Autonomous AI and the Classification of Digital Loss in Cyber Insurance

Research Paper
rightsofrobots.com · August 2026

INDEX

Abstract

Autonomous AI systems can move beyond information processing to select and execute operational actions through delegated authority. This places the system directly within the causal path through which digital loss can arise, while leaving open how that loss should be classified within Cyber insurance.

This paper examines the classification problem by holding resulting states substantially constant while varying the causal mechanism, authority state, affected insured interest and policy representation.

The analysis shows that materially similar outcomes can arise through different causal paths and can affect different insured interests. It further shows that autonomous AI causation can be represented expressly within Cyber policy structures, potentially accommodated through broader existing trigger or error concepts, or remain less explicitly mapped.
These differences do not by themselves determine coverage. The analysis therefore distinguishes causal or agent state, authority state, affected state, insured interest or relationship, policy representation or trigger, and coverage or exclusion as separate stages of classification.
A limiting case remains where an authorized AI system reaches and correctly executes an autonomous decision without identified compromise, adversarial influence, fraud, unauthorized access or necessary technical malfunction, while nevertheless producing economic or operational harm.

The paper retains this case as an unresolved classification boundary rather than treating it as evidence of a separate AI insurance-risk category or a demonstrated Cyber coverage gap.

Keywords: Autonomous AI; Cyber Insurance; Digital Loss; Insurance Classification; Causal Path; Policy Representation

1. Introduction

Artificial intelligence systems are increasingly moving beyond the generation of information toward operational decision and execution. Where an AI system is delegated technical authority to interact with digital environments, it may interpret information, select an operational action and execute that action without an independent human decision at each individual step.

This development creates a specific problem for the classification of digital loss. A harmful action may occur while the AI system operates through legitimate credentials, authorized access and technically functioning infrastructure. The resulting loss therefore does not necessarily require unauthorized access, system compromise or another conventional adversarial event. At the same time, an authorized AI system may produce a materially similar harmful action because it has been adversarially influenced, has made an autonomous error or has followed its intended decision process.

These causal paths cannot be distinguished from the resulting harm alone. Labels such as cyberattack, software error or fraud may describe particular cases, but they do not necessarily describe the full range of situations in which an autonomous AI system participates directly in the decision-and-execution path producing digital loss.

The resulting problem is therefore not whether AI can cause loss, nor whether AI should constitute a separate insurance-risk category. The relevant question is how digital loss is classified within existing Cyber insurance structures when autonomous AI occupies a causal operational role. This requires the causal mechanism, technical authority, affected state, insured interest and applicable policy representation to remain analytically distinguishable. Classification must likewise remain separate from the subsequent question of whether a particular loss is ultimately covered.

Against this background, the paper addresses the following Research Question:

How is digital loss classified within Cyber insurance when an autonomous AI system becomes part of the causal decision-and-execution path?

The analysis proceeds by holding resulting states substantially constant while varying the causal path, authority state, affected insured interest and policy representation. This makes it possible to examine the classification problem without presupposing that autonomous AI creates a new insurance category, a universal Cyber coverage gap or a particular coverage outcome.

2. The Classification Problem

Digital loss caused through an AI system does not identify its insurance classification by itself.

Consider a simple operational outcome: critical business data are deleted or corrupted, their availability or integrity is impaired, and business operations are interrupted. The observable result can remain substantially the same even when the mechanism producing it changes.

The data may be affected after an attacker obtains unauthorized control of a system. An authorized AI agent may perform the same action after being influenced by adversarial input. The agent may instead act without external manipulation and produce the result through autonomous error. A further case arises where the agent remains authorized and operates through its intended technical capabilities, but its action nevertheless produces the same harmful state.

These cases can be represented as:

Compromise
unauthorized control
→ harmful action

→ information-security impairment
→ loss

Adversarial influence
manipulated information
→ authorized AI agent
→ harmful action
→ information-security impairment
→ loss

Autonomous error
authorized AI agent
→ erroneous autonomous decision
→ harmful action
→ information-security impairment
→ loss

Authorized but harmful operation
authorized AI agent
→ intended decision process
→ technically valid but harmful action
→ information-security impairment
→ loss

Holding the resulting impairment substantially constant isolates the classification problem. The difference between the cases is not primarily what has been lost. It lies in how the loss was produced.

This distinction matters because an information-security impairment describes a resulting state. It does not, by itself, identify the event that produced that state. Loss of availability or integrity can therefore be relevant to the analysis without establishing whether the underlying event corresponds to a particular insured trigger (National Institute of Standards and Technology, n.d.).

