A CUBICAL FORMALISATION OF TOPOS CAUSAL MODELS: INTERVENTION, SHEAF GLUING, AND THE INTUITIONISTIC DO-CALCULUS

KAREN SARGSYAN

Institute of Chemistry, Academia Sinica, Taipei, Taiwan
e-mail address: karen.sarkisyan@gmail.com

Abstract. Topos causal models recast causal inference inside a topos: a causal world is a presheaf, an intervention is a characteristic map into the subobject classifier, and reasoning is carried out in the intuitionistic internal language. We give the first machine-checked account of this 1-topos core, in Cubical Agda, over a previously verified probability monad and do-calculus. We build the classifier of sieves and realise the intervention $\mathrm{do}(X := x_0)$ as a characteristic map with its classification theorem; prove the sheaf gluing of independent mechanisms, which the source asserts but never proves; and machine-check the Kripke-Joyal forcing clauses of the internal language. In the modal layer we find and repair a gap: the three standard Lawvere-Tierney axioms do not force a closure operator. With the missing law restored, we exhibit the double-negation topology as a concrete instance and show that interventions and Pearl's rules are stable under every topology. Transportability of a counterfactual across a cover of regimes then coincides with this j-stability, understood as invariance across the cover. We further add a phenomenon the programme does not consider: a machine-checked contextuality obstruction, where pairwise-consistent local data admit no global model. The development assumes no axioms and typechecks under Agda's `--safe` flag, with the ordered field discharged concretely at $\mathbb{Q}$; the scope is the presheaf (1-topos) fragment, with type-level sheafification and the directed lift left to future work.

1. Introduction

A structural causal model is a system of equations: each variable is a function of its causes and an independent noise term, and an intervention replaces one equation by a constant assignment. Mahadevan's *topos causal models* (TCMs) [Mah25c, Mah25b] recast this picture categorically. A causal world is a presheaf over a base category of contexts; a mechanism is a morphism of presheaves; and an intervention is no longer an edit to a syntax tree but a *characteristic map into the subobject classifier* Ω, the object of truth values of the topos. Because a presheaf topos has an intuitionistic (Heyting, not Boolean) internal logic, causal statements live in that internal language, interpreted by Kripke-Joyal forcing. A companion line of Mahadevan's work [Mah25a] adds an *intuitionistic j-do-calculus*, in which a Lawvere-Tierney topology $j : \Omega \to \Omega$ acts as a modality picking out the causal statements that are stable under localisation.

The appeal of this reformulation is that several pieces of causal machinery become standard topos-theoretic universal properties: intervention is the universal property of Ω; the assembly of independent local mechanisms into a global one is sheaf gluing; and

counterfactual reasoning is forcing in the internal logic. Two further phenomena live on the same machinery: the j-do-calculus modality that isolates the causal facts invariant across regimes, and — new to this account — a cohomological obstruction, the failure of pairwise-consistent local models to assemble into a global one. The framework is, however, developed on paper, and some of its central claims are stated rather than proved.

This paper supplies that verification: a machine-checked account of the 1-topos core in Cubical Agda [CCHM17, VMA19]. It is a direct sequel to a companion development [Sar26], which verifies a probability monad as a higher inductive type together with Pearl's do-calculus rules and the soundness of d-separation (the graphical criterion for which conditional independences a DAG entails); that development supplies the mechanisms (probability kernels) that the present one organises into a topos. Throughout we work in the presheaf topos $\mathbf{PSh}\,\mathcal{C} = \mathbf{Set}^{\mathcal{C}^{\mathrm{op}}}$ over an abstract base category $\mathcal{C}$ of contexts, exactly the setting in which Mahadevan's concrete constructions live.

Contributions. Our account is faithful to the source where the source is precise, and in several places it does more than transcribe: it proves a central claim the source leaves informal, repairs a gap in the standard modal axioms, and adds an obstruction phenomenon the programme does not consider.

- *The intervention classifier* (Section 3). We build Ω as the presheaf of sieves and realise $\mathrm{do}(X := x_0)$ as the characteristic map $\chi : X \Rightarrow \Omega$ of the subobject "$X = x_0$". We prove the classification theorem $\chi_c(b) = \top \iff b = x_0\, c$ in both directions, and connect χ to the operational, kernel-level intervention of the companion paper. We also prove the general universal property — every restriction-closed sub-presheaf is classified by a *unique* map into Ω — so Ω is the subobject classifier and the intervention is one instance. A small design point falls out: a topos-internal intervention must fix X to a *natural* global element, not merely a context-indexed family.
- *Sheaf gluing, proved* (Section 4). Mahadevan states that local mechanisms "collate" into a unique global mechanism but proves no gluing theorem for the claim; the Grothendieck-topology vocabulary the paper introduces is never instantiated. We prove the limit form: the pullback of two mechanisms agreeing on a shared overlap, with its universal property (existence and uniqueness of the glued mechanism). In a presheaf topos this is a pointwise limit and needs neither a topology nor sheafification.
- *The internal language* (Section 5). We define Kripke-Joyal forcing and machine-check the clause for each connective and quantifier: $\top, \bot, \wedge, \vee, \Rightarrow, \forall, \exists$, together with the locality (monotonicity) property that defines a Kripke semantics.
- *A corrected and instantiated do-calculus* (Section 6). We find that the three standard Lawvere-Tierney axioms do not entail inflationarity $S \leq j\, S$ (a three-element chain satisfies all three yet collapses a truth value), so as stated they permit operators that are not closure operators. We add the missing law, so that j becomes a closure operator. We exhibit the double-negation topology $\neg\neg$ as a concrete non-degenerate instance (the source keeps j abstract), present the modality $\bigcirc = j$ as a reflective subuniverse, and prove that the intervention classifier and Pearl's Rules 1, 2 and 3 are j-closed for every topology, and hence stable under any sheaf localisation.
- *An obstruction* (Section 7). Two mechanisms agreeing on an overlap always glue, but three contexts can obstruct: we machine-check that pairwise-consistent local causal data may admit no global model, the sheaf-theoretic structure of contextuality and a phenomenon the programme does not consider (Section 11 positions it against prior contextuality work).

The whole development typechecks under Agda's `--safe` flag — no postulates, holes, or unsafe features anywhere in it — with the underlying ordered field discharged concretely at $\mathbb{Q}$. It is openly available at `https://github.com/karsar/cubical-topos-causal`.

