

Observers, Records, and Verification

A computational theory of time's arrow, memory, and audit — from W. J. Sidis's The Animate and the Inanimate (1925) to the auditing of self-modifying AI. Synthesis of the Sidis Programme, Parts A–H.

The question

In 1925 William James Sidis — the Harvard prodigy, then twenty-one — published a speculative monograph arguing that life is a local reversal of the second law of thermodynamics, and that the second law itself is “a mental law”: the direction in which any organized mind conceives time to flow. This programme took that claim seriously enough to formalize it, test it, and follow it where it broke. The result is a small but coherent theory of observers, records, and verification that ends, unexpectedly, in the auditing problem for self-improving artificial intelligence.

The arc of the programme

Part	Question	Model	Key result	Law
Engine	Formalize Ch. XVI	exact Ehrenfest chain	2×2 slopes ± 0.2204 ; $e^{-\Delta S} = 7.3 \times 10^{-12}$	Sidis Selection Theorem
A	Do fluctuations host pseudo-minds?	equilibrium urn valleys	71 valleys; fall 3.576 = rise 3.556	fluctuation time-symmetry
B	Can a tape run backward?	reversible 2nd-order CA	bit-exact reversal; 1-bit echo decorrelates	arrow is in the boundary
C	What does memory detect?	boundary detector	$H(\text{past} \text{now})=0$ at the anchor	Anchored-Arrow Theorem
D	Does a window restore the cone?	multi-bit register	cone = $L-1$ stored; $h_f - h_b = \Delta H$	Trade-off Law
E	Does t^* emerge in a self-mod machine?	16-bit Sidis Machine	$t^* = 20$; two horizons	$\min()$ (corrected in F)
F	Which constraint binds?	9 interfaces $\times$ 6 budgets	spike vs long-tail; code carries the tail	$S_R(t, W) \geq \Theta$
G	Audit vs alignment?	tagged chain + spec Φ	clock drives violation; $I(\text{ever}) \rightarrow 0.004$	present or provenance, never path
H	Can a temporal hash defeat G3?	rolling MAC audits	integrity $\neq$ detection; internal root forgeable	root of trust is external

Every number above was reproduced exactly by the master script (part_I_master.py) in this bundle.

The five theorems

The programme converged on five statements, each verified in an exact or large-sample model.

- **1. The Selection Theorem.** The probability of causally assembling an observer whose memory is anti-aligned with its local entropy gradient is suppressed by $\exp(-\Delta S)$. Every constructible observer therefore experiences entropy increasing along its subjective time — Sidis's “mental law” is a selection theorem, not a law of mind. (Engine: $e^{-\Delta S} = 7.3 \times 10^{-12}$.)
- **2. The Anchored-Arrow Theorem.** A record-forming agent in a reversible world with one special boundary detects not entropy increase but the boundary itself: $H(\text{past}|\text{now}) = 0$ exactly at the anchor, and the arrow is the direction of anchoring, not of gradient. (Part C.)
- **3. The Trade-off Law.** A window of L stored states restores a certainty cone of exactly $L-1$ steps — the cone is stored history, not computed history — and recording forward is cheaper than backward by exactly the entropy rise: $h_f - h_b = \Delta H$, the information-theoretic H-theorem. The anchor is affordable only while near: $t \ll B/h$. (Part D.)

- **4. The Two-Regime Law.** Anchor loss has a phase structure. Coarse audit interfaces give a spike K-profile and $t^* \approx B/h$ (storage-governed); interfaces that resolve the self-modifying code give a long-tail K-profile and t^* far beyond both Part-E bounds (detection-governed). The boundary is structural: *the code remembers the anchor longest*. (Parts E–F.)
- **5. The Audit-Separation Law.** Audit and alignment decouple in kind. You can audit the present (if the spec is a function of audited variables) or the provenance (while the window reaches the anchor), but never the path: a violation before the audit window opens is undetectable by any state-based audit. A temporal hash adds integrity, not detection — and only under an external root of trust. (Parts G–H.)

The unified picture

Read together, the five theorems describe a single object from five sides: the relationship between a physical process and the records it leaves. The second law is not violated by life (Sidis's error) but is the precondition for records at all (his intuition, corrected). The arrow of time is the direction in which records accumulate toward an anchor. Memory is storage, not inference. And verification — the audit problem — is the same relationship viewed from outside: an auditor is just another record-forming system, subject to the same horizons, the same resolution limits, and the same inability to reconstruct a path from a state.

Sidis asked whether a mind could face the other way in time. The programme's answer: the question is well-formed (the pseudo-living mind is logically possible and dynamically lawful), but such a mind is unassemblable by any causal process in a universe with a low-entropy boundary, and — the engine's mutual-opacity result — we could not recognize it if it existed. The deep reason is the one Sidis glimpsed and the programme made exact: an observer and its records are the same kind of thing, and neither can step outside the arrow that constitutes it.

What this means for RSI

The programme began as philosophy of physics and ended as a constraint theory for the audit of recursive self-improvement. The mapping, now fully quantified:

Sidis result	RSI construct	Constraint
Loschmidt echo (B)	rollback / version recovery	bit-exact records or no return
Anchored-Arrow + Trade-off (C–D)	alignment verification horizon	provenance decays; $t \ll B/h$
Two-Regime Law (E–F)	audit interface design	resolution before budget; see the code
Audit-Separation (G)	continuous monitoring	check the present; history is unauditale
Integrity $\neq$ detection (H)	trusted computing base	the root of trust must be external

The central practical conclusion: an auditor of a self-modifying system faces three separate constraints — resolution, budget, and spec-alignment — and above all of them, the root of trust cannot live inside the audited system. A self-modifying machine that hashes its own log is not audited; it is merely consistent. This is the computational content of Sidis's Chapter XVI insight that no observer can step outside its own time-orientation to verify it.

Limitations and honest caveats

These are toy classical models: a 40-ball urn, a 256-cell automaton, a 16-bit register machine. The “anchor” is a chosen start state, not a moral specification; h is an audit-visible information cost, not a physical Landauer energy; the trusted accumulator of Part H is an oracle assumption. The five theorems are exact within their models and structural in form — the two-regime structure, the code-carries-the-tail finding, the integrity/detection split, and the internal-root collapse are robust — but the specific numbers are parametric. The value of the programme is that it makes each of these constraints visible, measurable, and separable, in a domain (RSI audit) where they are usually conflated.

Coda

Sidis ended his book by presenting his theory “as a possibility to be considered,” listing its own objections and declining to answer them, leaving the reader to weigh the evidence. This programme took him at his word. The answer is that he was wrong about the physics and right about the deep structure: the arrow of time is indeed about the recorder, not the recorded — and a century later, that single insight turns out to govern not only the philosophy of time but the verifiability of any system that rewrites itself.