tested. The monograph does not cite skill-acquisition data that discriminate among the pathways, and it does not claim that such data are absent: existing work on skill acquisition may already bear on them, and a systematic review of that literature is needed before the predictions are treated as more than hypotheses.

2.5 What a partition claim asserts

The inference from differences in speed to differences in cognitive kind requires additional assumptions, and the vocabulary is used with at least three different commitments. On one interpretation, tasks themselves belong to history-independent classes, so that a thought is fast or slow whoever thinks it. On a second, the classes are mechanisms whose engagement depends on the agent and the situation, and a given thought is handled by one or the other according to circumstance. On a third, the vocabulary is merely descriptive of speed and effort and specifies no architecture. These interpretations must be kept apart, because evidence against an intrinsic partition of tasks does not by itself establish that only one mechanism operates. The popular form of the doctrine adds a further element, which is that the slow mechanism is the one associated with awareness, effort, and the self, and the fast mechanism is what remains once those are subtracted, and a dispatcher that assigns each thought to its mechanism is often posited. The two-systems reading has also been criticized as a conflation of several independent dimensions [14, 17].

The most careful current statement of the dual-process view separates the defining characteristics of the two types of processing from the attributes that merely tend to accompany them [9]. On that account Type 1 processing is defined by autonomy, meaning that it does not require working memory or controlled attention, and Type 2 processing is defined by its dependence on working memory and by cognitive decoupling, the capacity for hypothetical thinking. Speed, effort, high capacity, and association with consciousness are treated as correlates and not as definitions. The target of this monograph is narrower than the whole of that position. The reversal argument of Chapter 4 bears on classifications inferred from cost, since a ranking by speed or effort is a ranking by correlated attributes, and the argument shows that such rankings depend on history. It leaves the proposed defining characteristics, autonomy and the use of working memory for decoupling, as questions for further investigation, with two cautions. A response that begins as a decoupled, working-memory-dependent computation

can become autonomous through practice, so autonomy is itself history-dependent. And the defining characteristics are properties of processes, so attributing them to tasks independently of history would require a further argument.

Chapter 3

Aspect Relegation

3.1 The core definition

Aspect Relegation Theory begins from the observation that the same task can be carried out with very different amounts of explicit attention depending on the history of the person carrying it out. It proposes that this variation is explained by the compression of previously explicit structure, and it states the core idea in functional terms.

Definition 3.1 (Relegation)**.** *A process is* relegated *when it continues to operate reliably while its internal distinctions are no longer explicitly tracked. The distinctions that are relegated may include intermediate representations, conditional branches, evaluative criteria, and error checks.*

Several features of this definition are deliberate. It is stated in terms of what continues to operate and what ceases to be tracked, so that the process is not said to stop and to be replaced by a different one. It is mechanism-neutral: it describes a functional transformation that could be implemented on different substrates, and it makes no claim about neural architecture or about which formalism best models the process. And it ties automaticity to a history, since relegation is something that happens to structure that was previously explicit.

Definition 3.2 (Resolution)**.** *The* resolution *at which a task is exercised on an occasion is the grain of its structure that is represented and tracked explicitly and is available to control. Higher resolution means that more of the task's structure, including intermediate representations, conditional branches, evaluative criteria, and error checks, is explicitly tracked.*

Relegation is then a reduction of resolution for a routine whose operation continues, and promotion is an increase. Resolution is a property of how a task is being exercised and not a fixed property of the task or of the agent, and it can differ across the aspects of a single task, which is the sense in which automaticity is local.

3.2 Learning as relegation

When a task is novel, many constraints must be coordinated explicitly. A beginner at a musical instrument must attend to which finger plays the next note, how the hand is positioned, what the next phrase requires, and how the rhythm is to be kept, and these demands compete for a limited attentional budget. With repetition, strategies stabilize, errors decline, and distinctions that once needed supervision become reliable enough to be left to themselves. The system economizes by pushing these distinctions below the level at which they are attended. The practiced performer is then free to attend to higher-level features, such as the shaping of a phrase, that the novice could not afford to consider.

The account predicts that relegated structure can return to attention when conditions require it. If the fingering begins to fail at higher tempo, it becomes an object of deliberate attention again. If a familiar route is blocked, the commuter's relegated navigation structure is partly promoted, and distinctions among streets and turns that had ceased to be tracked become explicit again. What happens in such cases is not a change of cognitive faculty. The same operations that ran in the background move into the foreground, and attention widens its aperture to include finer-grained distinctions.

3.3 Reversibility

A relegated routine remains operational; what ceases is the explicit tracking of its distinctions, and the structure behind them is accessible in the sense that tracking can be resumed. This claim distinguishes the account from one on which practice simply replaces explicit reasoning with a stored association that has no further connection to the reasoning that produced it. On the relegation account, the routine continues to run and the distinctions that it embodies can be re-promoted to explicit control when expectations fail, errors accumulate, or the environment becomes unpredictable. The practiced performer who is asked to explain how a technique works can often make its steps explicit, and the practiced performer who meets an unfamiliar case can often unfold the compressed routine into its constituents and work out which of them no longer applies.

The reversibility claim has an evidential consequence that is taken up later. If relegation is reversible, then the transition from fast to slow performance should be predictable from the occurrence of conditions that signal a breakdown, such as surprise, error, conflict, or novelty, and not from a fixed property of the task. It also has a conceptual consequence: the transition is one of resolution. Resolution, as defined above, is what varies between fast and slow performance, and the variation is in how much of the structure is held explicitly, not in which kind of process is running.

3.4 What persists and what returns

Three things need to be kept apart if the account is to be testable: prior explicitness, subsequent automatic execution, and later access. Prior explicitness is the claim that the distinctions now embedded in a routine were at some stage held and coordinated explicitly. Subsequent automatic execution is the claim that the routine runs reliably without such tracking. Later access is the claim about what returns to attention when the routine is scrutinized. The third admits at least three readings. The performer may recover a previously explicit rule, as when a pianist retrieves a fingering that was once consciously worked out. The performer may reconstruct a workable account of what the hand is doing, which was never explicit in that form. Or the performer may devise a new procedure for the difficult passage. These readings make different claims about continuity between the earlier explicit structure and what is later accessed, and the relegation account as defined here is committed to continuity only in the first. For the second and third, what it claims is the weaker point that the routine's operation can be brought under explicit control in a way that is sensitive to what has gone wrong. The empirical claims of Chapter 8 concern the extent of what returns to control and not its identity, so they do not depend on which reading holds.

The distinction also bears on scope. The fact that a competency was learned does not establish that its constituent operations were ever explicit, since learning can proceed by exposure without an explicit phase. Relegation as defined applies where the history includes prior explicit coordination, as in instruction and deliberate practice, and the claim that practice history explains ease and effort is broader than the claim that compression explains it.

3.5 Intuition and effort

On this account intuitive judgments feel immediate because their supporting structure has been compressed out of view. Intuition is a derivative source of judgment, and it depends on earlier structure that was explicit at some stage of learning. This does not make intuition unreliable. A relegated routine that has been exercised under stable conditions, with feedback, over many occasions can be highly reliable, and it can be more reliable than explicit reasoning on the same task, since explicit reasoning is subject to errors of its own. The account removes the presumption of unreliability that the dual-process vocabulary attaches to the fast system and replaces it with a question about the conditions under which the compression was achieved: whether the environment was stable, whether feedback was available, and whether the cases now being met resemble those on which the structure was formed.

Effort, on this account, is a signal that representational detail is being actively maintained. It is not itself a mark of rationality. Effort falls as relegation proceeds because the task has been reorganized, not because it has been simplified, and a person who finds a task effortless after

long practice is performing the same task at lower attentional cost. The moral valence that the dual-process vocabulary attaches to effort, as the mark of the careful and the conscientious, is therefore unwarranted as an account of cognition. An effortful judgment may be a sign of inexperience, and an effortless one a sign of mastery.

3.6 Graded and local automaticity

Automaticity on this account has no sharp boundary, and different aspects of a task can be relegated at different rates and to different depths. The experienced driver handles routine traffic with little attentional oversight, but remains alert to construction zones, to unusual noises from the vehicle, and to changes in road conditions, so that some aspects of driving are deeply relegated and others are held at a higher resolution because their stability has not been established or because their cost of failure is high. This local character matters for the contrast with dual-process accounts. A dual-process account has to say which system is operating when the driver is both coasting through traffic and watching for a hazard, and the natural answer, that both are operating at once, pushes the account toward the claim that the two are not exclusive. On the relegation account no special answer is needed: different aspects of one task are held at different resolutions at the same time.

3.7 A formal sketch of relegation and promotion

The informal account can be given a minimal formal statement in terms of the attention that a task receives. Let the structure of a task be a finite set $X = \{a_1, \ldots, a_n\}$ of aspects, each corresponding to a constraint, sub-operation, or dependency relevant to successful execution. Let $\mathcal{K}(u) \subseteq X$ be the known aspects, those that the agent has represented at time u whether or not they are currently monitored, and let the active attention set $\mathcal{A}(u) \subseteq \mathcal{K}(u)$ be the aspects under explicit monitoring, subject to a bounded capacity $|\mathcal{A}(u)| \leq k$. The formulation separates three things that are easily conflated: the full task structure, what the agent has come to represent about it, and what the agent is currently attending to. The separation matters because a novice's characteristic difficulty is that the task has more aspects than can be attended to at once, so that the full task cannot simply be placed in the attention set at the start.

Definition 3.3 (Stabilization)**.** *An aspect a_i is* stabilized to level $1 - \eta$ under policy σ and environment class $\mathcal{E}$ *if*

$$\Pr(\textit{constraint satisfied} \mid a_i, \sigma, \mathcal{E}) \;\geq\; 1 - \eta.$$

Stabilization is thus probabilistic and relative to a context, and not a binary property. This

is what allows a stabilized aspect to fail outside the environment class in which it was learned, and it is why automaticity transfers unevenly across contexts: a routine that is reliable under $\mathcal{E}$ carries no guarantee under a different class $\mathcal{E}'$.

Definition 3.4 (Attention update rule)**.** *The attention set evolves by*

$$\mathcal{A}(u+1) \;=\; (\mathcal{A}(u) \setminus R_u) \cup N_u \cup Q_u,$$

where $R_u \subseteq \mathcal{A}(u)$ is the set of aspects that have met their stabilization threshold and are relegated out of active attention, $N_u \subseteq \mathcal{K}(u) \setminus \mathcal{A}(u)$ is a set of newly attended higher-order aspects that become tractable once capacity has been freed by R_u, and Q_u is a set of aspects, possibly previously relegated, that are reactivated by uncertainty, novelty, an unfamiliar context, or a detected error.