Significance. Mechanising a paper framework is worthwhile in two ways here. One is assurance: the central claims are now certified rather than asserted, which matters where the framework is being driven by automation, since a verified core is what one checks machine-generated causal models against. The other is that carrying out the proofs revealed two findings that matter in practice. First, the modal layer's guarantees rest on j being a closure operator, and the stated axioms do not force this (Section 6): the invariance-and-discovery programme built on "j-stability" needs the corrected law for that stability to mean anything. Second, the gluing on which modular, map-reduce assembly of local mechanisms rests can be obstructed — over three contexts, locally consistent pieces can share no global model at all (Section 7). The practical implication is that modular causal discovery needs a global-consistency check, not merely pairwise agreement; and because the obstruction is cohomological, that check is in principle computable. We verify it on a minimal three-context example rather than on data; what this establishes is that the failure mode is real and exactly located, not that it is common.

Scope. This is the *1-topos* (presheaf) account: truth values are sieves, the modality acts on truth values, and the internal logic is the standard Kripke-Joyal one. We do not formalise type-level sheafification (the associated-sheaf reflector on the whole topos), which is a larger classical construction, nor the directed / ∞-categorical lift of the framework, which would require modal type theory outside current Cubical Agda. Section 10 states the boundary precisely.

2. The presheaf topos and its subobject classifier

We fix a base category $\mathcal{C}$ of *contexts* (Mahadevan's "regimes"): objects are contexts, morphisms are admissible context maps. A *causal world* is a presheaf $X : \mathcal{C}^{\mathrm{op}} \to \mathbf{Set}$: a family of value-sets $X(c)$ with a restriction map $X(f) : X(c) \to X(d)$ for each $f : d \to c$, functorial in f. We write composition in diagrammatic order, $f \star g$ meaning "f then g", so that restriction composes left to right. A *mechanism* is a natural transformation $X \Rightarrow Y$.

The classifier of sieves. A *sieve* on a context c is a downward-closed family of morphisms into c: a predicate S on arrows $f : d \to c$ such that $f \in S$ implies $k \star f \in S$ for every k. Sieves on c form a set, and pulling a sieve back along $h : c' \to c$ (keeping the arrows f with $f \star h \in S$) makes the assignment $c \mapsto \{\text{sieves on } c\}$ a presheaf Ω. Over the two-object poset $0 \to 1$, for instance, there are two sieves on 0 but *three* on 1: the empty sieve, the single arrow $\{0 \to 1\}$, and the maximal sieve. That middle one — true only after restriction to 0 — already shows the non-Boolean, intuitionistic character of Ω. This is the *subobject classifier*: a sub-presheaf $A \hookrightarrow B$ is named by a unique map $\chi : B \Rightarrow \Omega$, with $b \in A$ exactly when $\chi(b)$ is the maximal sieve $\top$ (all arrows). This is precisely Mahadevan's definition (his sieves are subobjects of the representable $\mathcal{C}(-, x)$, and his $\Omega(X)$ is their collection [Mah25b, Def. 8,11]); we restate it as an Agda record and verify the presheaf laws. Membership of a sieve is a proposition, which keeps $\Omega(c)$ a set — needed for Ω to be a presheaf of sets — and lets us reason about sieve equality by their membership alone.

Internal mechanisms. The mechanisms we intervene on come from the companion paper [Sar26]. There, a two-variable structural model on sets X, Y is a prior p_X on X together with a kernel $k_Y : X \to \texttt{FDist}\, Y$ into the probability monad, and the intervention $\mathrm{do}(X := x_0)$ replaces p_X by the point mass at x_0. Internally, a model on presheaves X, Y is such a model contextwise — one at each context c — and the internal intervention is the contextwise point-mass substitution. The marginal of the outcome Y is a context-indexed element of the internal distribution object. We take all of this as given; the present paper studies how interventions sit inside the topos, not how kernels compute.

3. Intervention as a characteristic map

Fix a value presheaf X and a value x_0 to which we wish to fix it. For the construction to be topos-internal, x_0 must be a *global element* $x_0 : \mathbf{1} \Rightarrow X$, that is, a value $x_0\, c \in X(c)$ at each context, natural in c: $X(f)(x_0\, c) = x_0\, d$ for every $f : d \to c$. Naturality is what makes "$X = x_0$" a sub-presheaf, and hence classifiable. A bare context-indexed family $c \mapsto x_0\, c$ that ignored restriction would not cut out a subobject, and the construction below would fail to be natural.

The classifier. Define $\chi : X \Rightarrow \Omega$ by sending $b \in X(c)$ to the sieve of arrows along which b restricts to the fixed value:

$$\chi_c(b) \;=\; \{\, f : d \to c \;\mid\; X(f)(b) = x_0\, d \,\}.$$

This is a sieve (closure under precomposition is naturality of x_0 together with functoriality of X), and χ is a mechanism: it commutes with restriction, again by functoriality. The Agda statement is the record below (where `pt d` is the global element x_0 evaluated at d); the proof obligations are the two equations just named.

```
χ-sieve : (c : Ob) → F₀ X c → Sieve c
χ-sieve c b = (λ d f → (F₁ X f b ≡ pt d) , isSetF₀ X d _ _) , χ-closed

χ : Nat X Ω
χ = (λ c b → χ-sieve c b) , χ-nat
```

Classification. The map χ classifies the subobject "$X = x_0$":

$$\chi_c(b) = \top \quad\Longleftrightarrow\quad b = x_0\, c.$$

Both directions are short. If $b = x_0\, c$ then for every $f : d \to c$ we have $X(f)(b) = X(f)(x_0\, c) = x_0\, d$ by naturality, so every arrow lies in $\chi_c(b)$ and the sieve is maximal. Conversely, if $\chi_c(b)$ is maximal then in particular the identity lies in it, which says $X(\mathrm{id})(b) = x_0\, c$, i.e. $b = x_0\, c$. Specialising the first direction to $b = x_0\, c$ gives that the value the intervention forces is always classified true: $\chi_c(x_0\, c) = \top$.

Bridge to the operational intervention. This χ is the characteristic map of the companion paper's intervention, not a parallel construction. The internal $\mathrm{do}(X := x_0)$ sets the prior at context c to the point mass at $x_0\, c$, whose support is the singleton $\{x_0\, c\}$. Two lemmas tie this to χ: `do-prior` shows the intervened prior is exactly that point mass, and `do-classified` shows $x_0\, c$ is the point χ classifies true. Together they exhibit χ as the topos-internal name of the intervention the companion paper computes with, the correspondence the source asserts.

