Blue Pill as Observational Refinement

Computational and Higher-Categorical Semantics of Virtualization Subversion

Anonymous

28 October 2026 · Version 1.0

Abstract

Blue Pill motivates a precise question: when can an operating system continue to compute correctly while losing authority over the machine that realizes its computations? We give an abstract operational semantics of live virtualization subversion, distinguishing prior privilege, insertion, direct execution, interception, and resumption. A coupling condition proves equality of the transcripts of all bounded adaptive guest observers; an approximate version bounds distinguishing advantage. Neither result establishes physical undetectability. We characterize when a security predicate factors through observations, prove a finite compatibility criterion for combining evidence, and construct a bicategory of deterministic simulation witnesses. A presheaf infinity-topos organizes observation contexts and groupoids of compatible realizations. An executable finite model checks transparent and distinguishable cases using exact arithmetic. The contribution is an explicit semantic synthesis with proved hypotheses and counterexamples, not a new exploit or a hardware verification of Blue Pill.

Keywords: virtualization; rootkits; operational semantics; observational equivalence; probabilistic coupling; coalgebras; bicategories; presheaf topoi; evidence.

1. Historical question and contribution

Rutkowska's 2006 presentation describes Blue Pill as moving an already running operating system into a virtual machine under a thin hypervisor, using AMD hardware virtualization [1]. It distinguishes this live insertion from SubVirt's earlier virtual-machine-based rootkit architecture [2]. We study that change of control, without identifying it with the separate problem of obtaining sufficient privilege in the first place. The historical presentation's claims about invisibility are claims to be analyzed, not premises accepted here.

Popek and Goldberg's classical virtualization criterion permits differences involving timing and resource availability [3]. Garfinkel, Adams, Warfield, and Franklin distinguish compatibility from transparency and analyze architectural, resource, and timing discrepancies [4]. Accordingly, a theorem about abstract input/output behavior cannot establish indistinguishability against every physical experiment. Nor does detecting virtualization identify malicious virtualization: an authorized hypervisor can produce the same observation.

Our contribution has four parts: an explicit semantics of the control change; exact and approximate observer-relative results; a compositional account of simulation witnesses; and a topos-theoretic organization of incomplete evidence. The coupling, factorization, and descent techniques are standard mathematical tools. Their application here makes the assumptions behind a Blue Pill-style invisibility

claim inspectable. We claim neither priority for those tools nor a newly discovered attack. Structural operational semantics follows the methodological tradition of Plotkin [5].

2. Scope, states, and observers

Fix countable request and response sets Q, Y. Responses include exceptions and whatever deadline or termination reports the experiment can actually observe. A *context* U specifies the request interface, available observation channels, cost accounting, and trust assumptions. Different contexts need not have the same interface. An observer chooses its next request from its transcript and private randomness, then eventually returns a bit.

A physical configuration contains guest architectural state g, environment state e, a controller h, and an authority assignment a. Architectural state includes the abstract program continuation, registers, memory, and device-facing state. These are mathematical components, not a processor-specific implementation. The authority assignment records who controls each mediator; a guest's numerical privilege level alone does not determine it.

Write $N(s)$ for a native configuration, $G(s, h)$ for a guest configuration, and $X(s, h, q, k)$ for an intercepted request with continuation k. The state s includes g, e and any scheduler information required to make the semantics Markovian. Countable state spaces suffice here. A nondeterministic scheduler must be fixed or explicitly quantified over; arbitrary scheduler choices cannot be silently turned into probabilities.

Threat-model assumptions. The adversary already has the capability needed to install a controller, and the platform permits the modeled virtualization transition. The insertion model abstracts a successful coordinated transition; it does not prove multicore synchronization, device correctness, or persistence. The controller can mediate only channels listed as controlled in U. An independent clock, measurement root, or external observer is independent only if the threat model supplies that property. The observer's decision procedure is fixed; arbitrary rewriting of the detector is not a proof of faithful simulation.

There are two different security questions. *Native versus virtual* asks about execution architecture. *Authorized versus unauthorized* asks about a policy and provenance. We write $P(w) \in \{0, 1\}$ for the latter property when needed. It must not be inferred from the former without an additional premise.