Learning is not monotonic in this model. As lower-level operations stabilize, new higher-order aspects can enter the attention set through N_u, and previously stabilized routines can be recruited back through Q_u when they fail outside their training distribution or when the agent detects an anomaly. The attention set therefore need not shrink to nothing. Skilled performance typically retains some attention at some level, such as goal selection, anomaly monitoring, strategic coordination, and the use of feedback, and is better modeled as a change in what is monitored than as the disappearance of monitoring. In terms of the definition of resolution above, the resolution at which a task is exercised on an occasion corresponds to the extent of $\mathcal{A}(u)$ within $\mathcal{K}(u)$ and X, relegation is the removal of aspects through R_u, and promotion is their return through Q_u.

Remark 3.1 (Bounded capacity and selective promotion). Because capacity is bounded, a task with more than k aspects cannot have all of its relegated aspects promoted at once as an attentional state. This gives a reason, internal to the model, to expect that promotion triggered by a breakdown is selective in extent, and so it bears on the prediction developed in Chapter 8. The bound is an assumption of the model, and the prediction remains a hypothesis to be tested.

Two further consequences follow. The first concerns depth of skill. Freeing capacity through R_u makes room for higher-order aspects to enter through N_u, namely coordination, abstraction, and integration across sub-tasks that were not representable while lower-level aspects still required monitoring. The result is a hierarchical attentional structure in which lower-level operations become largely automatic, mid-level structures are intermittently attended, and higher-level planning remains actively monitored, with Q_u able to reactivate any level when context demands it. Skill is then not merely efficiency but depth of compositional control, the capacity to operate at a higher level of abstraction because lower levels have been stabilized, though not irrevocably.

The second concerns reconstructability. Relegation introduces a partial loss of it. Once aspects have left the attention set, intermediate steps may no longer be explicitly representable,

explanations may become compressed or incomplete, and reconstructing the full history of the skill may require deliberate re-expansion. Expert knowledge is often operationally accessible but descriptively compressed, so that the expert can perform a task while struggling to narrate how. This is a structural feature and not a failure, and it explains why expert advice is so often correct for the expert and useless for the novice: the relegated aspects are no longer visible to the advisor, so the advice skips steps that have not yet been stabilized for the listener. It also bears on the distinction drawn earlier among the readings of later access, since what returns to attention may be a reconstruction and not a recovery.

3.8 Ease is not simplicity

That a task feels easy to an expert is evidence that the expert has performed enough exploration to stabilize its recurring structure. It is not evidence that the task is simple. The complexity has not disappeared. It has been internalized, and Q_u is the model's reminder that the internalization is conditional and not permanent. The divergence between expert and novice perception of a disordered system illustrates the point. An experienced builder surveying a derelict building perceives a tractable sequence of operations: assess structural integrity, identify constraint violations, decompose into sequenced sub-tasks, and estimate materials and labor. A novice surveying the same building perceives disorder whose resolution is not apparent. Both perceptions are accurate about the building's current state. What differs is not information about the building but the availability of a generative pathway toward a sound structure. The expert's known aspects already contain stabilized decompositions for this class of problem, so the perceived task is tractable; the novice's do not, and the system is, from the novice's position, genuinely difficult to find a path through, and appropriately so, given what has and has not yet been explored.

Chapter 4

Difficulty Is a Relation

This chapter states the first formal result of the monograph, a reversal argument showing that a partition of tasks into easy and hard cannot be read off from speed, and it develops the one-process account of practice against which the partition claim is to be measured. The claims in it are of two strengths. The formal results establish that a partition of tasks or of mechanisms cannot be inferred from speed alone. The further position, that no partition of mechanisms is needed, is argued from parsimony and from worked models, and it is not offered as a theorem.

4.1 Setting

Let $\mathcal{T}$ be a set of tasks and $\mathcal{H}$ a set of histories, where a history $h \in \mathcal{H}$ records everything an agent has previously done that bears on its performance. Let $C : \mathcal{T} \times \mathcal{H} \to \mathbb{R}_{>0}$ assign to each task and history the resource cost of producing an adequate response, measured in whatever units are convenient, such as time, error-corrected attempts, or arithmetic operations. A history is a partial record of past episodes, and for a task t the number of prior episodes in h that are relevant to t is written $n_h(t)$.

Definition 4.1 (History-independent cost separation)**.** *A set $E \subseteq \mathcal{T}$ is* cost-separating at h *if $C(t, h) < C(t', h)$ for all $t \in E$ and $t' \notin E$. It is* intrinsic *if it is cost-separating at every $h \in \mathcal{H}$.*

A partition of tasks into fast and slow that is read off from observed speed would need to be a set of this kind, since otherwise the grouping changes when the agent's history changes. A theory might instead classify processes, or classify tasks relative to an agent and a history, and the proposition below does not touch such theories directly. Its target is the inference from observed speed to a partition of tasks.

4.2 Reversal

Definition 4.2 (Reversal pair)**.** *Tasks t, t' form a* reversal pair *if there are histories h, h' with $C(t,h) < C(t',h)$ and $C(t,h') > C(t',h')$.*

Proposition 4.1 (No intrinsic partition across a reversal)**.** *If E contains exactly one member of a reversal pair, then E is not intrinsic.*

Proof. Say $t \in E$ and $t' \notin E$. At h' we have $C(t,h') > C(t',h')$, which violates the requirement that every member of E be cheaper than every non-member. Hence E is not cost-separating at h'. The argument is symmetric if $t' \in E$ and $t \notin E$, using h. □

The proposition does not say that every set is non-intrinsic. A set containing both members of every reversal pair, or neither, escapes it, and a population whose histories are drawn from a common practice distribution may contain a set that is cost-separating for every history in that population. Such a set would be a fact about the population's training, and it could fail when the distribution of histories changed. Nor does the proposition rule out two mechanisms whose relative engagement changes with practice. What it establishes is narrower: no history-independent cost ordering separates the members of a reversal pair, so a partition of tasks drawn from observed speed cannot be a fact about the tasks alone. The step from there to a conclusion about mechanisms is not made by the proposition. It requires the further argument of this chapter and the next, which shows that distinct strategies, or a single process with a store, can produce the same smooth performance curves.

Example 4.1 (The keypad)**.** A telephone arranges its digits with 1 2 3 in the top row, and a calculator arranges them with 7 8 9 in the top row. Let t be entering a fixed four-digit number on the telephone layout and t' be entering the same number on the calculator layout. Let h be the history of someone who has used telephones for years and calculators rarely, and h' the converse. Under these stipulated histories, which are plausible and not measured, $C(t,h) < C(t',h)$ and $C(t,h') > C(t',h')$. The task is identical in every respect that bears on its content, namely the same digits in the same order, and the ordering of cost flips with the history. Any set E that contains one of the two entries and not the other is therefore history-dependent by Proposition 4.1.

The same structure appears with a door that pushes or pulls, an appliance whose controls are laid out differently from the one previously used, or a language whose word order differs from the speaker's own. In each case the cost of the task is a joint property of the task and the practice behind it. The reversal argument applies wherever practice is task-specific. Where two tasks differ mainly in how much they have been rehearsed, and not in irreducible demands or attainable performance, their cost order can be exchanged by exchanging the rehearsal, and not every pair of tasks is of this kind.

4.3 Practice curves and the absence of a threshold

The standard empirical regularity is the power law of practice [19], according to which cost after n relevant episodes is well described by

$$c(n) = a + b\, n^{-\beta}, \qquad a \geq 0,\ b > 0,\ \beta > 0. \tag{4.1}$$

Other forms, such as exponentials, fit some data better, and it has been argued that a power law in averaged curves can arise from averaging over participants and items whose individual curves are exponential, so that the aggregate form does not identify the form of individual learning [10]. The curve is therefore a stipulated illustration and not a universal empirical law, and nothing below depends on the particular form. What matters is that c is continuous, strictly decreasing, and without a distinguished point.

Proposition 4.2 (No privileged threshold)**.** *For any threshold $\tau > a$, the number of episodes at which the stipulated cost first falls to τ is $n^*(\tau) = \big(b/(\tau - a)\big)^{1/\beta}$. The map $\tau \mapsto n^*(\tau)$ is a continuous bijection from (a, ∞) onto $(0, \infty)$, and the derivative $c'(n) = -b\beta n^{-\beta-1}$ is continuous and nonvanishing on $(0, \infty)$.*

Proof. Solve $a + b\, n^{-\beta} = \tau$ for n. The remaining claims are direct computations. □

The bijection holds for a continuous experience variable on $(0, \infty)$. If n counts actual episodes, the first episode count at which the cost is at or below τ is $\max\{1, \lceil n^*(\tau) \rceil\}$, and the conclusion that no threshold is distinguished is unaffected.

The argument fails if the empirical cost curve has a discontinuity or a change of regime at a putative switch point, which is a testable prediction of any account on which a threshold marks a change of mechanism.

No value of τ corresponds to a distinguished point of the curve, since c' is continuous and nonvanishing, and the choice of τ belongs to the observer and not to c. Any statement of the form that the task became automatic after n^* repetitions is therefore a statement about the threshold chosen by the observer, since different choices move n^* over the entire positive half-line, and the stipulated curve itself contains no step. This bears on a particular family of hypotheses, namely those that posit a discontinuous and behaviorally observable switch from one mechanism to another, which would predict a gap in the cost or practice axis. It does not bear on all two-mechanism accounts. Two mechanisms whose probabilities of engagement change gradually, or which operate concurrently with changing relative contributions, also produce continuous performance curves, and a single mechanism can show abrupt transitions. Continuity therefore does not exclude two mechanisms. What it shows is that crossing a threshold is not evidence of a switch.

4.4 A memoization model

The mechanism behind the curve can be made explicit with a memoization model, in the spirit of the instance theory of automatization, in which automaticity is the retrieval of stored episodes [16]. A complementary account describes knowledge compilation, in which proceduralization and composition convert knowledge that is executed interpretively into productions that run directly, so that practice can change a response by compiling a procedure and not only by storing an instance [1]. Neither account, taken alone, establishes that cognition uses one mechanism. The model has to be executable and not merely an envelope of attainable costs, because an agent cannot take the cheaper of two routes without first paying to find out which is cheaper.

Let M be a store of solved instances. Let $S(t)$ be the cost of solving t from scratch, including the checking of the result. Retrieving the nearest stored case costs ℓ, and assessing whether that case applies to the present task costs α. If it applies, which happens with probability $p(t, M)$, the case is adapted at expected cost $\bar{\gamma}(t, M)$, where adaptation cost is a nondecreasing function $\gamma(d)$ of the distance d between the task and the stored case, with $\gamma(0) = 0$, and γ is assumed continuous. If it does not apply, the agent falls back on solving from scratch after having paid for retrieval and assessment. The expected cost is

$$C(t \mid M) = \ell + \alpha + p(t, M)\,\bar{\gamma}(t, M) + \big(1 - p(t, M)\big)\,S(t). \tag{4.2}$$

Equation (4.2) includes the overhead of unsuccessful reuse attempts, namely retrieval and assessment that end in fallback, and so it is not the minimum of the costs of the two routes.