The classifier in general. The intervention is one instance of a general fact we also prove: Ω is the subobject classifier. For *any* restriction-closed predicate P on a presheaf B — a sub-presheaf — the map $\chi^P_c(b) = \{\, f : d \to c \mid B(f)(b) \in P \,\}$ is natural, its $\top$-fibre is exactly P in both directions, and it is the *unique* such map: any natural $\chi' : B \Rightarrow \Omega$ whose $\top$-fibre is P equals χ^P. Uniqueness turns on a single fact about sieves — one containing the identity is maximal, by closure under precomposition — which forces the membership of $\chi'_c(b)$ at each arrow f to be "$B(f)(b) \in P$", the defining clause of χ^P. The value-fixing classifier above is the case $P = \{x_0\}$ (`Topos.Classifier`, with the intervention recovered as `do-classifier`).

4. Sheaf gluing of independent mechanisms

The second universal property is that local mechanisms which agree where they overlap assemble into a unique global mechanism. In Mahadevan's paper this is the motivation for invoking sheaves [Mah25b, §6]: "local functions can be collated together to yield a unique global function". But the paper proves no gluing theorem for it, and the Grothendieck topology it defines is never instantiated on a causal model. In a presheaf topos the collation is simply a *limit*, computed contextwise; no topology and no sheafification are needed.

The pullback. Let $p : X \Rightarrow Z$ and $q : Y \Rightarrow Z$ be two mechanisms with a common target Z, the "overlap" on which agreement is measured. Their pullback $X \times_Z Y$ has, at each context c, the pairs that agree on the overlap:

$$(X \times_Z Y)(c) \;=\; \{\, (x, y) \in X(c) \times Y(c) \;\mid\; p_c(x) = q_c(y) \,\}.$$

The agreement condition is a proposition (equality in the set $Z(c)$), so the fibre is a set and restriction preserves it: restricting an agreeing pair gives an agreeing pair, using naturality of p and q. This makes $X \times_Z Y$ a presheaf with two projections π_1, π_2 to X and Y, and the square $p \circ \pi_1 = q \circ \pi_2$ commutes by construction: the commuting witness is just the stored agreement.

The universal property. A *cone* is a context A with two mechanisms $a : A \Rightarrow X$, $b : A \Rightarrow Y$ that agree on Z, i.e. $p \circ a = q \circ b$ pointwise. It collates to a *unique* global mechanism $\langle a, b \rangle : A \Rightarrow X \times_Z Y$:

```
glue      : Nat A Pullback
glue-π₁   : (c)(z) → fst π₁ c (fst glue c z) ≡ fst a c z
glue-π₂   : (c)(z) → fst π₂ c (fst glue c z) ≡ fst b c z
glue-uniq : (u : Nat A Pullback)
          → (π₁ ∘ u ≡ a) → (π₂ ∘ u ≡ b) → u ≡ glue
```

Existence is immediate: send z to the agreeing pair $(a_c(z), b_c(z))$, whose agreement is the cone's hypothesis; naturality is naturality of a and b on the two components, with the agreement coherence automatic because it is a proposition. The projection equations hold by definition. Uniqueness is the only step with content: a competing mechanism u whose projections are a and b must equal $\langle a, b\rangle$, because a pair is determined by its two components and the agreement witness is again propositional. This is the precise form of the informal statement that mechanisms agreeing on an overlap glue to a unique global mechanism.

5. The internal language

A *predicate* on a presheaf B is a mechanism $\varphi : B \Rightarrow \Omega$, i.e. a B-indexed family of truth values, equivalently a subobject of B. Kripke-Joyal semantics interprets the connectives of the internal language as operations on Ω and determines, for each context c and each generalised element $a \in B(c)$, whether c *forces* φ at a.

Forcing. We define forcing of a truth value S at a context c by "the identity of c lies in S":

$$c \Vdash S \;:=\; (\mathrm{id}_c \in S).$$

Since a sieve containing the identity is the maximal sieve, this agrees with "$S = \top$"; the identity formulation is prop-valued and keeps the clauses at the value level. Forcing a predicate at an element is forcing its truth value: $c \Vdash_\varphi a := c \Vdash \varphi_c(a)$.

Locality. The defining property of a Kripke semantics is that forcing is stable under restriction: if $c \Vdash S$ and $f : d \to c$, then $d \Vdash S{\cdot}f$ (the pullback of S). This holds because a sieve is downward closed. At the level of predicates it says: if a satisfies φ at c, so does every restriction of a; this internal-logic form is proved from naturality of φ.

The clauses. We verify the Mac Lane-Moerdijk clauses [MM92, VI.6] specialised to a presheaf topos. Conjunction and disjunction are local at the context, as in any Kripke model:

$$c \Vdash S \wedge T \iff (c \Vdash S) \text{ and } (c \Vdash T), \qquad c \Vdash S \vee T \iff \|(c \Vdash S) + (c \Vdash T)\|.$$

Both hold definitionally once the sieve operations are pointwise meet and join; the join carries a propositional truncation, which is why disjunction is recorded up to $\|-\|$. Implication is the characteristic Kripke-Joyal clause, quantifying over all future contexts:

$$c \Vdash S \Rightarrow T \iff \text{for all } f : d \to c,\ (d \Vdash S{\cdot}f) \text{ implies } (d \Vdash T{\cdot}f).$$

Falsity is forced nowhere, $c \Vdash \bot$ is empty. For the quantifiers, over a value presheaf D, a predicate φ on $B \times D$ yields predicates $\forall_D\varphi$ and $\exists_D\varphi$ on B, with clauses

$$c \Vdash \forall_D\varphi\,(a) \iff \forall (f : d \to c)\,(t \in D(d)),\ d \Vdash_\varphi (a{\cdot}f,\, t),$$

$$c \Vdash \exists_D\varphi\,(a) \iff \Big\| \textstyle\sum_{t \in D(c)}\ c \Vdash_\varphi (a, t) \Big\|.$$

The universal quantifier ranges over all restrictions and all elements; the existential is local-existential at the context (the image of the projection), in the presheaf-Kripke style. Each quantifier object carries full sieve structure (downward closure and naturality), which we verify; most of the formalisation effort lies there rather than in the clauses themselves.

6. A Lawvere-Tierney do-calculus

The modal layer interprets a *Lawvere-Tierney topology* $j : \Omega \to \Omega$ as a causal modality: $j\,S$ is the closure of S, and the j-closed truth values are those stable under the localisation j names.