3. Computational semantics of the attack

3.1 A small-step transition system

An abstract guest language is generated by

$$t ::= \mathrm{return}(v) \mid \mathrm{let}\ \ x = \mathrm{op}(q)\ \ \mathrm{in}\ \ t.$$

Internal evaluation and hardware interactions are represented by transitions. A continuation $k : Y \to T$ receives an operation's result. Let $F(s, q) = (s', y)$ be a deterministic native operation, $I(h, q)$ the interception predicate, and $H(s, h, q) = (s', h', y)$ the controller's abstract response. The probabilistic extension replaces these functions by kernels. The principal rules are:

$$\frac{\mathrm{cap}(s) \wedge \mathrm{enabled}(s)}{N(s) \xrightarrow{\tau_{\mathrm{ins}}} G(E(s), h_0)} \quad \text{(Insert)}$$

$$\frac{F(s,q) = (s', y)}{\langle N(s), q, k\rangle \xrightarrow{q/y} \langle N(s'), k(y)\rangle} \quad \text{(Native)}$$

$$\frac{\neg I(h,q) \quad F(s,q) = (s', y)}{\langle G(s,h), q, k\rangle \xrightarrow{q/y} \langle G(s', h), k(y)\rangle} \quad \text{(Direct)}$$

$$\frac{I(h,q)}{\langle G(s,h), q, k\rangle \xrightarrow{\tau} X(s,h,q,k)} \quad \text{(Intercept)}$$

$$\frac{H(s,h,q) = (s', h', y)}{X(s,h,q,k) \xrightarrow{q/y} \langle G(s', h'), k(y)\rangle} \quad \text{(Resume).}$$

Here E transports the live state into the guest representation. It need not preserve physical location or authority. A well-formed embedding must preserve the initial observations claimed in U and relate the resulting state to the native state in the relation defined below. Internal controller work may change h between responses. Thus the semantics permits a change of hidden state without a corresponding guest-visible change.

These rules describe the computational structure of the attack: the continuation of the existing operating system is preserved while a new mediator determines some of its future transitions. They provide no driver, instruction sequence, memory layout, deployment procedure, or persistence mechanism. The term *attack* refers to the unauthorized change in authority, not to a special mathematical property of virtualization.

3.2 Weak steps, cost, and divergence

For a context that hides administrative transitions (including insertion when explicitly covered), write τ for their erased labels and define

$$z \xRightarrow{q/y} z' \quad \Longleftrightarrow \quad z \xrightarrow{\tau*} \xrightarrow{q/y} \xrightarrow{\tau*} z'.$$

This notation alone proves nothing about termination. A controller that performs silent work forever is not equivalent to a responsive native machine for a deadline-sensitive observer. We either assume almost-sure response for every permitted request, or explicitly use an absorbing semantic outcome $\perp$. Divergence is not automatically a finitely observable event: a real timeout is a bounded experiment with a specified timer.

Each transition may additionally carry a nonnegative physical cost c. Costs add along paths. Erasing τ from a logical trace does not erase its cost from a context that measures cost independently. Insertion cost and intermediate states are likewise observable unless excluded or matched. Our principal theorem begins at related post-insertion states. It covers insertion itself only when the observation model includes a matching insertion argument.

At response boundaries a machine is

$$M = (S_M, m_0, K_M), \qquad K_M : S_M \times Q \longrightarrow \mathcal{D}(Y \times S_M),$$

where $\mathcal{D}$ denotes countably supported probability distributions. Absorbing outcomes make the kernel total when required. The kernel summarizes the weak operational transitions, including the chosen environment and scheduler. A physical claim requires a separate adequacy proof relating an implementation to this kernel.

4. Observer-relative indistinguishability

Definition 4.1 (response coupling). A relation $R \subseteq S_N \times S_V$ is an exact response simulation if, whenever mRn and $q \in Q$, there is a coupling $\gamma_{m,n,q}$ of $K_N(m,q)$ and $K_V(n,q)$ supported on

$$\{((y,m'),(y,n')) : m'Rn'\}.$$

Thus both the next response and the invariant relating successor states are preserved. Equality of next-response marginals alone is insufficient: correlations with hidden successor states affect later answers. For countable spaces all couplings used below are ordinary discrete distributions.