Proposition 4.3 (Gradual improvement under growing coverage)**.** *Fix $\ell, \alpha, \bar{\gamma}$ and $S > \ell + \alpha + \bar{\gamma}$, so that the stored route is cheaper than solving afresh, and let p_n be the probability that the nearest stored case applies at the nth encounter, with p_n nondecreasing in n and $p_n \to 1$. Then the expected cost at the nth encounter is nonincreasing in n and converges to $\ell + \alpha + \bar{\gamma}$, which is the cost when a stored case always applies. The limit is forced by $p_n \to 1$, and the improvement is gradual exactly to the extent that p_n is, so that if p_n is continuous and strictly increasing in a continuous measure of experience, the expected cost is continuous and strictly decreasing in it.*

Proof. From Equation (4.2), the expected cost is $\ell + \alpha + \bar{\gamma} + (1 - p_n)(S - \bar{\gamma})$, which decreases as p_n increases, and $S > \bar{\gamma}$ follows from the stated inequality. The limit follows from $p_n \to 1$. □

Remark 4.1. In the special case where the first encounter pays S and every later encounter retrieves the stored result at cost ℓ, the cumulative mean cost per episode is $\ell + (S - \ell)/n$. This decreases smoothly even though the per-episode costs are $S, \ell, \ell, \ldots$, so it describes average expenditure and not the current performance of the agent. Gradual improvement in current performance requires an assumption of the kind in Proposition 4.3, such as growing coverage, falling retrieval latency, or rising applicability.

Item-level evidence bears on the coverage assumption. Tests of candidate models of the transition from algorithmic computation to memory retrieval in skill learning treat learning history as acting on particular items and not on a task category as a whole [21]. That supports treating the applicability probability as a property of items and cases, and it supports history-sensitive change in which strategy is used. It concerns which strategy is used on which item, and it does not by itself settle whether the strategies are separate architectural mechanisms.

Equation (4.2) describes one process with one store and two ways of producing an answer, in which the route taken is determined by how close the nearest stored case is and whether it applies. Where applicability remains reliable and adaptation cost varies continuously with distance, small differences from a stored case can incur small additional costs. Continuity of adaptation cost does not by itself ensure continuity of the applicability probability, and an assessment that fails can trigger an expensive fallback, so the cost of a task near a stored case can change abruptly where applicability does. A task far from every stored case costs approximately $\ell + \alpha + S(t)$.

The verification attributed to stored responses requires care. A stored answer was verified for the situation in which it was produced. Reuse is warranted only when the present instance preserves the conditions on which that verification depended, and a similar but different instance may fail them, which is what the assessment cost α pays for and what the fallback covers. Habitual responses can also be acquired without any verified deliberate solution, by imitation or by reinforcement of outcomes, and nothing here denies that. The claim is conditional. Where a stored response derives from an evaluated solution and the conditions of that evaluation are preserved, its present speed is no evidence that evaluation never occurred, and the idea that a habitual response was never checked cannot be inferred from its speed.

Chapter 5

Amortized Deliberation

5.1 Reuse and the cost of relegation

The previous chapter described instance-based reuse: one process with one store, in which cost falls as stored results come to cover more of the cases met. The present chapter describes procedure-based reuse, in which what is kept is a compressed routine and not a set of solved cases. The two models complement one another and neither replaces the other, since the first states how far stored cases reach and the second states what it costs to relegate a routine and keep it under monitoring. The present chapter addresses the second question: what it costs to relegate a routine and to keep it under monitoring, and under what conditions relegation pays. Deliberation is a cost incurred when a task is first met. If its result is stored in a form that can be reused, later occasions of the same demand do not need to pay for the deliberation again, and the cost per occasion, averaged over the occasions on which the stored result is used, falls as the number of occasions grows. A task that feels hard is frequently one whose costs have not yet been amortized, and a task that feels easy is frequently one whose costs have been spread over many prior occasions. On this view the classification of tasks by speed or effort has stopped doing the work that the dual-process vocabulary was supposed to do, since it classifies a task by an agent's history and not by any intrinsic property of the task.

5.2 A cost model

The notion of amortization can be given a minimal formal statement. The statement below is a proposal intended to make the cost claims precise and testable, and it is not drawn from an empirical fit. Consider a demand of a fixed type that recurs on n occasions. Let c_f denote the cost of meeting an occasion by fresh deliberation at full resolution, with no stored structure to draw on. Let I denote the one-time investment required to produce a relegated routine for the demand, comprising the exposures, errors, and corrections through which the routine is

formed. Let c_r denote the cost of executing the relegated routine on an occasion. Because a relegated routine runs without tracking its internal distinctions, it needs some monitoring if it is to be abandoned when it fails, and let m denote the per-occasion cost of that monitoring. Let q denote the probability per occasion that the routine breaks down in a way that requires promotion of relegated structure, and c_p the cost of the resulting re-promotion and repair.

Definition 5.1 (Amortized cost)**.** *The amortized cost per occasion of meeting the demand through a relegated routine, over n occasions, is*

$$A(n) \;=\; \frac{I}{n} + c_r + m + q\,c_p.$$

Proposition 5.1. *Relegation is economical over n occasions, in the sense that $A(n) < c_f$, if and only if $c_f - c_r - m - q\,c_p > 0$ and*

$$n \;>\; \frac{I}{c_f - c_r - m - q\,c_p}.$$

The proposition follows by direct rearrangement and its content is as modest as its proof. It states the conditions under which compression pays. Relegation is worthwhile when the demand recurs often enough to repay the investment, when the routine is cheap enough relative to fresh deliberation, and when breakdowns are rare enough or cheap enough that the expected cost of repair does not absorb the saving. Two consequences deserve emphasis. The first is that the benefit of relegation depends on the stability of the environment through the parameter q: in an environment where demands of the same nominal type vary so that routines often fail, the expected repair cost increases and the saving shrinks. The model above compares fresh deliberation with one specified routine and has no variable for the depth of relegation, so it supports this point only for a given routine; the extended model below, which includes depth, shows under stated assumptions that optimal depth falls as instability rises. That is the intuition behind the local and graded automaticity described in the relegation chapter, in which stable aspects of a task are deeply relegated and volatile or high-stakes aspects are held at higher resolution. The second is that the condition is stated for a particular agent facing a particular demand. There is no place in it for a classification of tasks as intrinsically fast or intrinsically slow.

Remark 5.1. Setting $m = 0$ in the equation assigns monitoring no cost. It does not show that monitoring is absent, since a monitor may be embedded in a routine at negligible marginal cost. The substantive point is that a routine with no sensitivity to its own failure would run to completion regardless of whether the conditions that justified it still held, and the expected cost of breakdowns would be borne without any chance of reducing it. Sensitivity to failure, whatever its marginal cost, is the formal counterpart of the reversibility claim. A relegated routine is not a blind routine.

5.3 Depth, missed failures, and false alarms

The basic model omits three things that the later chapters make important: the depth of relegation, the loss from failures that the monitor does not detect, and the cost of interventions that were not needed. An extension can include them. Let $d \in [0,1]$ denote relegation depth, the proportion of a routine's distinctions that are not tracked, and let $\theta \geq 0$ index the instability of the environment. Let μ denote monitoring effort, with cost $m(\mu)$, detection probability $\delta(\mu) \in [0,1]$, and $\phi(\mu)$ the unconditional probability per occasion of an unnecessary intervention. If a false-alarm rate conditional on non-failure is used instead, the last term below becomes $(1-q)\phi(\mu)c_u$. Let L denote the loss from a failure that goes undetected and c_u the cost of an unnecessary intervention. The expected cost per occasion is then

$$A(n; d, \mu, \theta) \;=\; \frac{I(d)}{n} + c_r(d) + m(\mu) + q(d,\theta)\big[\delta(\mu)\, c_p + (1-\delta(\mu))\, L\big] + \phi(\mu)\, c_u.$$

Proposition 5.2. *The failure term $q\,[\delta c_p + (1-\delta)L]$ equals $q\,[L - \delta(L - c_p)]$, so for fixed $q > 0$ it strictly decreases in the detection probability δ if and only if $L > c_p$. A more sensitive monitor therefore lowers expected cost, despite triggering more repairs, whenever an undetected failure costs more than a repair, provided its own cost $m(\mu)$ and false-alarm cost $\phi(\mu)c_u$ do not offset the gain.*

Proposition 5.3. *Suppose, with monitoring effort held fixed, that A is twice differentiable, that $d^*(\theta)$ is an interior minimizer satisfying the second-order condition $A_{dd}(d^*(\theta), \theta) > 0$, and that instability raises the marginal cost of depth, so that $\partial^2 A / \partial d\, \partial\theta > 0$ at $(d^*(\theta), \theta)$. Then d^* is locally decreasing in θ.*

The second proposition follows from the first-order condition $A_d(d^*(\theta), \theta) = 0$ by implicit differentiation, which gives $\mathrm{d}d^*/\mathrm{d}\theta = -A_{d\theta}/A_{dd} < 0$. Its assumptions carry the content. That instability raises the marginal cost of depth is plausible if deeper relegation raises the breakdown probability more steeply in less stable environments, but this is a hypothesis about the form of $q(d,\theta)$ and is subject to test. The assumption is satisfiable, as the example below shows. The extension also repairs a defect of the basic model. Without the missed-failure term, a monitor that catches more failures would appear more expensive simply because it triggers more repairs, and the comparison between monitors would reward silence. With the term included, the model connects to the argument about expertise: skilled performance is well-validated relegated structure together with a monitor tuned so that missed failures are rare relative to their cost.

Remark 5.2 (A satisfiable specification). Let $q(d,\theta) = \kappa\theta(\varepsilon + d)$, where $\varepsilon > 0$ is a baseline failure term so that a routine with no relegation can still fail, with $\kappa\theta(\varepsilon+1) \leq 1$ on the range considered, let $I(d) = I_1 d^2$, and let $c_r(d) = c_f - c_1 d$ with $c_1 > 0$. Write $B = \delta c_p + (1-\delta)L$ for the expected cost of a breakdown at fixed monitoring effort. Then $A_{dd} = 2I_1/n > 0$ and $A_{d\theta} = \kappa B > 0$, and the baseline term does not enter the first-order condition, so both conditions of the second

proposition hold. The first-order condition $2I_1 d/n - c_1 + \kappa\theta B = 0$ gives $d^* = n(c_1 - \kappa\theta B)/(2I_1)$ whenever this lies in $[0, 1]$. Otherwise the optimum is at a boundary, $d^* = 0$ or $d^* = 1$, and the comparative statics hold only on the interior range. Depth rises with recurrence n and falls with instability θ and with breakdown cost B, as the account requires. The specification is chosen for tractability and not for realism. In this example $d = 0$ denotes an unrelegated routine within the extended model, which still pays for monitoring and for residual failures at the baseline rate, so c_f here is the execution cost at $d = 0$. The benchmark c_f of the break-even comparison earlier in the chapter is a separate quantity, the cost of fresh deliberation, and the two coincide only if the unrelegated routine is taken to be fresh deliberation with the same monitoring and failure charges.