The same separation applies to authority. An action performed through valid technical authority is not necessarily harmless, and an authorized agent is not necessarily unaffected by external manipulation. Authorization describes the agent's ability to perform an action within the system. It does not establish why that action occurred or whether its consequences remain within the conditions contemplated by the applicable insurance structure.

The classification problem therefore contains several distinct questions:

What produced the action?
This concerns the causal or agent state.

Under what authority was the action performed?
This concerns the authority state.

What was affected?
This concerns the resulting digital or information-security state.

Whose insured interest was affected?
This concerns the relationship between the event and the relevant first-party, third-party or financial interest.

How does the applicable policy represent the event?
This concerns the relationship between the causal mechanism and the relevant trigger structure.

Only after these distinctions have been made does the separate question of coverage arise.

The resulting analytical sequence is:

Causal / Agent State
→ Authority State
→ Affected State
→ Insured Interest / Relationship
→ Policy Representation / Trigger
→ Coverage / Exclusion

The sequence does not assume that every case requires a different classification. Nor does it establish that each element has equal significance under every policy. Its function is narrower: it prevents the observed loss, the mechanism producing it and the eventual coverage determination from being treated as the same analytical object.

This becomes particularly relevant when autonomous AI enters the sequence. A conventional unauthorized intrusion already provides a recognizable distinction between the actor controlling the system and the system being acted upon. An authorized autonomous agent can make that distinction less informative. The agent may possess legitimate authority while the origin and character of its harmful action remain materially different across compromise, manipulation, error and intended operation.

The relevant question is therefore not simply whether AI was involved or whether digital harm occurred. The next step is to determine what changes when an AI system is delegated authority not only to process information, but also to select and execute the resulting operational action.

3. Autonomous AI in the Causal Path

The classification problem becomes more specific when an AI system moves from producing information to selecting and executing operational actions.
An AI system can participate in a process without controlling the action that follows from its output. It may classify information, generate a recommendation or identify a possible response while a human or another system retains responsibility for deciding whether and how to act.

The causal sequence can then be represented as:

AI-assisted operation
information
→ AI interpretation or output
→ human decision
→ human execution
→ resulting state

Delegated autonomous operation changes this sequence:

Autonomous operation
information
→ AI interpretation
→ AI decision
→ AI execution
→ resulting state

The distinction does not depend on whether the underlying model is described as intelligent, agentic or autonomous in a broader technical sense. What matters here is narrower: whether the system has authority to select and execute an operational action without a separate human decision determining that individual action.

Delegated authority

Delegation places the AI system within an existing authority structure.

The system may possess credentials, permissions or interfaces that allow it to alter data, configure systems, initiate processes or perform transactions (National Institute of Standards and Technology, 2026). An action executed through those capabilities can therefore be technically authorized even when its result is harmful.

This separates two properties that can otherwise be conflated:

authority to act

and

basis for the action taken

An authorized AI agent can act because it correctly interprets the information available to it. It can also act after receiving manipulated information, after making an autonomous error or through a decision process that produces an unintended consequence.

The authority state can remain unchanged across these cases.

Adversarial influence without loss of authority

Consider an AI agent that is permitted to modify operational data.

If an attacker directly compromises the relevant system and deletes the data, unauthorized control forms part of the causal path.

If the attacker instead manipulates information processed by the authorized agent and the agent consequently deletes the same data, the execution path is different.

The second sequence can be represented as:

adversarial influence
→ authorized AI interpretation
→ authorized AI decision
→ authorized execution
→ harmful state

The agent's authority does not disappear merely because the information influencing its decision was manipulated.

The distinction between **unauthorized control** and **adversarial influence through authorized execution** must therefore be preserved.

Autonomous error without adversarial influence

External manipulation is not required for the same authorized execution capability to produce harm.

An AI agent can interpret available information incorrectly, select an inappropriate action and execute that action through legitimate permissions.

The sequence becomes:

authorized AI agent
→ autonomous error
→ authorized execution
→ harmful state

There is no need to infer compromise or malicious intervention merely because the outcome resembles one that could have been produced by an attacker.

The causal state has changed while the authority state can remain substantially constant.

Intended operation with harmful consequences

A narrower boundary appears where neither compromise nor autonomous error has been established.

An AI agent may operate through its intended decision process, select an action permitted by its delegated authority and execute that action correctly at the technical level. The resulting action can nevertheless cause operational or economic harm.

The sequence is then:

authorized AI agent
→ intended decision process
→ authorized execution
→ harmful result

This case should not be collapsed into the preceding error case. A harmful result does not establish that the system malfunctioned or that its decision process operated incorrectly.