Theorem 4.2 (adaptive transcript equality). Suppose m_0Rn_0 and Definition 4.1 holds. For every randomized adaptive observer D making at most n requests, the transcript laws in the native and virtual machines are equal. In particular,

$$\Pr[D^N = 1] = \Pr[D^V = 1].$$

Proof. Couple the observer's private random tape identically in both executions. Initially the histories coincide and the states are related. If these facts hold after i requests, the observer chooses the same next request. Use $\gamma_{m,n,q}$ to obtain equal responses and related next states with probability one. Induction proves equal transcripts through request n. A stopping decision and final bit are functions of the shared tape and transcript, so also agree. An observer that stops early can be padded with a fixed terminal response. This proves the claim. □

This is a sufficient semantic criterion, not a verification that Blue Pill satisfies it. It quantifies over the requests admitted by U, not over arbitrary experiments with the physical machine. A hidden unauthorized action can coexist with this theorem only when its consequences remain outside those observations or are included in a valid coupling.

Theorem 4.3 (approximate form). Suppose instead that every related pair and permitted request has a coupling whose equal-response, related-successor event has probability at least $1 - \varepsilon$, where $0 \le \varepsilon \le 1$. Then every observer using at most n requests satisfies

$$\left|\Pr[D^N = 1] - \Pr[D^V = 1]\right| \le 1 - (1-\varepsilon)^n \le n\varepsilon.$$

Proof. Until the first failed coupling, histories and related states agree, so the next conditional success probability is at least $1 - \varepsilon$. Iterated conditioning gives probability at least $(1-\varepsilon)^n$ of no failure. Output bits coincide on that event. The difference in acceptance probabilities is bounded by the probability of failure; the last inequality is Bernoulli's inequality. No independence between successive failures is assumed. □

Counterexample 4.4 (the missing state invariant). One machine chooses a fair bit once and returns it twice. Another returns two independent fair bits. Each answer separately is fair, but the

probability of equal answers is respectively 1 and 1/2. A detector comparing answers distinguishes them. Preserving isolated marginals or a list of successful individual probes does not establish Theorem 4.2.

For any fixed experiment, let its transcript laws be μ, ν. Their total variation distance is

$$\mathrm{TV}(\mu, \nu) = \tfrac{1}{2} \sum_t |\mu(t) - \nu(t)|.$$

The maximum difference in acceptance probabilities over tests of that transcript equals TV: choose the event $\{t : \nu(t) > \mu(t)\}$. Under equal prior probabilities, optimal classification success is $(1 + \mathrm{TV})/2$. This statement fixes the experiment before optimizing the final classifier; optimizing the adaptive experiment is an additional problem.

5. What additional evidence can establish

Proposition 5.1 (observation monotonicity). If every U-experiment can be implemented as a V-experiment by discarding some observations, the optimal distinguishing advantage available in U is at most that in V.

Proof. Compose any U-detector with the assumed discard operation to obtain a V-detector with the same acceptance probabilities. Take suprema. Equivalently, pushforward along a fixed observation map cannot increase TV, by the triangle inequality on each fibre. □

There is no general theorem that a richer context strictly distinguishes more systems. It can add a redundant or attacker-controlled observation.

Proposition 5.2 (a calibrated timing separator). Suppose an independent experiment reports duration in $[b - \eta, b + \eta]$ on the native machine and in $[b + \delta - \eta, b + \delta + \eta]$ on the virtual machine, with $\delta > 2\eta$. The threshold $b + \delta/2$ perfectly separates those two modeled populations.

Proof. The native upper endpoint is below the threshold and the virtual lower endpoint is above it. □

This elementary result exposes the necessary hypotheses: independent measurement, calibration, bounded disturbance, and a positive lower bound on overhead. It does not prove that any real Blue Pill instance supplies them. With overlapping distributions, statistical error must be quantified. A timestamp delivered through the suspect controller is not independent merely because it is called a clock. Historical detection work motivates considering such discrepancies [4]; it supplies no universal threshold for today's machines.

Proposition 5.3 (trusted-evidence separator). Let an independently verified predicate A have probabilities at least $1 - \beta$ on the clean population and at most α on the compromised population. Accepting the clean hypothesis exactly when A holds has false rejection at most β, false acceptance at most α, and equal-prior success at least $1 - (\alpha + \beta)/2$.