5.4 Reuse and the preservation of verification conditions

Amortization presupposes that what has been paid for can be reused, and reuse is safe only where the conditions under which the stored result was validated continue to hold. A relegated routine carries with it, implicitly, the conditions of its formation: the range of cases on which it was exercised, the feedback it received, and the stability of the environment during that period. When a routine is applied outside those conditions, it can fail without any signal of failure unless the monitoring that accompanies it is sensitive to the departure. This points to what is lost in the usual contrast between the careless fast system and the careful slow one. The error is not that a fast process was used where a slow one was required. It is that a stored result was applied outside the conditions that validated it and nothing detected the departure.

This reframing has a practical advantage. A recommendation to engage the slow system when the stakes are high gives no guidance about when the fast result should be distrusted. The amortization account gives a more specific one: the result should be distrusted when the present case differs from the validating cases in ways that matter, and the capacity that matters is the capacity to detect such differences. It also explains why expertise, which is largely a matter of well-validated relegated structure together with finely tuned monitoring, is characteristically fast and reliable within its domain and sometimes fails sharply at its edges.

5.5 A cached architecture as an illustration

A concrete case in which the split between reusable and request-specific computation is explicit is a recent family of typed decision models, which choose among a set of options, judge a condition, or assign a score, and which have been described by their authors with the language of System 1, a description this monograph does not adopt [15]. The architecture is used here only as an instance of amortization, and no claim is made about its performance or about the authors' evaluations; what makes the illustration clean is the exactness of the substitution and

not any property of the trained model. In it, each of K options is encoded independently of the request into a vector together with per-layer key and value projections, and these entries are stored. A request is encoded fresh, and a small decision stage refines the request state by attending over the stored entries. The reported property relevant here is that substituting cached entries for recomputed ones leaves every output unchanged in exact arithmetic, because an entry depends only on the option.

Let c_{opt} be the cost of encoding and projecting one option, and C_{req} the cost of encoding the request and running the decision stage over prepared entries. If a fraction $\rho \in [0, 1]$ of the K options is already in the store,

$$C(\rho) = C_{\mathrm{req}} + (1 - \rho)\, K\, c_{\mathrm{opt}}. \tag{5.1}$$

At $\rho = 0$ every option is computed afresh, and at $\rho = 1$ only the request-specific work remains. The two ends correspond to what would be called slow and fast operation, and they are produced by the same weights, the same procedure, and the same decisions, differing only in how much of the work has been done in advance. The cost is linear in ρ, so there is no value of ρ at which the model changes into another kind of model. Unlike the memory of Equation (4.2), retrieval here is exact, since an option is either in the store or it is not, and the applicability step is trivial because an entry depends only on the option. The decision stage is also run on every request even at full reuse, which is the architectural counterpart of the observation that a habitual response is still a computation on the present situation, with some of its inputs precomputed.

The illustration separates two things that the label runs together. The low latency at high reuse comes from the cache, and it is a cost property. The accuracy of the model on its decision tasks comes from training on a very large number of decisions of that type, which is practice, and the architecture exposes the amortized part as an explicit memory. The sensible name for such a model describes what it does, namely typed-output decision with cached option representations. The name of a supposed cognitive system adds a commitment, that the category exists independently of the training and the caching that made the tasks cheap, which the construction does not support.

Chapter 6

Routing and the Appearance of Two Systems

6.1 The routing argument

A two-systems theory needs an assignment of tasks to systems. Let $\lambda : \mathcal{T} \to \{1, 2\}$ be the assignment and let $r(t)$ be the cost of computing it. Suppose the routed agent pays $r(t)$ and then pays $C_{\lambda(t)}(t)$ under the selected system, so that its total cost is

$$C_{\text{routed}}(t) = r(t) + C_{\lambda(t)}(t). \tag{6.1}$$

Proposition 6.1 (Routing must be cheaper than the saving)**.** *If the speed advantage of System 1 on a task t is to survive routing, then $r(t) < C_2(t) - C_1(t)$.*

Proof. The routed cost on a System 1 task is $r(t) + C_1(t)$, which beats the System 2 cost $C_2(t)$ only if $r(t) < C_2(t) - C_1(t)$. □

The proposition permits cheap routing between distinct mechanisms, and it does not show that every cheap router is a memory lookup. The argument for parsimony is narrower. In the model of Equation (4.2), the operations that determine whether a stored response is used are retrieval and applicability assessment, whose cost $\ell + \alpha$ is paid on every encounter and which already settle whether the fast route succeeds. A routing policy could use a cheap cue, a task identifier, or a learned estimate of retrieval success without performing retrieval itself, so a separate dispatcher is not impossible. The model shows only that a dispatcher is unnecessary for its specified execution policy, in which the success of the fast route is settled by steps already paid for. A router added to it would need to be justified by evidence that its cue carries information the retrieval steps do not, and the speed evidence discussed here does not supply that.

A serious alternative is the expected value of control account, in which control is allocated by weighing the expected benefit of a control signal against its effort cost, and the signals selected vary in intensity as well as in identity [22]. This is a specified mechanism of control allocation and not an unexplained supervisor, so it is not open to the objection below, and it fits the position taken here to the extent that the allocation it describes is graded. It also shows the shape that a defensible two-mechanism account would take, namely a stated cost-benefit policy whose inputs and outputs can be tested, with the retrieval and assessment policy of Equation (4.2) as one competitor, in place of a categorical sorter justified by the speed of the outcome.

This is also where a supernatural reading can enter, though only under one condition. Two mechanisms need not be coordinated by an agent-like supervisor, and a learned mechanical routing policy raises no difficulty of this kind. The objection applies when the sorter is taken to be a supervisor that surveys each thought, judges it easy or hard, and dispatches it, while its own intelligence is left unexplained, because the supervisor is then the figure the dichotomy was meant to explain. A retrieval and assessment step against a store of prior solutions requires no such supervisor, and it is itself a candidate routing policy, of the graded kind discussed above.

6.2 Why the partition looks real

Three features of the evidence make a continuum look like two groups, and they induce three different kinds of error: the first is an error of selection, the second of salience, and the third of inference.

The first is the selection of examples. Studies of fast intuitive judgment use tasks that are easy for a typical adult, meaning tasks for which $n_h(t)$ is large in the sampled population, and studies of deliberate reasoning use tasks for which it is small. If the label System 1 is then applied to the tasks with large n, the label predicts low cost by construction, and the finding that System 1 tasks are cheap is a restatement of how the tasks were chosen.

The second is the observer's attention to the ends of the practice curve. A person remembers the first, halting attempt at a new appliance and the thousandth effortless one, and no episode in between is marked. Naming the two ends gives the impression of two machines, whereas the intermediate points exist and are unremarkable.

The third is that the shape of a cost distribution underdetermines the architecture that produced it. Suppose practice counts, treated as a continuous approximation to what are in fact discrete counts, are distributed with density f_N and cost is $Y = c(N)$ with c as in Equation (4.1). Since c is strictly decreasing and differentiable, the density of Y at $c(n)$ is

$$f_Y(c(n)) = \frac{f_N(n)}{|c'(n)|} = \frac{f_N(n)\, n^{\beta+1}}{b\beta}. \tag{6.2}$$

The factor $n^{\beta+1}$ is nonconstant, so it moves the locations of the density maxima and can change their number. A unimodal distribution of practice counts can therefore yield a bimodal distribution of costs, and a bimodal distribution of practice counts can yield a unimodal one. A mixture of two mechanisms can likewise appear unimodal. Neither the presence nor the absence of bimodality in cost, taken by itself, identifies the number of mechanisms, and it follows that a bimodal histogram of response times does not, by itself, identify two systems. This conclusion serves the broader argument better than a claim that a single process cannot produce bimodality, since it shows that distributional appearance underdetermines architecture in both directions.

Chapter 7

Routes to Automaticity

This chapter has a limiting role in the argument. The preceding chapters explain contrasts in performance by relegation and amortization, and the present chapter states what that explanation does not cover. It marks the boundary of the claim, so that the account can be assessed at the level at which it is made. The boundary matters for the argument in one further way. If a substantial fraction of the learned competencies on which the monograph relies were found to be implicitly acquired, the claim that compression explains their automaticity would narrow, but the limited thesis would survive, because it concerns what profiles of speed, effort, and reportability do and do not license and not whether the underlying structure was ever explicit.

7.1 The objection from innate automaticity

The most natural objection to the relegation account is that some automatic processes appear to be automatic from the start. Low-level perceptual binding, reflexive threat responses, and basic affective valences do not seem to be compressed deliberation, either developmentally or evolutionarily. If so, the claim that System 1 is relegated System 2 cannot be general. The objection is correct against the stronger thesis, and the account developed here accepts the distinction that it requires.

Two sources of automaticity are distinguished. Hardwired automaticity is structure fixed by evolution and development, which operates automatically without having been acquired through practice. Relegated automaticity is structure that was acquired and stabilized through experience and compressed out of attention through repetition. The two can be distinguished by their developmental course, by their sensitivity to training, and by how their operation changes when attention is directed to them, though the evidence available for any given process may be incomplete.

7.2 Where the weight falls

The claim made on behalf of relegation is a claim about scope. The phenomena that are central to debates about rationality, bias, expertise, and artificial intelligence are mostly learned, and are relegated in the sense defined here where their acquisition included explicit coordination. Route selection, linguistic interpretation, social inference, probabilistic expectation, and strategic choice are learned competencies that move from effortful to effortless, and they are the competencies that dual-process discussions use as examples when they contrast the fast system and the slow. A person who understands a sentence in a language learned in childhood does so with little attention, but the understanding rests on a long history of acquisition, and the ease of the performance reflects that history. That example supports the claim that the competency was acquired through experience. It does not by itself show that its constituent operations were ever explicit, so it needs a separate argument to count as relegation as defined here; where acquisition proceeded by implicit learning, the appropriate claim is the weaker one, that practice history and not a second kind of process explains the ease. A person who judges that a stranger's expression signals irritation does so quickly on the basis of social learning, and the speed does not require that the process be of a different kind from the one that would be used to work out the same conclusion deliberately.

The scoping is not a retreat. It states the claim at the level at which the evidence bears on it. For hardwired processes the relegation account says nothing about their origin, but it says something about their integration with higher cognition: even innate processes can be attentionally modulated. Affective responses can be highlighted, suppressed, reinterpreted, or overridden depending on context, which shows that automaticity at the level of the process does not exclude access at the level of the system's control.

7.3 Compression and its dependence on exploration

A further distinction concerns where the structure behind a fast and fluent performance came from. Let Ω be a space of configurations of some system, and let $\Omega_{\text{valid}} \subseteq \Omega$ be the configurations that satisfy its constraints, as working software, load-bearing structures, and valid proofs do. A bare mapping from valid configurations to a smaller representation is too weak to capture compression, since a constant function satisfies it while preserving nothing, so a decoder and a preservation condition are needed.