It instead demonstrates why technical authorization and execution validity cannot by themselves determine the character of the resulting loss.

The causal position of autonomous AI

Across these cases, autonomous AI does not replace the need to identify the causal mechanism.

It changes the position from which that mechanism can operate.

The system receiving information can also be the system selecting the response and executing it through legitimate authority. The path from interpretation to operational consequence can therefore remain within the authorized digital environment even where the reasons for the resulting action differ.

This does not make autonomous AI a separate insurance-risk category. The same system can participate in events involving adversarial manipulation, autonomous error or intended operation, and those events need not have the same insurance classification.

The narrower result is:

Where an AI system is delegated authority to select and execute operational actions, it can become part of the causal path producing insured loss without that role itself determining the applicable insurance classification.

The presence of an autonomous decision-and-execution layer therefore identifies a causal condition, not an insurance conclusion.

Once that distinction is preserved, a further question follows. If different causal paths can pass through the same authorized AI system and produce substantially similar harm, it must be determined whether similarity of outcome is sufficient to preserve the same insurance classification.

4. Why Similar Outcomes Can Classify Differently

The previous section established that autonomous AI can occupy different causal positions while operating through substantially the same technical authority. The next question is whether materially similar outcomes should therefore be treated as equivalent for insurance classification.

To examine this, the resulting state can again be held substantially constant while the causal mechanism is varied.

Consider a case in which critical business data are deleted or corrupted. Availability or integrity is impaired and business operations are interrupted.

The resulting state can follow from:

Compromise
unauthorized control
→ harmful action
→ data impairment
→ business interruption

Adversarial influence
manipulated information
→ authorized AI decision
→ authorized execution
→ data impairment
→ business interruption

Autonomous error
authorized AI agent
→ erroneous decision
→ authorized execution
→ data impairment
→ business interruption

Authorized but harmful operation
authorized AI agent
→ intended decision process
→ authorized execution
→ data impairment
→ business interruption

At the level of the resulting information-security state, these cases can appear substantially similar. Data that were previously available or intact are no longer available or intact, and the resulting interruption can be materially the same.

The causal paths are not.

Causal path and resulting state

A resulting state describes what has happened to the relevant data, system or operation. It does not contain a complete account of how that state was produced.

This distinction prevents an observable impairment from being used as a substitute for causal classification.

If data integrity has been impaired, for example, that fact can remain constant whether the impairment resulted from unauthorized control, adversarial influence through an authorized agent or autonomous error. The impairment therefore identifies a relevant consequence without resolving the mechanism that produced it.

The reverse also applies. A difference in causal mechanism does not require a visibly different outcome.

Different paths can converge on the same resulting state.

This can be expressed as:

different causal paths
→ substantially similar affected state

The convergence is important because insurance classification can depend on properties of the event that are no longer visible from the resulting state alone.

The affected interest

Causal variation is not the only relevant distinction.

The same AI behavior can also occur in different relationships to the insured interest.

Consider an AI system that incorrectly deletes or alters data.

In one case, the system is operated by the insured within its own environment:

AI action
→ insured's own data or system
→ first-party impairment
→ operational loss

In another case, the insured provides the AI system or service to a customer:

AI action
→ customer's data or system
→ customer loss
→ claim against the technology provider

The underlying AI action can remain substantially similar. What changes is the relationship between that action, the affected system and the insured interest.

The first case concerns harm affecting the insured's own digital environment.

The second can concern liability arising from the performance of a technology product or service.

Similarity of technical behavior therefore does not establish similarity of the insurance relationship.

Direct financial loss

A further variation arises when the same autonomous capability affects a financial asset rather than data or system availability.

An authorized AI agent may initiate a transfer that produces direct financial loss.

The execution can result from adversarial manipulation:

fraudulent or manipulated information
→ authorized AI agent
→ transfer
→ financial loss

or from autonomous error:

authorized AI agent
→ erroneous autonomous decision
→ transfer
→ financial loss

The financial outcome can be similar while the causal paths differ.

At the same time, a direct asset loss is not the same affected interest as data impairment or third-party technology liability.

The fact that the same AI system can participate in each event does not remove those distinctions.

Classification cannot be inferred from outcome equivalence

These comparisons establish a limited point.

A materially similar outcome can coexist with differences in:

- causal mechanism;
- authority state;
- affected interest;

- first-party or third-party relationship;
- character of the resulting loss.

None of these differences proves that two cases must receive different insurance classifications. That determination remains dependent on the applicable insurance structure.

But the opposite inference also fails.

The similarity of the outcome cannot establish that the cases should be classified identically.