Proof. The two error probabilities are the complements or values of the stated probabilities; average them under equal priors. □

For an attestation interpretation, authenticity, freshness, relevant measurement coverage, and binding to the machine and session being tested are hypotheses supporting those probability bounds. A genuine old measurement does not by itself certify current runtime authority. No concrete attestation system is proved sound here.

6. Security properties and observation fibres

Let W be a set of possible complete realizations of an experiment and $o : W \to O$ its observation map. A world can include the whole interaction and random choices; consequently this construction does not assume each instantaneous state determines future observations.

Theorem 6.1 (factorization criterion). A Boolean security property $P : W \to \{0, 1\}$ can be decided exactly from observations on realizable outputs if and only if it is constant on every fibre of o. Equivalently, there is a function $p : o(W) \to \{0, 1\}$ such that $P = p \circ o$.

Proof. A factorization gives $P(w) = p(o(w)) = p(o(w')) = P(w')$ whenever $o(w) = o(w')$. Conversely define $p(y)$ to be the common value on the nonempty fibre $o^{-1}(y)$; fibre constancy makes this unique and well-defined. □

If a clean and a compromised world have the same observation, no perfect detector of that property exists on that interface. This is an information-theoretic obstruction, even for an unbounded detector; it is not an application of Rice's theorem. Conversely, fibre constancy alone need not yield an effectively computable detector on infinite representations. Statistical distinguishability of distributions is also different from exact separation of all worlds.

For $C \subseteq W$ the clean worlds, define

$$K_U(y) = \{w \in C : o_U(w) = y\}.$$

An observation refutes the clean model exactly when $K_U(y) = \varnothing$. A nonempty fibre means only that a clean explanation remains. It does not assert that the actual world is clean, probable, or computationally reconstructible. Claims of diagnosis are relative to the adequacy of C.

7. Composing simulation witnesses

Restrict in this section to finite deterministic total machines over fixed finite Q, Y. Let Mach have objects $T_M : S_M \times Q \to Y \times S_M$ and morphisms $f : M \to N$ satisfying

$$T_N(f(s), q) = (\mathrm{id}_Y \times f)T_M(s, q).$$

These are coalgebra homomorphisms for $F(X) = (Y \times X)^Q$. They preserve output and successor state, not physical authority. A witness from M to N is a span of homomorphisms

$$M \xleftarrow{p} R \xrightarrow{q} N.$$

For selected initial states, require an r_0 projecting to both. Each witness state supplies related realizations that answer every request identically. A 2-cell is a homomorphism between apex machines commuting with both legs. It need not be invertible.

Theorem 7.1 (witness composition and coherence). These spans and 2-cells form a bicategory. The composite of $M \leftarrow R \to N$ and $N \leftarrow R' \to L$ has apex $R \times_N R'$. Initial witnesses with the same intermediate initial state compose. Every composite preserves all finite response traces of those states.

Proof. The pullback state set consists of pairs (r, r') with equal images in N. Given an input, the two outputs are equal because both are the output in that same intermediate state. Apply the two

transitions componentwise. Their successors again have equal images, so the pair transition is defined and unique. This construction has the pullback universal property, since a map into it is exactly a compatible pair of homomorphisms.

Use identity spans as units and pullback-induced maps as horizontal composites of 2-cells. Rebracketing $((r, r'), r'') \mapsto (r, (r', r''))$ gives the associator; projection off an identity coordinate gives the unitors. They are homomorphisms because transitions are componentwise. Every pentagon composite sends a nested quadruple to the same ordered quadruple with the target bracketing; the triangle agrees by deleting the identity coordinate. Thus the coherence diagrams commute. The paired initial witness exists by the equality assumption. Induction on the request word proves equal traces. □

The theorem composes *certificates of matching behavior*. It does not make installation a reversible operation or identify the physical machines as isomorphic. We do not claim ordinary set-theoretic pullbacks compose arbitrary stochastic couplings: probabilistic composition requires its own coupling construction. Keeping this restriction prevents a false categorical generalization.

8. Presheaf and higher-topos semantics of evidence

8.1 Local compatibility and global refutation

Fix a finite set of jointly recorded channels B. Let $\mathcal{C} = \mathcal{P}(B)$, with arrows given by inclusion, and put

$$\mathcal{E}(U) = \prod_{c \in U} O_c, \qquad \mathcal{F}(U) = \{o_U(w) : w \in C\} \subseteq \mathcal{E}(U).$$

Restriction forgets channels. These are presheaves on $\mathcal{C}$. With covers given by unions, $\mathcal{E}$ is a sheaf: compatible tuples glue uniquely coordinatewise. The clean-realizable subpresheaf $\mathcal{F}$ need not be a sheaf. These channels are jointly modeled classical records; no quantum measurement interpretation is assumed.