Definition 7.1 (Compression)**.** *A* compression *is a triple* $(\text{Enc}, \text{Dec}, \sim)$ *consisting of an encoder* $\text{Enc} : \Omega_{\text{valid}} \to \Sigma$*, a decoder* $\text{Dec} : \Sigma \to \Omega$*, and a task-relevant equivalence relation* $\sim$ *on* Ω_{valid}*, such that* $\text{Dec}(\text{Enc}(s)) \sim s$ *for all* $s \in \Omega_{\text{valid}}$*, possibly subject to a bounded distortion condition in place of exact equivalence.*

Compression in this sense eliminates redundancy while preserving task-relevant structure,

encoding invariants discovered during exploration and suppressing intermediate derivations that are not needed to reconstruct behavior up to $\sim$. A universal claim that no valid compression can be constructed without prior exploration does not hold, since a mapping might be inherited from another agent, derived analytically from constraints already known, or found by accident. The defensible claim is epistemic and conditional.

Proposition 7.1 (Conditional dependency of compression)**.** *In an initially unknown task domain, constructing a reliable compression generally requires information about which variations preserve $\sim$-relevant validity. Absent an external source of this information, such as prior search by the agent, observation, instruction, or an inherited body of structural knowledge, it must be acquired through exploration of Ω near the boundary of Ω_{valid}.*

Sketch. A compression that removes components while preserving $\sim$ requires knowing which components are invariant under $\sim$ and which are incidental. Where this knowledge is not supplied externally, it can only be established by observing which variations preserve validity and which do not, and that is exploration of the boundary between valid and invalid configurations. □

The exploration in question need not have been performed by the agent who benefits from it. It can be carried out by another person, an institution, a culture, an evolutionary process, or an earlier generation of a system, and its results transmitted and not rediscovered. This observation motivates the two cases that follow, which show that fast and fluent output can arise without the agent's own trajectory of deliberate practice.

7.4 Inherited compression

Consider an agent whose entire competence in a domain consists of an inherited representation, where the agent performed none of the exploration that produced it. Large language models are the clean paradigm case. A model's fluent, low-latency completions are a compression over regularities discovered through the effortful, iterative, error-correcting production of the human text on which it was trained, which is genuine exploration performed by many agents over a long history. The model's forward pass shows the surface signature of relegated performance, namely speed, fluency, and the absence of visible deliberation, but nothing in that forward pass corresponds to an agent's own trajectory from deliberate monitoring to earned stabilization, because no such trajectory occurred for the system producing the output. This will be called *inherited compression*: a representation transmitted to an agent that performed none of the exploration behind it, as opposed to relegation, in which the agent that now executes automatically is the same agent that once explored deliberately.

The distinction reframes a question that recurs whenever dual-process language is applied to artificial systems, namely whether a model has System 1 processing without a developed System 2. Posed that way the question asks only about behavioral signature, whether the

output is fast and non-deliberative, and the answers built on that framing tend to list what the system lacks without engaging the structural question underneath. The structural question is not whether behavior looks automatic but whose trajectory produced its stabilization. Inherited compression can reproduce the behavioral signature of relegation without instantiating the process that the signature is usually taken to indicate, so treating fast, fluent output as evidence of relegated System 1 processing risks a category error, since it mistakes a property of the artifact's presentation for a property of the process that generated it. The same care applies in the other direction. Fluent output does not show that deliberate computation is absent, and it does not show that anything has been relegated either. This is why the discussion of artificial systems in Chapter 10 claims only that the structure of a trained model is hidden and does not claim that it is relegated.

7.5 Conditioning

Inherited compression is not the only mechanism that can produce automaticity without the agent's own deliberate exploration. When an advertiser transforms a brand into characters and repeats the pairing until consumers respond to the brand automatically, the consumer does possess the attentional machinery that relegation requires, namely an active attention set and a capacity for stabilization. But the stabilization was not produced by the consumer's own deliberate practice at solving a task of the consumer's own. It was produced by an external actor engineering the consumer's exposure history toward a fixed association, for that actor's objective. This will be called *conditioning*: stabilization that is the agent's own, but whose triggering exposure was authored by someone else.

7.6 Three routes stated on common ground

With the apparatus of Chapter 3 in hand, the three routes can be stated together. *Relegation proper* holds when the stabilization probability is estimated over the agent's own trajectory through the known aspects, the attention set, and the relegated set: the same agent that once represented an aspect under active monitoring is the one whose repeated attempts, under its own policy σ, drove the constraint-satisfaction probability above threshold. *Inherited compression* holds when a receiving agent is given a compression directly, without ever instantiating a trajectory through $\mathcal{E}$ of its own, so that there is no attention set from which anything was relegated, only an artifact whose stabilization, if the term applies at all, was earned by a different agent. *Conditioning* holds when the stabilization probability is genuinely estimated over the receiving agent's own responses, but the policy and the distribution of exposures constituting $\mathcal{E}$ were selected by an external actor for that actor's objective and not as a result of the agent's own deliberate attempts to solve a task. All three can produce output that is fast,

fluent, and free of visible deliberation. Only the first is relegation in the sense used here, and the differences among the three concern reliability, transfer, and whose interests the resulting automaticity serves. The diagnostic question for any fast performance is therefore a question about trajectory: whose exploration produced the stabilization, by what mechanism, and for whose objective.

7.7 Relegation as one route among several

Three claims need to be separated. The first is that controlled procedures can, through repetition and correction, become increasingly autonomous. The second is that all expert intuition has controlled, deliberative origins. The third is that all Type 1 cognition is compiled Type 2 cognition. The first is well supported and is the content of the account of relegation developed here, which explains a large and important class of expert automaticity. The second is doubtful, and the third is false. Research on automaticity finds that automatic higher-cognitive processes do not arise exclusively from a history of conscious skill acquisition, and that some have evolved substrates or are established through early childhood learning mechanisms that were never consciously decomposed by the individual who now exhibits them [4]. Research on motor learning finds that both explicit and implicit training can lead toward automatic performance that is robust to a second task, and the balance of evidence in some paradigms favors implicit acquisition, which produces routines never represented explicitly enough to count as relegated procedures in the first place [13]. Earlier work on automaticity had already found that no single process satisfies all the criteria bundled into the popular notion of the automatic, which undermines any picture of automaticity as one unified state that a system either has or lacks [3].

Proposition 7.2 (Relegation as one route to Type 1 automaticity)**.** *A substantial class of expert Type 1 performances consists of procedures that once required controlled attention, in that they were represented in the attention set and governed by explicit monitoring, and that became increasingly autonomous through the practice and probabilistic stabilization described in Chapter 3. This class does not exhaust Type 1 cognition, which also includes innate, perceptual, affective, associative, and implicitly acquired processes with no comparable history of explicit control.*

Where an expert performance does belong to the relegation class, what appears as immediate intuition is, for that performance specifically, the residue of prior deliberation. The statement does not generalize to Type 1 cognition as a category. It is this restricted claim, and not an identity between the two types, that the monograph defends. The proposition is a summary claim supported by the cited literature and not a derivation, and it is offered as a statement of scope and as the appropriate account of the cognitive competencies that figure in debates about rational judgment and in the comparison of human and artificial performance.

A neural remark is in order, and it is a modest one. If skilled automatic execution corresponds in part to withdrawal of active monitoring from stabilized aspects, some forms of flow-like performance might show reduced engagement of self-monitoring systems, which is close to the transient hypofrontality hypothesis that its proposer presents as a hypothesis and not as a finding [8]. Imaging during induced flow has found selective reallocation among frontal regions in place of uniform prefrontal withdrawal [25], and reviews stress network-level reorganization [26]. The account entails nothing about neural localization. What is worth keeping is the functional distinction: supervision can be withdrawn because a stabilized aspect no longer needs it, or it can be impaired while it is still needed, and the two differ in reliability and origin.

Chapter 8

Resolution and Its Regulation

8.1 The reframed question

If fast and slow performance differ in the resolution at which a capacity is exercised, then the question that the dual-process vocabulary frames as whether a system has System 2 becomes a question about transitions. How does a system decide that its current resolution is inadequate? What signals trigger a widening of the attentional aperture? What resources are deployed once the decision has been made? These questions are not answered by the dual-process vocabulary, which treats the switch as a primitive, and they are the questions on which the most interesting differences between kinds of cognitive systems are likely to lie.

The question has an asymmetry that the relegation account does not deny. Humans appear to have a distinctive capacity to recognize when intuition has become unreliable, and this capacity may be what the vocabulary of a second system has been used to name. The account accepts the asymmetry and treats it as the explanatory problem that the dual-process language obscures. Widening attentional scope and raising resolution require monitoring mechanisms that detect instability, surprise, error, uncertainty, novelty, or contextual mismatch. Such mechanisms can draw on affective, social, or uncertainty signals, and what matters is sensitivity to breakdown and not slowness.

8.2 Selective return to scrutiny

The central empirical claim of the account can now be stated. When a relegated routine breaks down, the distinctions that return to explicit control are selectively those relevant to the breakdown. A pianist whose fingering fails at a fast tempo scrutinizes the passage and the fingering, and does not thereby re-examine every other aspect of the performance. A driver who meets a diversion re-attends to navigation while continuing to steer with little oversight. If the transition were global, a wholesale change from one mode to another, then a perturbation

to one component of a routine should slow and disrupt the performance as a whole. If it is selective, the disruption and the return of explicit control should be localized to the affected component, with matched effects on error types, on the pattern of slowing, and on what the performer can report.

The claim needs to be stated in terms of the extent of the promoted structure and not only the location of the trigger. A global-switch theorist can accept that breakdowns are triggered locally, by a signal such as an error or a surprise, while holding that the response is global, so that the whole task moves to a different mode. What distinguishes the account is a prediction about the scope of re-promoted structure conditional on a local trigger: the distinctions that become explicit, effortful, and reportable are those within the affected component and those it depends on, and aspects of the task that are unaffected remain at their prior resolution. A global-switch account predicts that unaffected aspects also change, in slowing, in effort, and in reportability. The boundaries of the affected component and of what it depends on must be identified independently of the effect, since otherwise the prediction could not fail. The dependency structure therefore has to be specified in advance from an analysis of the task, and this constrains which tasks are suitable for the test. The claim concerns extent and not identity, so it holds under each of the three readings of later access given in Chapter 3: whether the returning structure is a recovered rule, a reconstruction, or a new procedure, the prediction concerns how much of the task is brought under explicit control.

This is a more discriminating prediction than practice-dependent speed. That performance becomes faster and easier with practice is compatible with several architectures, including accounts on which practice substitutes a different mechanism, and it therefore does not by itself favor relegation. Evidence of selective return would put weight on the account's own proposal, that a monitoring function localizes breakdowns and promotes just the structure involved. It would also give the human examples, the cost model, and the experiments on artificial systems a common question to answer, namely how a system detects which part of a routine needs renewed attention.