A missing axiom. A Lawvere-Tierney topology is usually given by three equations: $j\top = \top$, $j(S \wedge T) = j\,S \wedge j\,T$, and $j(j\,S) = j\,S$ (Mahadevan's Definition 10 [Mah25a] states exactly these). These make j idempotent, meet-preserving and truth-preserving, but they do not entail that j is *inflationary*, $S \leq j\,S$, which is what makes j a closure operator and supplies the unit of the modality. A three-element chain $\bot < a < \top$ with $j(a) = \bot$ (and j fixing $\bot, \top$) satisfies all three axioms yet collapses a below itself; we machine-check this counter-model (module `InflationarityIndependence`), so the independence of inflationarity is itself verified, not merely asserted. Hence the three axioms, as stated, admit operators that are not closure operators. We add inflationarity as a fourth law, obtaining a closure operator — inflationary, idempotent, and meet-preserving (a *nucleus*); both of our instances satisfy it.

```
record LawvereTierney where field
  jop    : (c) → Sieve c → Sieve c
  jnat   : IsNat Ω Ω jop          -- j is a mechanism Ω ⇒ Ω
  j-⊤    : jop c ⊤ ≡ ⊤
  j-∧    : jop c (S ∧ T) ≡ jop c S ∧ jop c T
  j-idem : jop c (jop c S) ≡ jop c S
  j-infl : (S d f) → f ∈ S → f ∈ jop c S    -- the added law
```

The modality as a reflector. Writing $\bigcirc = j$, the data above make $\bigcirc$ a *reflective* modality on truth values — one that sends each truth value to the least j-closed value above it (its reflector). The unit $\eta : S \leq \bigcirc S$ is inflationarity, $\bigcirc S$ is modal (j-closed) by idempotence, and $\bigcirc$ is monotone, which we derive from meet-preservation rather than assume. The reflector universal property follows: for a modal T,

$$S \leq T \quad \Longleftrightarrow \quad \bigcirc S \leq T.$$

So the j-closed truth values form a reflective sub-poset of Ω, with $\bigcirc$ the reflector; this is the propositional shadow of sheafification.

A concrete topology. The identity is a topology but a vacuous one. The source keeps j abstract, naming only the trivial (Boolean) topology as a special case [Mah25a]; we provide the double-negation topology $\neg\neg$ as a non-degenerate instance, whose sheaves are the classically-behaved truth values. This needs the Heyting structure on sieves — bottom, implication, negation — which we build, and then the three topology axioms plus naturality for $\neg\neg$. Two are short (truth-preservation, and idempotence via the triple-negation law $\neg\neg\neg = \neg$); meet-preservation is the substantial one, whose difficult direction $\neg\neg S \wedge \neg\neg T \leq \neg\neg(S \wedge T)$ is the intuitionistically-valid but non-trivial step. With $\neg\neg$ established, the modal results below are no longer vacuous.

Interventions are modal. The upshot is that interventions and the do-calculus survive localisation. The truth value classifying the intervened value is j-closed for every topology:

$$\chi_c(x_0\, c) \text{ is } j\text{-closed},$$

because it equals $\top$ (Section 3) and $\top$ is j-closed for every topology. Likewise each of Pearl's three rules, as verified in the companion paper, has a conclusion that is an equality of distributions; that equality is a proposition (distributions form a set), so it has an internal truth value, and that truth value is j-closed for every topology.

```
do-j-stable : (J)(c) → is-j-closed J c (χ-sieve c (pt c))
modal-rule1 : (J)(c) → is-j-closed J c (rule1-Ω c)
modal-rule2 : (J)(c) → is-j-closed J c (rule2-Ω c)
modal-rule3 : (J)(c) → is-j-closed J c (rule3-Ω c)
```

In causal terms, a do-fact and each do-calculus rule hold in the internal logic of every sheaf subtopos; they are invariant under any modality $\bigcirc$. At the double-negation topology this is concrete: the intervention sieve and Rule 1 are $\neg\neg$-closed (module `NonTrivialModal`), the Boolean-localisation instance of the generic statement.

What this does and does not show. These closure facts are true, but their proof carries no causal content, and we do not overstate them. Each rule's conclusion is a proved equality of distributions, hence a proposition with internal truth value $\top$, and $j\top = \top$ is a topology axiom; the same holds for $\chi_c(x_0\, c)$, which equals $\top$ by Section 3. Thus `modal-rule`$_\mathtt{i}$ would typecheck verbatim with the causal content deleted: the proof depends only on the conclusion's truth value being $\top$, not on the rule that produced it. Mahadevan's thesis, however, is that j *selects* the causal claims stable under localisation [Mah25a], which presupposes that some are not. The substantive statement therefore concerns *contingent* causal claims — conditional independences, which hold at some regimes and fail at others — whose truth value is a non-maximal sieve.

A contingent claim, and its stability. We exhibit one (module `ContingentCI`). Over two regimes, with values in a two-point type and rational kernels, take the claim that the Y-marginal is invariant under $\mathrm{do}(X{:=}1)$ — the very proposition `rule1-Ω` internalises, now *without* assuming $Y \perp X$. At the regime whose mechanism is the constant kernel it holds, so its truth value is $\top$; at the regime whose mechanism is the copy kernel $Y := X$, the intervention shifts the marginal from the point mass at 0 to the point mass at 1, so it *fails*. The failure is a computation in the rational weight algebra, not a postulate: the two marginals are distinct point masses, separated by their mass at 1 (`w1` $\neq$ `w0`), giving

$$\texttt{witness-non-maximal} : \Sigma\,(c : \text{regime})\ \neg\,(\texttt{ci-Ω}\, c \equiv \top).$$

This is the input the modal-rule theorems never had. The same non-maximal claim, taken through the classifier of Section 3 rather than regime-by-regime, is an internal subobject $\mathbf{1} \Rightarrow \Omega$ (module `CIObject`) — the subobject shape Section 8 gives counterfactuals. On it the stability question has content, and we answer it (module `ModalCI`): for the double-negation topology, `ci-Ω` is j-closed at *both* regimes — at one because its value is $\top$, at the other because its value is $\bot$ (the empty sieve, which is $\neg\neg$-closed), and $\bot$ is not $\top$. The verdict of a contingent independence, whether it holds or fails, therefore survives the Boolean localisation — j-stability applied to something other than $\top$. This is the j-do-calculus claim in the case where the statement is not already $\top$. The general soundness that would let `ci-Ω` range over

arbitrary graphical independences (Section 10) remains future work, and not by routine extension. The sieve structure holds only where refinement preserves the independence, and for a collider — a variable with two arrows into it, $X \to C \leftarrow Y$ — it does not: conditioning on the collision vertex C opens the path between its causes (Berkson's paradox). That failure of restriction-stability (module `CIObject`) is the obstruction the general case must address.

