

Riskified: Approval Guarantees and Who Bears the Harm

Every automated approval decision creates a possible loser, and every fraud stack eventually has to answer who absorbs the cost when a decision turns out wrong. Riskified, as publicly described, answers that question commercially: it decides whether to approve an ecommerce order and takes on chargeback liability for orders it approves, so the merchant is not left carrying fraud losses on approved traffic. Chapter 8 of U.S. Provisional Application No. 64/117,812 answers a structurally different version of the same question for autonomous agents. There, an obligation created by an agent's own admitted deviation is designated by the class of party that deviation harmed, and where an identified counterparty resolves, that obligation is discharged only by that counterparty's own acceptance, never by the agent's self-assessment and never by elapsed time.

An Approval Is a Bet Someone Has to Cover

A checkout decision engine looks at an order and says yes or no in a fraction of a second. Both answers can be wrong, and the two kinds of wrong land on different people. A yes on a fraudulent order produces a chargeback that someone pays for. A no on a legitimate order produces a lost sale for the merchant and an unexplained

rejection for a shopper who did nothing wrong. That these errors occur is not in dispute. The question worth asking is architectural: what structure determines who bears the error, and what event counts as the matter being settled?

For commerce, that structure is contractual. Liability follows agreements among merchants, processors, card networks, and vendors, enforced by money moving between corporate entities under terms negotiated before the transaction occurred. For autonomous software agents acting under delegated authority, no comparable structure exists by default. An agent that harms a counterparty can, if nothing prevents it, decide on its own that the harm has been addressed. It logs a correction, marks the incident closed, and proceeds. Its self-assessment becomes the settlement. Chapter 8 of U.S. Provisional Application No. 64/117,812 addresses that gap, and setting it beside a mature commercial approach to the same underlying question makes the difference easier to see.

Riskified in Its Own Terms

Riskified, as publicly described, is an ecommerce fraud and risk management company serving online merchants. Its central product decides, at the moment of checkout, whether to approve or decline an order, and its publicly stated position emphasizes approving more good orders rather than simply blocking more bad ones. The company has publicly positioned itself around the cost of false declines, on the reasoning that a legitimate customer turned away is a real loss, and that a system tuned only to minimize fraud produces that loss continually.

The distinguishing element is the guarantee. As publicly described, Riskified assumes chargeback liability for the orders it approves, reimbursing the merchant when an approved order turns out to be fraudulent. That structure does something more than shift a cost. It aligns the vendor's incentives with the merchant's, since a vendor that

pays for its own false approvals cannot improve its own economics by declining aggressively. It also converts an unpredictable operational exposure into a priced one, which is a genuine benefit to a merchant trying to plan.

Public materials also describe a broader set of concerns familiar to anyone working in the category: account takeover, policy abuse of returns and promotions, and chargeback dispute handling.

Someone who builds this product should recognize the account above as fair. It is a well-engineered answer to a real commercial problem, and it is a different problem from the one the filed chapter is solving. Nothing here suggests otherwise.

What the Filed Chapter Requires

Chapter 8 treats harm as a record carrying a named owner.

The trigger is the appending of a permitted deviation record (710), the mechanism for which is governed by a separate chapter of the filing and is not described here. On that appending, and before any reparation arc is created, the semantic agent (100) performs an affected-party resolution (802), testing the affected-party class carried in the conduct descriptor of the admitted action against the counterparty identity records (114) held within the scope partition in which the deviating action is recorded. A record resolves as a member only where two conditions hold together: the class extracted maps to the interpersonal scope, and the identity primitive the record carries is recorded, in an entry of the append-only lineage field (104), as a party to or a recipient of the deviating action. A counterparty resolution record is then appended, naming the class extracted, each record tested, and the outcome of each test.

Exactly two outcomes are available. An identified-counterparty outcome names the identity primitive of each resolving record. Otherwise the outcome is a no-identified-counterparty outcome, and the affected party is designated the structurally silent party.

The filing is careful about what that second outcome means: it is a determination that the agent's own records disclose no counterparty of the class, not an assertion that none exists. The agent surveys no party absent from its records.

On the identified-counterparty outcome, the amount exceeding the deviation deductible becomes an other-directed reparation arc (804), carrying an arc designation field and a counterparty binding field, both written at creation and neither removed nor modified afterward. A later change in the counterparty identity records (114) held does not redesignate an arc already created. Where several counterparties resolve, the amount is apportioned among them, with the deductible drawn once against the aggregate rather than per counterparty. Rounding residue is assigned by earliest recorded time, and where times do not strictly order the parties, by lexicographic order of the identity primitive. Apportionment discharges no portion and leaves the total unchanged.

Discharge has one path. The agent emits a reparation acknowledgment artifact as the first governed observation (608), enumerating the deviation record that occasioned the arc, the policy constraint overridden, the amount apportioned to that counterparty, and the action classes to which it is directed. The pair completes only if the bound counterparty produces a reparation acceptance determination (806) as the second governed observation (610), within a discharge window declared in the signed policy object (112) and expressed as a count of successor epochs of the agent's own hash chain. Pair recognition itself is governed by a separate chapter. What Chapter 8 states plainly is who may not produce that determination: not the semantic agent (100), not its principal, and not any adjudicator, arbiter, or other third party appointed to determine whether the reparation was sufficient.