8.3 A human study

Selective scrutiny tests the regulation of a routine. Establishing relegation requires evidence about how the routine was acquired as well, because a performer can selectively repair an implicitly acquired routine, or retrieve a stored answer and then work through only the step that failed, and either could produce local effects of the kind predicted above. A more telling human study has four parts. It documents which distinctions participants explicitly coordinate during training, for example through instructed strategies and verifiable intermediate responses. It shows that performance later requires less explicit tracking, for example through reduced dual-task cost, faster responding, and reduced reportability of the distinctions. It perturbs

one of the previously coordinated distinctions after practice and measures whether slowing, errors, and reportability are localized to that distinction while unperturbed distinctions stay unchanged. And it compares these participants with others who reach similar fluency through exposure, without explicit coordination. Fluency matching has to be verified at the time of perturbation, through accuracy, speed, and dual-task cost on the unperturbed task, because participants who coordinated distinctions explicitly might simply have encoded the task more robustly. The design should also show that any advantage of the coordination group is specific to the distinctions it coordinated, by perturbing a distinction that neither group coordinated and finding no group difference there. That control distinction has to be chosen so that the coordination training is not plausibly able to affect its encoding, for example a distinction orthogonal to both trained sets or one drawn from a structurally similar task with different content. If both groups show selective repair, selective scrutiny tells us about regulation and does not establish relegation. If the return of the specific previously coordinated distinctions is found only in the group that coordinated them, the result supports relegation as defined here.

8.4 The switch problem

The literature on dual-process models contains a closely related difficulty. If intuitive and deliberate processing are exclusive, so that some responses are available only to one mode, then a reasoner who is currently in the intuitive mode needs some way to detect that the deliberate mode is required, and detecting that need seems to require the capacity for the very kind of reasoning that has not yet been engaged. De Neys's analysis of switch accounts makes the point that the leading proposals are conceptually problematic because they presuppose exclusivity [6]. The relegation account removes the difficulty at the level of description. It does not treat the switch as a transition between two systems. It treats it as a change in the resolution at which a capacity is exercised, brought about by a monitoring function that runs alongside relegated routines and compares their operation with expectations, and a promotion function that makes explicit the structure the monitor has flagged. Neither function presupposes the capacity for reasoning that it is supposed to call forth. Whether the capacity, the monitor, and the promotion function are implemented by one mechanism or several is left open. The limited thesis requires only that the contrasts in performance can be explained without presupposing two exclusive systems, and the architectural question is a further one that the evidence considered here does not settle.

The cost model of Chapter 5 locates these functions precisely. The monitoring cost m is the price of sensitivity to breakdown, and the repair cost c_p is the price of promotion. A system that economizes on m by monitoring little will relegate cheaply and fail silently, while a system that monitors extensively will relegate more safely and pay more for each occasion. The design problem for any system, biological or artificial, that relegates structure is therefore the problem

of allocating monitoring effort across the aspects of its tasks according to how likely they are to fail and how costly the failures would be.

8.5 A rational counterpoint

Not every treatment of the architecture favors deflation. Milli, Lieder, and Griffiths have argued that a dual-process architecture can be understood as a rational response to the trade-off between the flexibility afforded by multiple systems and the cost of choosing among them, and they find a plausible range of environments in which the optimal architecture includes a fast system that performs no deliberation and a slow system that does [18]. This is a serious position and it deserves a direct reply. The reply has two parts. First, the result concerns the conditions under which a two-system architecture would be optimal, and it does not show that human cognition has such an architecture or that the observed contrasts in performance are produced by one. The cost trade-off that the analysis uses is, in structure, the trade-off among I, c_r, m, q, and c_p in the amortization model above. In the extended model the shape of the optimum depends on the shape of the costs. If the cost of depth is convex, as in the example in Chapter 5, optimal depth is intermediate and varies continuously with recurrence and instability. If the expected cost is concave in depth, for instance because the setup cost $I(d)$ is largely fixed and monitoring is all-or-nothing, then a concave function on an interval attains its minimum at an endpoint, and the optimal configuration is either fully fresh deliberation or a fully relegated routine, which is the two-system case. A two-system architecture then appears as a limiting configuration for environments and cost structures of that kind, and a continuum as the natural solution for others. The endpoint case establishes a choice between operating policies, and it leaves open whether distinct mechanisms implement them. Second, even if some cognitive architectures are well described as two-system, this does not license the reading of every fast and slow contrast as evidence of such an architecture, and it does not license the use of the labels as criteria for excluding systems from reasoning. The two positions are compatible with one another on the points that matter here.

Chapter 9

Model Transfer and Borrowed Interpretations

This chapter has a methodological role. It adds no step to the argument about relegation. It supplies a case study of how a vocabulary developed for one domain acquires interpretive content in another, so that the application to artificial systems in the next chapter can be read as an instance of a general pattern and not as an isolated complaint.

9.1 The notion of model transfer

Philosophers of science use the term model transfer for cases in which a formal model developed in one discipline is brought into another and used to understand phenomena that differ in important ways from those for which the model was originally introduced. The descriptive notion is neutral. The standard paradigm is the Lotka-Volterra model, whose oscillatory structure was developed in connection with chemical diffusion and then adapted to describe the periodic dynamics of predator and prey populations, where it proved explanatorily productive. Mel Andrews, in a conversation with Rahul Sam published on 2 November 2026, uses this case to illustrate transfer done well: the formalism travels, and the interpretation is rebuilt for the new domain without carrying over the features of the old one that do not belong [2]. The population dynamics model brings with it the idea of two quantities that oscillate at different rates, and it does not bring the idea that individual organisms have molecular weights.

The problem arises when the interpretive content of the original domain clings to the formalism. Andrews describes this as conceptual baggage, and she observes that it is most likely in fields where the phenomena are less well understood than those of the source domain and where the default level of conceptual care is low. In her account, the example is the extension of the formal models of conditioning and extinction, used to describe learned associations in dogs, monkeys, fish, mice, and humans, to plants such as Mimosa pudica, whose leaves fold on contact.

She reports that researchers in the area of plant cognition take the fit of the same formalism to be a reason to regard the plants as cognitive, and she calls the inference fallacious, since the applicability of a formalism does not license a conclusion about the fundamental nature of the phenomenon that it describes. The monograph relies on her argument only for this logical point, and it makes no claim of its own about what the plant experiments established. The fit of a model to a pattern of change over time is a resemblance, and whether the pattern is produced by anything that warrants the language of learning or cognition requires evidence about what produces it and what other explanations can be excluded.

9.2 Taxonomies as interpretive shortcuts

Andrews also describes taxonomies of models as hermeneutic shortcuts: schemata that hand the user a way of interpreting a model and its relation to the world, and that thereby simplify interpretation at the risk of imposing distinctions the models do not support. Her own example concerns the division of models into models of data and models of theory, which she argues are hybrid in their conceptual content. The description fits the dual-process partition closely. The vocabulary of Systems is a schema for interpreting the contrast between fast and slow performance. It tells the reader that fast performance is to be read as the output of one kind of process with one character, and slow performance as the output of another. Like other interpretive shortcuts it economizes effort, and like them it carries commitments that the evidence for the original contrast does not support.

9.3 Systems vocabulary and language models

What migrates into discussion of language models is better described as a taxonomy together with its interpretation than as a mathematical model. The fast and slow contrast is a classification of performance profiles: fast and fluent on one side, slow and effortful on the other. The closer parallel is therefore Andrews's account of taxonomies as interpretive shortcuts, and the Lotka-Volterra and plant cases show what goes wrong when interpretive content is carried across with a formal or classificatory structure. The transfer carries over the interpretation that goes with the taxonomy, namely that the contrast marks two kinds of process and that a system can have one without the other. The host of the conversation described, from his own experience, a tendency among some practitioners to treat the vocabulary as a folk ontology and to speak of models as having critical-thinking systems in a sense that their more careful colleagues explicitly disavow. This is an impression reported in the conversation and not a sourced finding. Andrews's diagnosis is that the confusion is not merely accidental. In her account, the labels attached to current systems, including the label of reasoning models and the claim of expert-level performance, are used in ways that create or exploit conceptual confusion,

because the confusion has commercial value. This is her assessment and not a conclusion that the framework developed here depends on. What the framework supports is the narrower observation that the labels carry interpretive content that the performance measurements do not establish.

The same point applies to the label of doctoral-level performance. A score on a set of difficult questions establishes how a system performed on those questions under the conditions of the test. It does not establish the capacities that the phrase invites the reader to infer: the formulation of an original research question, the recognition that new evidence undermines a project, or the revision of a project over time. These are capacities that a doctoral education is meant to cultivate, and a benchmark score is not a measurement of them. The inference from score to capacity is a transfer of interpretation of the same kind that Andrews describes for the plants.

9.4 Intelligence as a borrowed concept

A related issue concerns the concept of intelligence, and the method of this chapter can be applied to it directly. A measure of a specific capacity, such as accuracy on a defined class of problems under specified conditions, licenses a claim about performance on that class under those conditions and, given evidence about generalization, about nearby classes. What the word intelligence adds to such a measure is an aggregation across heterogeneous capacities and an implied single scale on which systems can be ranked, and neither is supplied by the measure. Andrews argues that intelligence is not a properly scientific construct, comparing predictions of the arrival of artificial general intelligence to asking a horticulturalist when a variety of green bean with exceptionally good vibes will be bred: the question has no determinate success condition. The broader claim is contested, since psychometric traditions treat intelligence as measurable in humans, and the monograph does not rest on it. What it needs is the narrower claim, that forecasts of arrival for something whose criterion of arrival is undefined do not have the status that the forecasts claim, and that a well-designed measure of a specific capacity supports a precise claim whose scope must be respected when the result travels. This is the same lesson as for model transfer: results that are carefully established in one setting are extended to another by means of a word, and the word carries more than the evidence does.

Chapter 10

Inference-Time Computation Reconsidered

10.1 What the practice is

Current systems built on large language models are often given more computation at the time of answering. They may be permitted to produce long intermediate outputs before the final answer, to sample several candidate solutions, to search over alternatives, or to check and revise their own results. These practices are frequently described by borrowing the vocabulary of Systems, so that the addition of extra computation is said to give the model a slow, deliberate mode in addition to a fast one. In this vocabulary, a model without such computation is a System 1 system and a model with it has acquired System 2.

The relegation account suggests a different description. A trained model embodies a large amount of learned structure that runs without being exposed in the output. Whether this structure is relegated in the strict sense of the definition, which requires a prior explicit phase, is not established by anything in the training process, and the account needs only the weaker claim that it is hidden, in that its distinctions are not tracked in the output. The vocabulary of resolution and cost applies to it in that weaker sense. The model's weights are fixed once training is complete, so the structure that can be exercised at answering time is that structure. Giving the model additional computation at answering time does not alter its weights, and it does not add a faculty. It can support scrutiny at a higher resolution of the model's existing structure on a given problem, by allowing intermediate results to be produced and fed back, by allowing alternatives to be held and compared, and by allowing results to be checked. Whether it does so is the question that the experiment below is designed to investigate. What the added computation provides is, at most, closer to a larger attentional budget for working a problem out afresh than to a second kind of mind.