The modality does causal selection. The stability above is limited. A *site* is a category of contexts together with a notion of which arrow-families *cover* an object; on the two-regime *discrete* site the only covering families are trivial, so j never *collates* across contexts, and the $\top$-value of $\texttt{modal-rule}_i$ is a degeneracy of that coverage rather than a feature of the framework. The selection Mahadevan's thesis wants — j closing a claim that holds at every context *below* up to holding at the one above — needs a site with a non-trivial coverage. We give the minimal causal one (modules `InterventionSite`, `InterventionModality`): the intervention poset $\texttt{do}_0 \to \texttt{obs} \leftarrow \texttt{do}_1$, with the observational context `obs` covered by its two interventions $\{\texttt{do}_0, \texttt{do}_1\}$ — a non-maximal covering sieve, impossible on the discrete site. The induced closure is discriminating: a conditional independence holding under *both* interventions is covered up to holding observationally, while one holding under only *one* survives as a proper sieve,

$$j\,\{\texttt{do}_0, \texttt{do}_1\} \equiv \top, \qquad \neg\,(\,j\,\{\texttt{do}_0\} \equiv \top\,).$$

On an intervention coverage the modality is thus non-degenerate and performs exactly the invariance-across-interventions selection the j-do-calculus intends; the $\top$-collapse is a property of the coverage, not an obstruction in the theory. The directed / ∞-categorical lift (Section 10) would let this range over the full space of interventions; the propositional, single-variable case is what we machine-check here.

7. An obstruction: causal contextuality

The gluing of Section 4 always succeeds: two mechanisms agreeing on an overlap glue, because a pullback is a limit and a limit exists. Obstructions need three contexts. Over a cover by three measurement contexts, a family of local sections can agree pairwise on every overlap and yet admit no global section — the sheaf condition fails. This is the sheaf-theoretic structure of *contextuality* of Abramsky and Brandenburger [AB11], and in causal terms it is the statement that pairwise-consistent local causal data need not come from any single global model. For the modular, "map-reduce" assembly of mechanisms from local data that motivates the gluing, this is a structural warning. Local pieces that agree wherever they overlap can still fail to assemble, so a procedure that stitches local causal models cannot certify a global one from pairwise agreement alone. The witness below is minimal and built by hand, not drawn from data; it shows the failure mode occurs and is exactly located, not that it is common.

Specker's triangle. The minimal witness has three Boolean observables A, B, C, measured pairwise, each pair perfectly anti-correlated: the context $\{A, B\}$ supports only $A \neq B$, and likewise $\{B, C\}$ and $\{A, C\}$. Each context is locally realisable, and the family is *compatible* (no signalling): every value of a shared observable occurs in both contexts that contain it, so the single-observable marginals agree. But there is no global assignment, because $A \neq B$ and $B \neq C$ force $A = C$, against $A \neq C$. We prove its three obligations: local realisability and

the absence of a global section (bundled as a record), and compatibility, i.e. no-signalling (a separate lemma); the no-global obligation is discharged by deciding the eight Boolean assignments.

```
no-global : ¬ GlobalSection
no-global ((true , true , _) , p , _ , _) = p refl
no-global ((true , false , true) , _ , _ , r) = r refl
...                                        -- six cases in all
```

The cohomological refinement. Abramsky, Mansfield and Barbosa [AMB12] refine strong contextuality from a possibilistic fact to a *cohomological invariant*: the obstruction is a non-zero class in the first Čech cohomology of the cover. We carry this out for the triangle. The idea is that a global model is a consistent labelling of the observables, and the cohomology class measures the extent to which the locally-consistent pairwise data fail to come from any single such labelling: the class vanishes exactly when a global labelling exists. Its cover is a loop on three observables, so the relevant cohomology is H^1 of a loop with $\mathbb{Z}_2$ coefficients. A global model is a labelling of the observables — a 0-cochain — whose coboundary δ realises the observed pairwise correlations, a 1-cochain; Specker's model is the cocycle that is anti-correlated on every edge. We define δ and the holonomy (the loop sum), and machine-check both inclusions $\operatorname{im}\delta = \ker(\mathrm{hol})$ — every coboundary has zero holonomy, and every zero-holonomy cochain is a coboundary — together with surjectivity of holonomy. These are the data that determine the cohomology $H^1 = C^1/\operatorname{im}\delta \cong \mathbb{Z}_2$: we machine-check the two inclusions and surjectivity, from which the isomorphism follows, rather than constructing the quotient as a separate term. Specker's cocycle has holonomy 1, so it is the non-zero generator and is not a coboundary; the obstruction is a computed, non-trivial cohomology class. A global model is exactly a coboundary witness, so this class recovers the possibilistic no-global-section fact established above.

The triangle is then a special case of a general theorem. Over an *arbitrary* set of observables and *arbitrary* abelian coefficients, we define the coboundary and the holonomy of a closed walk, and prove the telescope: a coboundary's holonomy is the difference of its endpoints, hence trivial around a loop. It follows that any 1-cochain with non-trivial holonomy around a closed walk is not a coboundary — a non-zero class in H^1. Specker's triangle is one instance. A four-cycle is another, but there the theorem says nothing: its holonomy is trivial, and the even cycle is moreover *satisfiable* — we machine-check a global two-colouring realising every edge (`square-satisfiable`), the global section the triangle provably lacks. The invariant thus separates contextual from non-contextual covers. This is the degree-one (contextuality) fragment for covers whose contexts are pairs; the full nerve of an arbitrary cover and the higher cochain degrees are the remaining generalisation.

8. Transportability: the invariance modality internalised

The modal layer of Section 6 shows that interventions and Pearl's rules are j-stable for every topology: invariant under the localisations a Lawvere-Tierney topology names. A natural question, which Section 10 flags, is whether the internal logic yields a *transportability* criterion — whether a causal effect established in one domain stays valid in another — of the kind Bareinboim and Pearl give graphically [BP16]. We answer it here: across a cover of regimes, transportability of a counterfactual coincides with its j-stability. This development

is in the same Cubical Agda, likewise typechecks under `--safe`, and reuses the classifier, the forcing layer, and the modal layer verified above.

Counterfactuals as subobjects. Fix a base category of regimes (environments) and a presheaf W of worlds over it, a world carrying the exogenous data against which a counterfactual is evaluated. A counterfactual query is a predicate on worlds that is *restriction-stable*: if it holds of a world it holds of every restriction. Restriction-stability is exactly sieve closure, so such a predicate is a subobject $\chi : W \Rightarrow \Omega$, a natural transformation whose naturality we prove. Forcing χ at a regime is the counterfactual holding there; we call this transport to that regime. Kripke locality then yields transport-invariance immediately: a counterfactual forced at a regime is forced at every regime restricting into it.