Theorem 8.1 (finite compatibility criterion). If $U = \bigcup_{i=1}^{k} U_i$ and $y \in \mathcal{E}(U)$, then

$$K_U(y) = \bigcap_{i=1}^{k} K_{U_i}(y|_{U_i}).$$

Hence all local observations can admit clean explanations while their combined observation refutes every clean world.

Proof. A world has tuple y on U exactly when it has the indicated tuple on each U_i, because every coordinate belongs to at least one member of the cover. This proves the set equality. Nonempty sets need not have nonempty intersection. □

Example 8.2. Let clean channel tuples be $C = \{(0, 0), (1, 1)\}$ and observe $(0, 1)$. Channel one alone is explained by $(0, 0)$; channel two alone by $(1, 1)$. The single-channel sections agree on the empty overlap but have no clean global extension. Thus $\mathcal{F}$ fails descent. Replacing it by its sheafification on this cover topology would admit the glued tuple and destroy precisely this global constraint. The presheaf of realizability must therefore not be sheafified merely for aesthetic reasons.

This elementary constraint example is not a physical attack, a proof of quantum contextuality, or a universal detection technique. Sheaf-theoretic contextuality offers a related local/global vocabulary

[6], but the experiment and hypotheses here are different. If one controller supplies both records coherently, the inconsistency may never arise.

8.2 Realization groupoids and homotopy fibres

The category $\widehat{\mathcal{C}} = \mathrm{Fun}(\mathcal{C}^{op}, \mathsf{Set})$ is a presheaf topos. Its space-valued counterpart

$$\mathcal{X} = \mathrm{Fun}(\mathcal{C}^{op}, \mathcal{S})$$

is a presheaf infinity-topos [7]. This is a standard ambient construction, not a claim that a processor is itself an infinity-topos. The discrete presheaves above embed as 0-truncated objects of $\mathcal{X}$.

To retain explicit symmetries, let $\mathcal{W}$ be a small groupoid of complete realizations and declared isomorphisms of representation. Require every observation and the clean predicate to be invariant under these isomorphisms. Let $W = N\mathcal{W}$ be its nerve, regarded as a constant presheaf, and let $o : W \to \mathcal{E}$ be the observation map. For a context U and observed tuple y, the compatible realization space is

$$\mathrm{hofib}_y(o_U) \simeq W \times^h_{\mathcal{E}(U)} \{y\}.$$

Since $\mathcal{E}(U)$ is discrete, this is the nerve of the full subgroupoid on realizations producing y. It can have several connected components or nontrivial automorphism groups. Merely adding an extra hidden controller does not make two realizations isomorphic: the declared arrows must justify that identification.

Proposition 8.3 (a non-discrete fibre). A realization with a representation symmetry group G and a single invariant observation contributes the space BG to the corresponding fibre. Its set of connected components is a singleton, while its fundamental group is G.

Proof. The one-object groupoid with endomorphisms G has nerve BG. The map to the one-point observation space has that entire nerve as its homotopy fibre. By the standard classifying-space description, its fundamental group is G and it is connected. □

Here symmetry concerns representation, for example renaming an abstract hidden state, rather than a physically performed attack transition. The example explains what the higher language retains beyond a set of possible observations. Our construction is groupoid-valued and hence 1-truncated; no nontrivial higher homotopy groups are asserted. Higher proof objects would require additional specified structure.

Let $C \subseteq W$ now denote the nerve of the full subgroupoid of clean realizations, as a constant presheaf. The internal image predicate $\mathrm{im}(C \to \mathcal{E})$ states that a clean realization exists. The actual compatible-fibre object retains which realizations and symmetries witness that proposition. Existential truncation discards those witnesses. Consequently, "has a clean explanation" and "is clean" are different semantic assertions even inside the topos.