10.2 A behavioral test

The description yields a test that does not depend on the vocabulary of Systems. For a given model and a family of subtasks, one can measure accuracy as a function of the inference budget allowed. Let $a_j(b)$ denote accuracy on subtask j with budget b. The shape of the curve is informative. Three patterns can be distinguished, each with a defeasible reading. Where accuracy is high at small budgets and does not improve as the budget increases, the pattern is consistent with the relevant structure being routine for the model, compressed in training so deeply that extra computation adds nothing; it is also consistent with a ceiling in the measure, or with an inference procedure that makes poor use of extra computation. Where accuracy rises substantially with budget, the pattern is consistent with a significant portion of the work being done afresh on each occasion. It establishes only that performance improved under the larger budget with the procedure used, and the improvement could come from constructing a solution, from sampling more candidates, or from running a practiced verification routine. Where accuracy remains low at all budgets, the demand may lie beyond what the model's structure can accomplish even with more resolution, or the procedure may fail to exploit the budget. The partition of subtasks into these groups is relative to a model, a procedure, and a training history, in the same way that the partition of tasks into the effortless and the effortful is relative to an agent and its practice history.

The test places the questions that matter where they can be investigated. It asks which operations a system can perform routinely and which require additional computation on the present occasion, and it asks how the answer depends on the relation between the subtask and the training distribution. It also supports a more discriminating comparison between systems than the label of Systems allows, because two models might perform equally well on a benchmark while differing in which subtasks are routine for them.

10.3 A discriminating experiment

A budget curve can answer a useful behavioral question on its own, but it does not bear on the account of selective return. A more discriminating experiment introduces a controlled mismatch into an otherwise familiar task. A step of a routine task is altered, or a datum is changed, so that the default routine yields a wrong answer, while matched variants of the task leave the routine valid and further variants alter an unrelated step. The question is whether additional computation helps selectively: whether accuracy on the mismatched variants rises with budget more than accuracy on matched variants and on variants with an unrelated alteration, and whether the gain depends on where in the task the mismatch occurs. Evidence of this kind is behavioral, defined by accuracy on variants that differ only at the affected step, and does not depend on the content of any intermediate text the system produces. If a system were to

show selective gains, that would bring the artificial case closer to the pianist example, where the location of the breakdown matters as much as the amount of extra work. The stronger version of the claim concerns the scope of what is promoted, so the design should also test scope. Variants can carry two independent mismatches, one signaled, in that it produces a conspicuous inconsistency, and one silent, in that it gives no cue. A global increase in resolution triggered by the signaled mismatch should help the silent one as well, whereas promotion that is local to the signal should not. The critical comparison is whether signaling a problem in component A changes performance on the independently scored, unsignaled problem in component B. Three conditions need to be kept apart. First, B must be shown to be recoverable under direct cueing or an instruction to review the entire task, since otherwise a null effect on B could mean that the silent mismatch was undetectable under either strategy. This verifies recoverability and is not the test of scope, because the instruction is itself an intervention and recovery under it shows only that B can be repaired. Second, the test of scope is whether B is recovered when A is signaled and no instruction is given, since a global increase in resolution predicts recovery and local promotion does not. Third, the condition with nothing signaled estimates the baseline rate at which B is recovered spontaneously, which need not be zero. Recovery of B under the instruction to review the whole task also needs to be established at the same budget as a positive control for global scrutiny, since the contrast between local and global promotion is meaningful only if the global instruction is shown to improve B when it is given. A fourth condition signals A on variants in which a further component is unaltered, and scores whether that component is changed or wrongly revised, since this separates promotion local to A from a general increase in scrutiny that any signal sets off. A factorial design that varies the presence of each mismatch and whether A is signaled, with the components scored separately, supports the comparison. Even then, a selective improvement supports a claim about the scope of correction and leaves the mechanism open, since a system could produce the same pattern by following a cue with a practiced repair routine. If gains were uniform across variants, the added computation would be better described as general resolution increase without localization, which is the deficit the account predicts for systems lacking endogenous control of resolution.

10.4 Matching requirements

The experiment is only as good as the matching of its variants. Matched, mismatched, and unrelated-alteration variants must be matched on difficulty, measured independently of the manipulation, for example by the accuracy of the same system at a very large budget on the unaltered task. They must be matched on distance from the system's training distribution, as far as that can be assessed, for example through the frequency and surface similarity of the content. And they must be matched on the amount of computation that the default routine would expend, for example the length or number of steps of the default solution. The position

of the altered step should be counterbalanced across variants, and the alterations should be matched in surface form and size. Without these controls, a selective gain could reflect difficulty, familiarity, or the length of the default solution and not localization.

10.5 Evidential roles of the human and artificial studies

The human and artificial experiments do different evidential jobs. The human study tests whether previously explicit distinctions are selectively brought back under control, which bears on relegation as defined. The model study tests whether correction is selectively allocated under controlled mismatch, which bears on how scrutiny is regulated and says nothing by itself about how the system acquired its routines. Their connection is the regulation of scrutiny, and the monograph can make that connection without assuming that the two kinds of system acquired their routines in the same way.

10.6 Practice that teaches a procedure

The cached architecture shows practice acting through stored components. A second case shows practice acting differently, and it matters because it blocks the inference that all learned speed is retrieval. Progress-aligned recurrence takes a pretrained language model, repeats a block of its layers a variable number of times, and trains it on sequential state-tracking tasks, such as following a series of swaps among five boxes and answering which item ends in a queried box [27]. Its central training choice is that iteration k is trained to predict the state after the first k operations, so that each iteration has a defined job, and its second is to rescale the recurrent state to a fixed norm after every iteration, which controls the magnitude of the state so that a drifting scale does not change what the repeated transformation computes. A learned halting signal lets the model decide when to answer.

The reported results are specific. Trained on short problems, a small model is reported to extrapolate to problems sixteen times longer with high accuracy, where an explicit chain of thought on the same backbone is reported to do far worse on the longer problems, and the reported ablations show that removing the norm anchoring or replacing the progress-aligned targets with final-answer supervision sharply reduces the extrapolation factor, as does adding depth or extra iterations without the alignment. These are the authors' reports on synthetic tasks and mostly on one small backbone, and the paper itself notes that results on larger backbones vary far more across seeds and that on graph reachability every recipe extrapolates much less, a limitation the monograph does not try to explain.

Four consequences bear on the argument. First, practice here does not act by retrieving complete solved instances, which is the distinction between storing instances and acquiring a procedure. It supports reusable sequential computation that can be applied more times than it

was applied in training, so an account of practice limited to amortized answers is too narrow and should be broadened to the acquisition of computational competence. The result does not show that the latent state alone implements a sufficient transition state, since the full prompt remains available at every iteration. Second, the amount of computation applied is graded, being the number of iterations, an integer-valued and variable depth, and a longer problem simply receives more of it. Third, the availability of extra computation is not what makes it useful. A model that can iterate without the learned organization gains little from more iterations, so the depth of computation depends on training history and not on the architecture's capacity to loop. Fourth, the extended computation runs without any generated intermediate text, and the explicit chain of thought on the same backbone is reported to do far worse on longer problems, so externalization is neither what makes extended computation possible nor what distinguishes it.

The paper also bears on the claim that a correct answer, once reached, simply persists. Its models, trained with the number of iterations matched to problem length, answer correctly at the matching iteration and degrade rapidly if run beyond it, which is why a halting rule is needed, and the learned stopping rules vary considerably in how well they hold up across seeds and lengths. Obtaining an adequate answer and preserving it through further computation are therefore different achievements, and a theory of cost that identifies the first with the second will misdescribe the second. This is a concrete instance of the monitoring problem of Chapter 8: a system needs some means of detecting that its current state is adequate, and of stopping.

10.7 Cautions

Several cautions apply. The first is that a budget curve measures how performance varies with the computation allowed and not how the computation was used. A larger budget tells us how much computation was available, and finding out what the system had to work out afresh requires evidence about what it did with the budget. The second is that visible intermediate output is weak evidence of the computation that produced the answer. There are reasons to doubt that the text a model produces while working on a problem tracks the operations that determine its answer, and so the curves of accuracy against budget are a more appropriate source of evidence than the content of the intermediate text. The third is that practice can relegate the practice of deliberation itself. A model trained to produce extended intermediate output may have compressed the production of that output into a routine, so that a long and apparently deliberate trace can be, in the relevant sense, compressed and hidden: the routine for producing it is deeply compressed, even though the trace is long. This is the counterpart for artificial systems of the observation that a practiced expert can recite a lengthy derivation without effort. Length of output is therefore not a measure of effort in the sense that matters here.

The fourth caution concerns what the test can establish. The budget curves can show whether performance on a subtask is sensitive to budget under a particular procedure, which is consistent with routine status or its absence and establishes neither, and they cannot show that the model has or lacks any further property, such as understanding, that the vocabulary of Systems was sometimes taken to imply. The test replaces the question of whether a system has a second faculty with questions that are answerable, and it does not claim that the answers settle every question that the original vocabulary was used to raise.

10.8 Deficits relocated

The deflation of the partition does not imply that current systems are free of limitations. It relocates them. Fluent output does not show that reasoning is absent, and, as the discussion of inherited compression showed, it does not show that reasoning was ever relegated either, so the interesting deficits need not be the absence of a slow faculty. They concern the regulation of resolution. The structure of a trained model is largely fixed at answering time, so it can rarely revisit, reinterpret, or restate that structure when it breaks down. What appears as a lack of reflection is better described as limited adaptive re-promotion: the return of relegated structure to active processing in response to signals of failure. The depth of a system's internal structure and the visibility of that structure are separate matters that critics tend to conflate. Humans reason slowly when they do because exposing fine distinctions takes time and attention, and slowness is not essential to rationality. The deficit in artificial systems is one of self-directed scrutiny, not speed, and it shows in degradation outside familiar input patterns and in difficulty in pausing, questioning, or revising a transformation in light of goals or unexpected results.

On this view the criticism that a system lacks System 2 reasoning is not so much false as misdirected. It points at a real limitation, the limited control over when internal structure should become explicit, and attributes it to a missing faculty, which suggests the wrong remedy. The remedy suggested by the dual-process picture is the addition of a module. The remedy suggested by the resolution picture is better mechanisms for detecting when default compression is inappropriate, for selectively raising internal resolution, and for holding several representations active while uncertainty is unresolved.

Chapter 11

Objections and Limits

11.1 The objection from reflexes

The first objection has been addressed in Chapter 7: some automaticity is hardwired. The reply is a scope restriction, and it should be restated with its consequence. If the account is read as a claim that everything automatic began as deliberation, it is false, and it is not offered as such. If it is read as a claim about the learned competencies that dominate debates about rational judgment and about the comparison of human and artificial performance, it is the appropriate account. A critic who rejects the account on the basis of reflexes has not engaged with the claim that is made.