Transportability is j-stability. The identification is as follows. A counterfactual transports to the global regime iff it is forced there, iff its internal truth value is the maximal sieve $\top$; and being $\top$ entails being j-closed for *every* Lawvere-Tierney topology, so transportability entails invariance (the converse is the soundness result below, under a density hypothesis). The middle step is restriction-stability carrying global truth down to the cover; the last is the truth-preservation axiom $j\top = \top$ — the same $\top$-collapse by which Section 6 proves do-facts and Pearl's rules j-stable. So the invariance modality the paper studies for interventions is, for counterfactuals, transportability:

$$\text{transportable} \iff j\text{-stable (invariant)}.$$

A counterfactual that holds only in an environment, not globally, is correspondingly *not* j-closed: forced at the environment but not at the global regime, exactly the failure of transport.

The probabilistic case. The identification is not special to deterministic outcomes. The finite-distribution monad `FDist` is a set, so a distributional statement is a proposition and a probabilistic counterfactual is again an internal truth value; the subobject and its forcing transfer unchanged. We carry this out for a point-mass outcome; lifting it to the companion paper's full interventional distribution, with abduction as Bayesian conditioning, is the natural next step and is not yet formalised.

Soundness. The converse direction is transport *soundness*: invariance entails transportability. We prove that for any topology in which the environment is j-dense — the closure of its sieve is $\top$ — a j-stable counterfactual that holds in the environment holds globally. The double-negation topology is dense in this sense: the negation of a counterfactual holding along the environment arrow is empty, so its double-negation closure is $\top$. This is soundness in the shape of the companion paper's d-separation soundness, and of Bareinboim and Pearl's transport soundness, both of which carry admissibility side-conditions. The matching completeness statement, an equivalence with their s-hedge criterion [BP13], we leave open; so too the discharge of the density side-condition into a single unconditional term, which the present finite regime category does not support cleanly.

Scope of this development. This is the deterministic and finite-regime account: the worlds are a presheaf of exogenous data, the cover is small, and transport is the internal-logic notion just defined, not yet proved equivalent to the complete graphical criterion. What it establishes is structural and, to our knowledge, new: counterfactuals are subobjects of a world presheaf, transport is forcing, and transportability coincides with the invariance modality of the regime cover.

9. The formalisation

The core development is thirty Cubical Agda modules layered over an eight-module companion probability layer. We indicate where the cubical setting is used essentially, which proofs were substantial, and two points that mechanisation forced.

Structure and reuse. The base modules fix a category of contexts and the presheaves over it (`Cat`, `PSh`), build the classifier of sieves Ω (`Omega`), and re-export the companion paper's finite-distribution monad as an internal distribution object $\mathrm{Dist}\,X = \texttt{FDist} \circ X$ (`InternalDist`). On top sit the five pillars — the intervention classifier (`DoClassifier`, with Ω's general universal property and uniqueness in `Classifier`), gluing (`Gluing`), forcing (`Forcing`), the modal layer (`LawvereTierney`, `Modality`, `DoubleNegation`), and the obstruction (`Contextuality`, `Cohomology`, `CechCohomology`) — together with the internalised do-calculus rules (`Rule1`–`Rule3`, `ModalRules`). The mechanisms themselves are not rebuilt: the internal kernels are the companion paper's, and the topos layer only organises them. The development assumes no axioms. The probability monad is built over an ordered field that earlier drafts left as a `postulate`; we now supply that field concretely as the cubical-library rationals $\mathbb{Q}$, each of its ring, order, and bounded-division laws a proved theorem. One subtlety makes this more than a substitution: a transparent $\mathbb{Q}$ unfolds to stuck set-quotient projections, which defeats the unifier and breaks implicit-argument inference at hundreds of downstream sites — the very reason the field had been left abstract. We recover the abstract behaviour by re-exporting the field through a single `opaque` block, which keeps the operations rigid for unification while concretely equal to $\mathbb{Q}$; every downstream proof then goes through unchanged. The whole development, probability layer and topos layer alike, typechecks under Agda's `--safe` flag, which rejects any postulate, hole, or bypass of the termination, positivity, or universe checks anywhere in the development. The account is therefore not merely postulate-free relative to an assumed interface: it is unconditionally machine-checked, with the ordered field discharged at $\mathbb{Q}$.

What the cubical setting provides. Sieve membership is valued in propositions, which keeps each $\Omega(c)$ a set and lets us prove sieve equality from membership alone; the same device makes the agreement condition in the pullback and the commuting square of the gluing proof hold automatically, because a proof of a proposition is unique. Two of the forcing clauses then hold *definitionally*: $c \Vdash S \wedge T$ and the pair of $c \Vdash S$ and $c \Vdash T$ are literally the same term, and likewise for $\vee$ against the truncated sum, so those soundness proofs are the identity. Propositional truncation $\|-\|$ carries disjunction, the local-existential quantifier, and the compatibility (no-signalling) witnesses of the contextuality model. Function extensionality and the propositional-extensionality path $\Leftrightarrow$-to-$\equiv$ give the naturality and extensionality lemmas that every universal-property argument runs through.

The substantial proofs. Three steps account for most of the difficulty. First, the double-negation topology needs the Heyting structure on sieves built by hand — bottom, implication, negation — and its substance is the intuitionistically-valid meet inclusion $\neg\neg S \wedge \neg\neg T \leq \neg\neg(S \wedge T)$: the witness restricts both double-negations along the test arrow and reassembles an $(S \wedge T)$-membership from an S- and a T-witness, reassociating every composite by hand; idempotence then follows easily from the triple-negation law $\neg\neg\neg = \neg$. Second, the quantifier objects of `Forcing` are not the one-line clauses they reduce to: each of $\forall_D$ and $\exists_D$ must be given full sieve structure — downward closure and naturality — with the substitutions along associativity and functoriality discharged explicitly, and $\exists_D$ threaded through the truncation. Third, the gluing universal property is a limit argument: existence is the agreeing pair, but uniqueness needs that a cone map is determined by its two components, with the agreement coherence automatic because it is propositional.

Two consequences of mechanisation. Formalising forced one design choice and collapsed one family of proofs. The intervention classifier typechecks only when x_0 is a *natural* global element: a bare context-indexed family does not cut out a subobject and χ fails to be natural, so the naturality side-condition of Section 3 is something the type system insisted on, not something we imposed. And the modal-stability results — that the intervention sieve and Pearl's three rules are j-closed for *every* topology — collapse to a single three-line argument: each conclusion is an equality of distributions, hence a truth value equal to $\top$, and $\top$ is fixed by every j. The contextuality obstruction, by contrast, is decided concretely: the no-global-section lemma rules out all eight Boolean assignments, and the cohomological refinement computes H^1 of the triangle's cover as $\mathbb{Z}_2$ over an arbitrary observable set and arbitrary abelian coefficients (`CechCohomology`). The Specker triangle then has holonomy 1 and fires, while a four-cycle has holonomy 0 and, correctly, does not.