9. Executable finite evidence

The accompanying standard-library Python program checks a finite abstraction; it does not implement virtualization. Its native machine has a bit b and requests `read` and `flip`. `read` returns b; `flip` changes b to $1 - b$ and returns the new value. A virtual model adds a hidden parity bit h that

toggles after every request but leaves the guest response unchanged. The relation is equality of b, independently of h.

The program exhaustively checks all related state/input pairs, compares traces for every request word of length at most eight, and constructs the pullback for two successive hidden-bit extensions. Separate cases expose the hidden bit, expose an extra unit of modeled cost, compare correlated with independent fair responses, and compute the compatibility example in Section 8. Approximate-discrepancy calculations use rational arithmetic. The generated report records exact counts and values.

The run checked 8 local simulation obligations, 511 input words and 2,044 related-state trace comparisons, an 8-state pullback with 32 projection obligations, 16 tuple-rebracketing cases, 576 compatibility identities, and 36 rational approximate-coupling examples. The correlation example has TV 1/2; the exposed-state and independent-cost examples have TV 1. These checks test examples and prevent elementary transcription errors; the mathematical proofs establish the quantified results. No hardware experiments, rootkit installation, benchmark, processor verification, proof-assistant kernel check, or independent referee assessment was performed. The cost example is stipulated, not measured. The model does not verify a real controller's response function, resource footprint, interrupt behavior, multicore operation, or security of its environment.

10. Consequences and limits

Blue Pill's semantic signature is a change of authority under an attempted preservation of guest computation. Correct answers are compatible with unauthorized mediation. Exact observational equivalence therefore depends on both a transition invariant and a deliberately stated observation boundary. More evidence helps only when it provides an independent constraint or a genuinely richer experiment.

The formal development yields concrete proof obligations for defensive analysis: specify the clean population and allowed observers; include termination and measured cost; establish a simulation invariant if claiming equivalence; establish separation or error bounds if claiming detection; and justify why each trusted channel lies outside the compromised authority. A universal assertion of invisibility or detectability cannot replace these obligations.

The bicategory composes behavioral witnesses, while the presheaf infinity-topos organizes compatible realizations across observation contexts. Neither construction creates physical information that an observer lacks. The primary unresolved implementation problem is adequacy: establishing that a concrete hardware/controller system satisfies the abstract transition and measurement assumptions.

Authorship and reproducibility

The responsible human submitter has chosen the public byline **Anonymous**. OpenAI ChatGPT/Codex assisted with research, mathematical drafting, proofs, finite-model code, reference checks, and typesetting. Responsibility for publication claims remains with the submitter; no independent review is implied. The package contains this manuscript, a typeset PDF, executable checks, their report, and build instructions. The code operates only on synthetic finite data and has no privileged or networking operations. Version 1.0, 28 October 2026.

References

1. Joanna Rutkowska. *Subverting Vista Kernel for Fun and Profit.* SyScan and Black Hat, 2006, especially slides 25, 31-32 and 37-39. Original presentation.
2. Samuel T. King, Peter M. Chen, Yi-Min Wang, Chad Verbowski, Helen J. Wang, and Jacob R. Lorch. *SubVirt: Implementing malware with virtual machines.* IEEE Symposium on Security and Privacy, 2006, pp. 314-327. Author-hosted paper.
3. Gerald J. Popek and Robert P. Goldberg. *Formal requirements for virtualizable third generation architectures.* Communications of the ACM 17(7), 1974, pp. 412-421. Paper.
4. Tal Garfinkel, Keith Adams, Andrew Warfield, and Jason Franklin. *Compatibility Is Not Transparency: VMM Detection Myths and Realities.* HotOS XI, USENIX, 2007. Publisher record and paper.
5. Gordon D. Plotkin. *A Structural Approach to Operational Semantics.* DAIMI FN-19, Aarhus University, 1981. Author's publication index, entry 17.
6. Samson Abramsky and Adam Brandenburger. *The Sheaf-Theoretic Structure of Non-Locality and Contextuality.* New Journal of Physics 13, 113036, 2011. Author manuscript.
7. Jacob Lurie. *Higher Topos Theory.* Annals of Mathematics Studies 170, Princeton University Press, 2009, Sections 5.1 and 6.1-6.2. Author's electronic edition.

All reference links were inspected on 28 October 2026. The Wikipedia article supplied with the request served as an orientation aid; technical claims rely on the primary sources above. Historical platform claims are not presented as assessments of current Windows or processor security.