11.2 The objection from metacognition

The second objection is that humans have a distinctive capacity to recognize unreliability in their own intuitions, and that this capacity is what a second system names. The reply is that the asymmetry is real and is the problem that the account identifies as central. The objection is therefore an objection to the dual-process vocabulary and not to the relegation account, since the vocabulary hides the problem by naming it. The account's contribution is to restate the capacity as a monitoring and promotion function that is part of a single system, to make its parameters explicit in the cost model, and to note that how well it works varies with the individual, the domain, and the history of practice.

11.3 Evidential status

The claims of the monograph are of three kinds with different evidential standing, and a reader should treat them differently. The limited thesis, that speed, effort, and reportability do not identify cognitive kinds, rests on the observation that these features come apart across tasks

and individuals and on the dependence of cost on practice history, and it is supported by the structure of the existing literature as well as by everyday cases. The relegation account of learned competencies is a functional hypothesis consistent with a large body of evidence about skill acquisition, and it is stated so that it can be tested, in particular through its predictions about reversibility and about the dependence of relegation depth on environmental stability. The cost model is a formal proposal that has not been fitted to data. Its role is to make the claims about amortization precise, and it could be falsified by findings that relegation depth does not track recurrence and stability in the way the model implies. The propositions of the formal chapters are consequences of their stated assumptions and carry no empirical weight beyond those assumptions. The parsimony argument they support shows that a classification by cost does not require two kinds, and it does not show that there is only one. The same holds of the reversal result, which bears on classifications by cost or history and leaves untouched any partition defined by a mechanism.

11.4 What would count against the account

An account that cannot be refuted is of little use, and the present account makes several risky predictions. If the transitions from fast to slow performance were found to depend on fixed properties of tasks and not on signals of breakdown, the reversibility claim would be in trouble. If relegation depth were found to be independent of the recurrence and stability of demands, the cost model would fail. If systems with comparable training histories were found to differ qualitatively, and not in degree, in the way their accuracy depends on the computation allowed at answering time, the claim that added computation increases the resolution at which a capacity is exercised would need revision. And if some mechanisms that supply the same functions as the monitoring and promotion described here were found to require a separate architecture that is exclusive in the sense that De Neys's analysis describes, the deflationary claim would be weakened. The most direct challenge would be evidence that breakdowns, even when triggered by a local signal, return explicit control to the task as a whole and not to the affected component. None of these findings is presently in hand, and each is the kind of result that could be sought.

Chapter 12

Conclusion

12.1 What remains of the distinction

The partition claim is not supported by the evidence usually offered for it, and the observations that motivated it remain. Some responses are produced more cheaply than others, some are accompanied by report and some are not, and some can be interrupted and revised while others run to completion. These are real differences, and they are best treated as distinguishable dimensions that fall into three families. The first family concerns history. One parameter is the number of relevant prior episodes, $n_h(t)$. A second is the distance from the present task to the nearest stored case, together with whether the stored result carries a warrant that applies. A third is whether the practiced competence takes the form of stored instances or of a reusable procedure. The second family concerns execution, and its parameter is the depth of computation actually executed on the present task. The third family concerns access to what was computed, and it comprises three properties that are often run together although they are conceptually distinguishable and need not be coupled. Externalization is a property of the system's states, namely whether intermediate states are written into a context where they can be reread. Decodability is also a property of states, namely whether they can be recovered from the system by a readout of a stated class. Reporting alone concerns an account, namely whether the agent produces a narrative of its process, and such a narrative need not correspond to what was computed. A progress-aligned recurrent model can be deep, decodable by its trained readout, and unexternalized, while a person's practiced arithmetic may be shallow and readily reported after the fact. These dimensions permit combinations that a single fast-slow axis obscures.

What has not been established is the stronger claim that every form of deliberation is retrieval with less work precomputed, and the monograph does not rely on it. The vocabulary of two systems should be retained, if at all, as shorthand for the ends of the practice axis, with the understanding that it names a region of a continuum, implies no mechanism of sorting, and

cannot be read off from speed.

12.2 Summary of the argument

The partition of cognition into a fast system and a slow system has had a long and useful career as a description of how performance varies. This monograph has argued that the career has outrun the evidence at the point where the description became an architecture. The relegation account explains many of the contrasts that the partition describes by compression through practice, a reversible process that leaves a routine operating while its distinctions go untracked. The amortization account shows why the cost of a task is a property of an agent's history and not of the task, and it states the conditions under which compression pays. The limited thesis is defended in full, the stronger thesis is set aside for reasons the account itself supplies, and hardwired automaticity is acknowledged as a separate source.

The deflation has costs and benefits. The cost is rhetorical: a simple two-system story is exchanged for a more precise picture of performance governed by shifting constraints of attention, compression, and control. The benefits are that automaticity is no longer suspect by default, that effort is no longer a sign of rationality, and that human and artificial systems can be placed within one conceptual framework, so that differences between them are stated as differences of functional capacity and not as categorical exclusions. The question the framework proposes to put in place of whether a system has a second faculty is how a system decides that its current resolution is inadequate and what resources it then deploys. That question is open for biological cognition and for artificial cognition alike. Three capacities in particular remain poorly understood: widening the attentional aperture, re-exposing hidden structure, and sustaining unresolved alternatives. The claim on which the account most depends is that a breakdown selectively returns the relevant distinctions to explicit control. The tests proposed here, which examine how performance varies with the computation allowed and how it responds to controlled mismatch, are a first step toward investigating them in artificial systems, and the cost model is a first step toward stating what such investigations would be expected to find.

The case of model transfer adds a lesson about method. When a vocabulary developed in one domain is carried into another, the formalism and the interpretation separate, and what the formalism licenses is much less than what the interpretation suggests. The right response is neither to abandon the vocabulary nor to accept it, but to ask in each new setting what evidence supports each part of the interpretation. For the vocabulary of Systems as it is now used of language models, the evidence supports a description of performance profiles and the way they vary with computation, and it does not support the claim that a second faculty is present or absent. The point is stronger than a lack of support. On the account developed here the question whether a second faculty is present is not well formed, since the term faculty would do independent work only if it picked out a mechanism whose engagement is necessary for a

class of performances and which cannot be replaced by practice history, routing among stored and computed results, and the regulation of resolution, and the argument of the preceding chapters is that no performance profile considered here requires such a mechanism, and the answerable questions are those about routine status, resolution, and control that have been set out above.

Bibliography

[1] Anderson, J. R. 1982. Acquisition of Cognitive Skill. *Psychological Review*, 89(4), 369–406.

[2] Andrews, M. and Sam, R. 2026. Machine Learning (Still) Has a Pseudoscience Problem. Video conversation, published 2 November 2026.

[3] Bargh, J. A. 1992. The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects. *American Journal of Psychology*, 105(2), 181–199.

[4] Bargh, J. A., Schwader, K. L., Hailey, S. E., Dyer, R. L., and Boothby, E. J. 2012. Automaticity in Social-Cognitive Processes. *Trends in Cognitive Sciences*, 16(12), 593–605.

[5] De Neys, W. 2021. On Dual- and Single-Process Models of Thinking. *Perspectives on Psychological Science*, 16(6), 1412–1427.

[6] De Neys, W. 2023. Advancing Theorizing About Fast-and-Slow Thinking. *Behavioral and Brain Sciences*, 46, e111.

[7] Dewey, A. R. 2022. Metacognitive Control in Single- vs. Dual-Process Theory. *Thinking & Reasoning*. doi:10.1080/13546783.2022.2047106.

[8] Dietrich, A. 2003. Functional Neuroanatomy of Altered States of Consciousness: The Transient Hypofrontality Hypothesis. *Consciousness and Cognition*, 12(2), 231–256.

[9] Evans, J. St. B. T. and Stanovich, K. E. 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3), 223–241.

[10] Heathcote, A., Brown, S., and Mewhort, D. J. K. 2000. The Power Law Repealed: The Case for an Exponential Law of Practice. *Psychonomic Bulletin and Review*, 7(2), 185–207.

[11] James, W. 1890. *The Principles of Psychology*. New York: Henry Holt.

[12] Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

[13] Kal, E., Prosée, R., Winters, M., and van der Kamp, J. 2018. Does Implicit Motor Learning Lead to Greater Automatization of Motor Skills Compared to Explicit Motor Learning? A Systematic Review. *PLOS ONE*, 13(9), e0203591.

[14] Keren, G. and Schul, Y. 2009. Two Is Not Always Better Than One: A Critical Evaluation of Two-System Theories. *Perspectives on Psychological Science*, 4(6), 533–550.

[15] Li, Q. and Wang, X. 2026. Jeva: System One Models from First Principles. arXiv:2610.18773.

[16] Logan, G. D. 1988. Toward an Instance Theory of Automatization. *Psychological Review*, 95(4), 492–527.

[17] Melnikoff, D. E. and Bargh, J. A. 2018. The Mythical Number Two. *Trends in Cognitive Sciences*, 22(4), 280–293.

[18] Milli, S., Lieder, F., and Griffiths, T. L. 2021. A Rational Reinterpretation of Dual-Process Theories. *Cognition*, 217, 104881.

[19] Newell, A. and Rosenbloom, P. S. 1981. Mechanisms of Skill Acquisition and the Law of Practice. In J. R. Anderson (ed.), *Cognitive Skills and Their Acquisition*, 1–55. Hillsdale, NJ: Lawrence Erlbaum.

[20] Posner, M. I. and Snyder, C. R. R. 1975. Attention and Cognitive Control. In R. L. Solso (ed.), *Information Processing and Cognition: The Loyola Symposium*. Hillsdale, NJ: Lawrence Erlbaum.

[21] Rickard, T. C. 2004. Strategy Execution in Cognitive Skill Learning: An Item-Level Test of Candidate Models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1), 65–82.

[22] Shenhav, A., Botvinick, M. M., and Cohen, J. D. 2013. The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, 79(2), 217–240.

[23] Stanovich, K. E. and West, R. F. 2000. Individual Differences in Reasoning: Implications for the Rationality Debate? *Behavioral and Brain Sciences*, 23(5), 645–665.

[24] Tversky, A. and Kahneman, D. 1974. Judgment Under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131.

[25] Ulrich, M., Keller, J., Hoenig, K., Waller, C., and Grön, G. 2014. Neural Correlates of Experimentally Induced Flow Experiences. *NeuroImage*, 86, 194–202.

[26] Van der Linden, D., Tops, M., and Bakker, A. B. 2021. Go with the Flow: A Neuroscientific View on Being Fully Engaged. *European Journal of Neuroscience*, 53(4), 947–963.

[27] Wang, P., Hu, Y., Dong, G., Lin, Y., Wang, H., Liu, Z., Wen, Z., Li, J., Yang, J.-M., Wu, W., and Tong, Y. 2026. PAR: Progress-Aligned Recurrence for Length Generalization in State Tracking. arXiv:2610.27628.

[28] Wason, P. C. and Evans, J. St. B. T. 1974. Dual Processes in Reasoning? *Cognition*, 3(2), 141–154.