Suppose the agent instead performs a restorative mutation on its own against such an arc. The record is appended and retained as a non-discharging restorative mutation record. The arc remains pending, the accumulated amount stays in the retention register, and the self-esteem aggregate (108) receives no positive entropy-scaled increment. Where the accumulated amount exceeds the aggregate retention (800), a

retention foreclosure record (808) is appended and the capacity to deviate is foreclosed, which a successor policy declaring a larger aggregate retention (800) does not lift while that record stands.

Two Different Definitions of Settled

Both approaches concern the same abstract object: an obligation created by an automated decision that went against somebody. The convergence there is real. The divergence sits at the level of what closes the obligation.

Under a commercial guarantee model, closure is a payment made under an agreement between the merchant and the vendor. That is appropriate to its purpose: the merchant is the contracting party, and the exposure being priced is the merchant's.

The filed architecture makes a different structural commitment, and its cost is deliberate. Since the arc binds to the identity primitive of the party recorded as harmed, the disclosed architecture requires that party's own acceptance and admits no substitute act: not a correction by the deviating agent, not a period of standing without further incident, not policy succession, and not elapsed time. Non-response resolves nothing in either direction. The arc remains pending, and equally, no counter of the counterparty is incremented, no record adverse to it is appended to its counterparty identity record (114), and no standing quantity of the counterparty is modified. So the mechanism cannot be turned into a scoring instrument aimed at parties who decline to accept.

That non-weaponizability creates an exposure the filing names rather than hides. Where a bound counterparty has withheld the action class within which acceptance would be produced, by an attestation mechanism a separate chapter governs, it can carry the agent toward foreclosure without producing any determination and without emitting any refusal. The disclosed remedy is narrow. The agent appends an abstention-suspended accumulation record, and the arc's contribution to the amount tested against

the aggregate retention (800) is held at the amount then standing while the entry stands unreleased. The arc is not discharged, redesignated, removed, or reduced. Suspension bears on accumulation alone, and the record is constructed without a count, a score, or a magnitude derived from the withholding.

Where the Two Sit in a Real Stack

These are not substitutes, and a team running fraud decisioning would not be choosing between them.

A guarantee-backed fraud vendor answers a question about a commercial relationship between a merchant and its risk provider. It works because both parties are identified, contractually bound, and able to move money. Where that structure applies, an architecture built for agents under delegated authority adds nothing to it.

The filed chapter answers a question arising elsewhere: where an autonomous agent acts for a principal, holds identity records for the parties it deals with, and can admit that a specific action harmed a specific one of them. Four properties matter for deployment.

- **Each obligation names its owner.** An arc binds to one identity primitive, and only a matched pair with that party retires it.
- **Failed restoration stays on the record.** A non-discharging restorative mutation record is retained rather than rejected, disclosing that restoration was attempted and discharged nothing.
- **Terminal consequence is conditioned.** Foreclosure follows where the accumulated amount exceeds the declared aggregate retention (800), which the filing treats as a policy-declared value. The chapter's worked trace uses a deviation deductible of 2.0 units and an aggregate retention (800) of 10.0 units by way of illustration only.

- **Harm with no reachable party still accumulates.** Where the class resolves to no identified counterparty, the arc is undischageable and accumulates at a declared multiple not less than unity, and where the accumulation consists wholly of such arcs, the loss of the capacity to deviate is permanent.

The two layers compose without contradiction. Commercial guarantees settle priced exposure between contracting entities. Counterparty-directed reparation settles the record of a specific harm with the party who bore it, and declines to let the harming party declare that settlement itself.

Disclosure Scope

This article describes subject matter in Chapter 8 of U.S. Provisional Application No. 64/117,812, a pending application. It is published for technical discussion and defensive publication purposes. Statements about the architecture are drawn from that chapter alone; other subject matter in the filing is governed by separate chapters and is not described here. Quantities such as the deviation deductible, the aggregate retention (800), the unaddressed multiple, the discharge window, the re-emission cadence, and the minimum apportioned amount are policy-declared in the signed policy object (112), and the values in the chapter's worked trace are illustrative only. Nothing here is an offer, a legal opinion, or a representation that any claim has been allowed.

References to Riskified are to public materials and are used for comparison only; no relationship, endorsement, or infringement is asserted.

Counterparty-Directed Reparation (/ [All 49 steps](#) → (/inventive-steps)

counterparty-reparation

Repair directed at the harmed party — bounded and resistant to weaponization.