10. Scope and limitations

We have formalised the 1-topos core: the classifier, gluing, the internal language, and the modal do-calculus, all over a presheaf topos. Three things are deliberately outside it.

First, the modality acts on *truth values*, not on arbitrary types. A full type-level reflector (sheafification of an arbitrary presheaf, with its descent) is a larger but still classical construction (the plus-construction, applied twice); it needs no machinery beyond what Cubical Agda offers, only more of it. We have built the propositional reflector that the do-calculus actually uses.

Second, we work over presheaves, not sheaves on an instantiated site. This matches the source, whose constructions are presheaf constructions and whose Grothendieck topology is defined but, as far as we can determine, never instantiated on a causal model. Pillars one (classifier) and three (internal logic) transfer to a sheaf topos unchanged; only the gluing pillar would acquire covers, and our pullback is the trivial-cover case of that.

Third, this is the *undirected* account, and we can now say precisely what a directed one keeps and what it costs. Causation is asymmetric, and the identity types of cubical type theory are symmetric: a causal arrow modelled as a path $x = y$ silently yields its inverse, collapsing "X causes Y" into "Y causes X". A faithful treatment instead models causal influence as a *non-invertible* arrow, and we have machine-checked the structural core of one. This is a companion development in the directed cousin of cubical type theory — the simplicial type theory of Riehl and Shulman [RS17], in which a type can carry a primitive

notion of directed arrow that, unlike a path, has no inverse — in the rzk proof assistant on the sHoTT library, verified under `rzk 0.8`, in the `directed/` directory of the artifact repository.

The translation is direct. A causal world is a *Segal type*, a type whose arrows compose; causal influence is one of those directed arrows, a `hom`, with no backward partner; transitivity of influence is composition of arrows. A mechanism — a downstream variable depending on an upstream one — is a *covariant family*, a family of types that transports forward along arrows, and interventional propagation is that forward transport. On this dictionary we machine-check the structural laws of causal reasoning. The trivial mechanism propagates a value unchanged. A causal history propagates by composition, so influence through a mediator factors into its two legs — a directed chain rule (the one place the directed core invokes an axiom: the library's extension extensionality, the directed analogue of function extensionality). A morphism between mechanisms commutes with propagation, so mechanisms transform naturally. In a complete (*Rezk*) world, one in which isomorphic variables are already equal, causal identity coincides with causal isomorphism, and a variable is pinned down by the totality of its downstream effects — an identifiability statement of Yoneda type. A counterfactual, finally, is a twin network: two runs of the world sharing a single exogenous source. The asymmetry is faithful throughout — a forward `hom` yields nothing backward, where a symmetric path would have yielded the reverse automatically.

What this directed development cannot reach is exactly the do-operator. An intervention $\mathrm{do}(X := x_0)$ mutilates the causal world, and to recover it, the directed account has to treat the universe of types as itself a directed world and extract the intervention as an arrow inside it. That step is *directed univalence*: the principle that an arrow between two types in the universe is the same data as a function between them, $\hom_{\mathcal{U}}(A, B) \simeq (A \to B)$. It is the construction of Gratzer, Weinberger and Buchholtz [GWB24], carried out in a modal extension of the theory (triangulated type theory). rzk 0.8 provides the experimental modal substrate that construction needs but does not build the universe itself, and no formalisation of directed univalence exists in any system; we can *state* the goal in rzk but not inhabit it. The two accounts thus meet at a precise boundary: the present, undirected one machine-checks the full 1-topos do-calculus and transportability, though its causal arrows are symmetric and invertible; the directed one keeps the asymmetry faithfully but stops short of the do-operator. A machine-checked directed do-calculus, joining the two, requires a directed-univalent universe.

Invariance, not transportability. The modality is easy to over-interpret, and warrants one caution. A j-closed truth value is one that holds identically across the contexts j glues: it is *invariant*. This is the "globally valid" interpretation the source intends for j [Mah25a], and it is the natural home for invariant-prediction approaches to discovery, whose target is exactly the mechanism that does not change across environments. It is *not* transportability in the sense of Bareinboim and Pearl. A causal effect can be identifiable in a target domain from source-experimental and target-observational data (hence transportable) while taking a different value in each domain, so it is not invariant and not j-closed. Our modal-stability theorems (Section 6) therefore certify invariance, the direct-transportability case, not transportability in general; the same caveat applies to the source, whose j is the same invariance modality. General transport is a marginalisation of a source factor against a target factor: the arithmetic is available (it is monadic bind in the underlying probability monad), and the selection structure separating shared from domain-specific

mechanisms is exactly naturality. Section 8 carries out the invariance case of this development, identifying transportability across a regime cover with j-stability and proving soundness; the completeness equivalence with the graphical s-hedge criterion [BP13], which is where effects taking a different value in each domain are decided, remains open.

11. Related work

The source of the framework is Mahadevan's topos causal models [Mah25c, Mah25b] and the intuitionistic j-do-calculus [Mah25a]; our development is a formalisation and, in the gluing and modal layers, a correction and completion of that work. The categorical-probability background — mechanisms as kernels, do-calculus rules as abstract theorems — is that of Markov categories [Fri20, CJ19], and our internal mechanisms are instances of the probability monad verified in the companion paper [Sar26]. The topos-theoretic constructions follow Mac Lane and Moerdijk [MM92]: the classifier of sieves, the Kripke-Joyal clauses, and Lawvere-Tierney topologies are all theirs. The treatment of a topology as a reflective modality is the propositional case of the modalities of Rijke, Shulman and Spitters [RSS20]; the directed lift we leave to future work draws on directed type theory. The topos infrastructure we build on has been formalised before, though never for causal models: Quirin and Tabareau machine-check Lawvere-Tierney topologies and sheafification in homotopy type theory [QT16], and the Cubical Agda library and its scheme-theoretic developments [ZM24] formalise presheaves and the sheaf condition. The universal property of the subobject classifier, which we verify here to ground the intervention and for self-containedness, is likewise standard and has been mechanised in general topos-theory libraries; we claim it as a completion, not a novelty. What is new here is the causal interpretation of this machinery — and, in the gluing and modal layers, its correction — not the machinery itself.

The book-length development of the framework [Mah26] includes a Lean 4 formalisation of its own. That development and ours do not overlap in substance. It is declaration-level and, by its authors' own verification map, partial: the topos results record only a finite-limit core (the coalgebra topos is marked as future work), and the Grothendieck–Lawvere-Tierney correspondence is formalised in a single direction. It is classical, built on Mathlib, and does not aim at Agda's `--safe` discipline. It does not touch the intervention classifier's classification theorem, the gluing universal property, or the modal-stability results. Where it does address related material — it states a Kripke-Joyal semantics declaration — our contribution is not to be first to name the clauses but to prove all of them, constructively and axiom-free, for the causal setting.

A complementary line studies do-calculus *derivations* combinatorially rather than semantically. Yvernes, Devijver, Clausel and Gaussier [YDCG26] organise the rewrites of Pearl's rules into a derivation graph, presupposing the soundness our companion paper mechanises, and show that do-calculus-equivalent estimands can differ in statistical efficiency. That distinction lives below the population layer any faithful semantics identifies, and is hence orthogonal to what a truth value in Ω can see. Their open problem, extending the derivation graph with probabilistic rewritings that trivial identities render infinite, is a sheafification: precisely the type-level modality that Section 10 defers to future work.

The obstruction of Section 7 belongs to a well-developed line that we do not claim to originate. That contextuality is an obstruction to gluing local sections to a global one, with the obstruction measured by Čech cohomology, is due to Abramsky and Brandenburger [AB11] and Abramsky, Mansfield and Barbosa [AMB12], in quantum foundations. Gogioso and

Pinzani [GP23] carry the sheaf framework to causality and identify a "causally-induced contextuality", but in the operational, indefinite-causal-order setting rather than for classical structural causal models. For classical causal models the closest work is very recent: global counterfactuals can be obstructed by the homology of the causal graph [WXL26], formalised there through cellular sheaves over optimal-transport spaces, a route different from ours and not machine-checked. Quantum no-go inequalities have been machine-checked before — the CHSH inequality and Tsirelson's bound in Isabelle/HOL and Lean, for instance — but those are bounds on correlations, not the sheaf-theoretic obstruction we treat here. Against this background, our contribution is a machine-checked realisation of the obstruction in its sheaf-theoretic form (compatible local sections with no global section), and a topos-internal one: to our knowledge it is the first such global-section obstruction verified in a proof assistant. It lives on the same presheaf gluing as the rest of the development and connects Mahadevan's j-stability programme, where obstructions do not appear, to the contextuality line.

12. Conclusion

Topos causal models recast intervention, mechanism assembly and counterfactual reasoning as three topos-theoretic universal properties. We have machine-checked the 1-topos core of this picture in Cubical Agda over a verified probability monad: the intervention classifier with its classification theorem (an instance of Ω's universal property, which we also verify), sheaf gluing as a pullback with its universal property, the Kripke-Joyal forcing clauses for every connective and quantifier, and a Lawvere-Tierney do-calculus in which interventions and Pearl's rules are stable under every topology. Relative to the source, the development proves the gluing property that was only asserted and corrects the Lawvere-Tierney axioms by adding the missing inflationarity law; the double-negation topology makes the modal layer concrete where the source stays abstract. The development is machine-checked under Agda's `--safe` flag, with the ordered field discharged concretely at $\mathbb{Q}$. A directed account, in which the asymmetry of causation is primitive, remains future work.

References

[AB11] Samson Abramsky and Adam Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics*, 13(11):113036, 2011.

[AMB12] Samson Abramsky, Shane Mansfield, and Rui Soares Barbosa. The cohomology of non-locality and contextuality. In *Proceedings 8th International Workshop on Quantum Physics and Logic (QPL 2011)*, volume 95 of *EPTCS*, pages 1–14, 2012.

[BP13] Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013. doi:10.1515/jci-2012-0004.

[BP16] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi:10.1073/pnas.1510507113.

[CCHM17] Cyril Cohen, Thierry Coquand, Simon Huber, and Anders Mörtberg. Cubical type theory: A constructive interpretation of the univalence axiom. *Journal of Applied Logics (IfCoLog)*, 4(10):3127–3170, 2017. Conference version: TYPES 2015, LIPIcs 69, 5:1–5:34, 2018.

[CJ19] Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019. doi:10.1017/S0960129518000488.

[Fri20] Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020. doi:10.1016/j.aim.2020.107239.

[GP23] Stefano Gogioso and Nicola Pinzani. The topology of causality. *arXiv preprint arXiv:2303.07148*, 2023.

[GWB24] Daniel Gratzer, Jonathan Weinberger, and Ulrik Buchholtz. Directed univalence in simplicial homotopy type theory. *arXiv preprint arXiv:2407.09146*, 2024.

[Mah25a] Sridhar Mahadevan. Intuitionistic j-do-calculus in topos causal models. *arXiv preprint arXiv:2510.17944*, 2025.

[Mah25b] Sridhar Mahadevan. Topos causal models, 2025. `arXiv:2508.08295`.

[Mah25c] Sridhar Mahadevan. Universal causal inference in a topos. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

[Mah26] Sridhar Mahadevan. *Categories for AGI*. Draft manuscript, 2026. Lean 4 formalization companion at `https://github.com/sridharmahadevan/catagi`.

[MM92] Saunders Mac Lane and Ieke Moerdijk. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Universitext. Springer, 1992.

[QT16] Kevin Quirin and Nicolas Tabareau. Lawvere-tierney sheafification in homotopy type theory. *Journal of Formalized Reasoning*, 9(2):131–161, 2016. `doi:10.6092/issn.1972-5787/6232`.

[RS17] Emily Riehl and Michael Shulman. A type theory for synthetic ∞-categories. *Higher Structures*, 1(1):147–224, 2017.

[RSS20] Egbert Rijke, Michael Shulman, and Bas Spitters. Modalities in homotopy type theory. *Logical Methods in Computer Science*, 16(1), 2020.

[Sar26] Karen Sargsyan. A cubical formalisation of conditional independence, Bayesian conditioning, and Pearl's d-separation soundness, 2026. Companion paper. URL: `https://arxiv.org/abs/2606.20351`, `arXiv:2606.20351`.

[VMA19] Andrea Vezzosi, Anders Mörtberg, and Andreas Abel. Cubical Agda: A dependently typed programming language with univalence and higher inductive types. In *Proceedings of the ACM on Programming Languages (ICFP)*, volume 3, pages 87:1–87:29, 2019. `doi:10.1145/3341691`.

[WXL26] Rui Wu, Hong Xie, and Yongjun Li. Cohomological obstructions to global counterfactuals: A sheaf-theoretic foundation for generative causal models, 2026. `arXiv:2603.17384`.

[YDCG26] Clément Yvernes, Emilie Devijver, Marianne Clausel, and Eric Gaussier. Unveiling the structure of do-calculus reasoning via derivation graphs. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*, 2026. URL: `https://arxiv.org/abs/2606.03719`.

[ZM24] Max Zeuner and Anders Mörtberg. A univalent formalization of constructive affine schemes. *arXiv preprint arXiv:2212.02902*, 2024.