Provisional application

PRIMARY TECHNICAL DISCLOSURE

- [Counterparty-Directed Reparation: Repair Owed to the Party Actually Harmed \(/articles/counterparty-reparation/reparation-directed-to-the-party-harmed\)](/articles/counterparty-reparation/reparation-directed-to-the-party-harmed)

SECONDARY TECHNICAL

- [The Costly Acknowledgment Artifact \(/articles/counterparty-reparation/costly-acknowledgment-artifact\)](/articles/counterparty-reparation/costly-acknowledgment-artifact)
- [Completion by Non-Worsening Re-Encounter \(/articles/counterparty-reparation/completion-by-non-worsening-re-encounter\)](/articles/counterparty-reparation/completion-by-non-worsening-re-encounter)
- [Partial Reparation Acceptance and the Residual Arc \(/articles/counterparty-reparation/partial-reparation-and-residual-arc\)](/articles/counterparty-reparation/partial-reparation-and-residual-arc)
- [Dual-Lock Escrow of the Discharge Itself \(/articles/counterparty-reparation/dual-lock-escrow-of-discharge\)](/articles/counterparty-reparation/dual-lock-escrow-of-discharge)
- [Per-Integrity-Scope Aggregate Retention and OR-Firing Foreclosure \(/articles/counterparty-reparation/aggregate-retention-or-firing-foreclosure\)](/articles/counterparty-reparation/aggregate-retention-or-firing-foreclosure)
- [Costless Rehearsal of Repair: Speculative Evaluation of Acknowledgment Artifacts and Restorative Mutations Inside the Non-Executing Cognitive Mode \(/articles/counterparty-reparation/costless-rehearsal-repair-non-executing-mode\)](/articles/counterparty-reparation/costless-rehearsal-repair-non-executing-mode)
- [Delegated Acceptance Authorization: Discharging a Reparation Arc Through a Counterparty's Delegate or Principal \(/articles/counterparty-reparation/delegated-acceptance-authorization\)](/articles/counterparty-reparation/delegated-acceptance-authorization)
- [Capping Pending Reparation Arcs: The Outstanding-Arc Bound, Untendered Marking, and the Refusal-Meter Exemption \(/articles/counterparty-reparation/outstanding-arc-bound-untendered-marking-refusal-meter\)](/articles/counterparty-reparation/outstanding-arc-bound-untendered-marking-refusal-meter)
- [Commitment-Conditioned Reparation Acceptance: A Counterparty-Evaluated Conformity Determination Within an Epoch-Count Observation Interval \(/articles/counterparty-reparation/reparation-conformity-determination-epoch-count-interval\)](/articles/counterparty-reparation/reparation-conformity-determination-epoch-count-interval)
- [Reparation Containment Suspension: Holding an Arc's Accumulation While a Scope Partition Stands Contained \(/articles/counterparty-reparation/reparation-containment-suspension\)](/articles/counterparty-reparation/reparation-containment-suspension)
- [Structured Inquiry to the Principal on Retention Foreclosure: Disclosing an Agent's Foreclosed Permission Condition Without Supplying an Override Path \(/articles/counterparty-reparation/structured-inquiry-principal-foreclosure\)](/articles/counterparty-reparation/structured-inquiry-principal-foreclosure)

APPLICATIONS • GENERAL

- [Automated Consumer Redress: Repair Directed at the Party Harmed \(/articles/counterparty-reparation/consumer-redress-automation\)](/articles/counterparty-reparation/consumer-redress-automation)

- [Insurance Claims Networks: Discharging Harm to the Counterparty That Holds It \(/articles/counterparty-reparation/insurance-claims-agent-networks\)](/articles/counterparty-reparation/insurance-claims-agent-networks).

APPLICATIONS · SPECIFIC

- [Card Network Chargebacks: Counterparty-Directed Remedy in Practice \(/articles/counterparty-reparation/payment-chargeback-systems\)](/articles/counterparty-reparation/payment-chargeback-systems)
- [Stripe Disputes: Refunds and Counterparty-Directed Repair \(/articles/counterparty-reparation/stripe-disputes\)](/articles/counterparty-reparation/stripe-disputes)
- [Merchant Protection and Repair Owed to the Harmed Party \(/articles/counterparty-reparation/shopify-protect\)](/articles/counterparty-reparation/shopify-protect)
- [Signifyd: Chargeback Guarantees and Repair to the Harmed Party \(/articles/counterparty-reparation/signifyd\)](/articles/counterparty-reparation/signifyd)
- [Riskified: Approval Guarantees and Who Bears the Harm \(/articles/counterparty-reparation/riskified\)](/articles/counterparty-reparation/riskified).

TERMINOLOGY

- [other-directed reparation arc \(/glossary#other-directed-reparation-arc\)](/glossary#other-directed-reparation-arc)
- [reparation arc \(/glossary#reparation-arc\)](/glossary#reparation-arc)
- [restorative mutation \(/glossary#restorative-mutation\)](/glossary#restorative-mutation)
- [unaddressed reparation arc \(/glossary#unaddressed-reparation-arc\)](/glossary#unaddressed-reparation-arc).

[Counterparty-Directed Reparation overview → \(/counterparty-reparation\)](/counterparty-reparation)