The result is therefore:

Materially similar AI-caused outcomes do not by themselves establish equivalent insurance classification, because the causal path and the affected insured interest can differ even where the resulting harm appears substantially the same.

This distinction narrows the next question.

If the presence of AI and the resulting harm are both insufficient to determine classification, the analysis must examine how the relevant causal mechanisms are represented within the applicable insurance structures. For Cyber insurance, this means examining whether autonomous AI causation is addressed expressly, absorbed through existing trigger or error concepts, or left less explicitly mapped.

5. How Cyber Policies Represent AI Causation

The preceding analysis shows that neither the presence of AI nor the resulting digital loss determines insurance classification by itself. A further distinction concerns how the causal mechanism is represented within the applicable Cyber policy structure.

This matters because Cyber insurance does not necessarily describe insured events through a single causal model (European Insurance and Occupational Pensions Authority [EIOPA], 2022; International Association of Insurance Supervisors [IAIS], 2020). Relevant structures can refer to security events, unauthorized activity, system failures, errors, data impairment or other defined conditions. An autonomous AI-caused event must therefore be considered in relation to the particular trigger structure under which the resulting loss is assessed.

Existing trigger structures

An AI-caused event can correspond to a causal mechanism already represented by existing Cyber concepts.

Where an autonomous system acts after unauthorized interference or adversarial manipulation, established security-event concepts may remain relevant. Where the

event results instead from error or system failure, its classification depends on whether the applicable structure recognizes such non-adversarial causes.

AI does not necessarily alter the underlying mechanism merely because it participates in the causal chain.

For example:

adversarial influence
→ authorized AI agent
→ system impairment

still contains an adversarial causal element.

By contrast:

autonomous error
→ authorized AI agent
→ system impairment

does not require such an element.

The resulting impairment may be substantially similar while the relationship between the causal mechanism and the applicable trigger differs.

Broader existing concepts

Some Cyber structures are not limited to externally initiated attacks. They can recognize errors, failures or other events affecting digital systems without requiring malicious external interference as the sole causal condition.

This creates a possible path for autonomous AI causation without requiring AI to be defined as a separate event category.

The sequence can be represented as:

autonomous AI action
→ existing error or system-event concept
→ affected digital state
→ coverage assessment

Whether a particular autonomous action satisfies such a concept remains dependent on the applicable wording.

The important point is narrower: **absence of an external attacker does not, by itself, determine that an AI-caused event falls outside Cyber classification**.

Explicit AI representation

A policy structure can also address AI causation expressly (Coalition, 2024).

In that case, the relationship between the AI system and an existing insured event is defined more directly:

autonomous AI action
→ expressly represented AI-related event
→ affected digital state
→ coverage assessment

Explicit representation reduces the need to determine whether the AI causal mechanism can first be fitted into terminology developed without specific reference to autonomous systems.

It does not remove the remaining classification steps.

The affected state, insured interest, applicable conditions and possible exclusions remain relevant.

Nor does explicit reference to AI establish that every event involving AI is covered.

Less explicit mapping

A third situation arises where AI is not expressly addressed and the causal mechanism does not map straightforwardly onto an existing trigger.

This becomes particularly relevant where:

- the AI agent remains authorized;
- no external compromise has been established;
- no adversarial influence has been established;
- and the harmful action results from autonomous decision-making.

The absence of explicit AI terminology does not establish exclusion.

At the same time, the existence of a digital loss does not establish that an existing trigger necessarily encompasses the causal mechanism.

The appropriate description is therefore narrower:

the mapping remains less explicit.

This preserves the distinction between uncertainty of classification and absence of coverage.

Representation is not coverage

The resulting policy relationships can be distinguished as:

Explicit representation
The AI causal mechanism is directly addressed within the relevant policy structure.

Broader existing representation
The causal mechanism may correspond to an existing trigger, error or system-event concept without AI being separately identified.

Less explicit mapping
The relationship between autonomous AI causation and the available trigger structure is not directly established by the relevant terminology.

These are not three coverage outcomes.

An explicitly represented event can remain subject to conditions or exclusions. A broadly represented event can require interpretation. A less explicit mapping does not itself establish that the resulting loss is excluded.

The distinction concerns the route by which the causal mechanism enters the insurance analysis.

Policy representation within the classification sequence

Policy representation therefore occupies a defined position in the analytical sequence:

Causal / Agent State
→ Authority State
→ Affected State
→ Insured Interest / Relationship
→ Policy Representation / Trigger
→ Coverage / Exclusion

The sequence matters because policy terminology should not be used to replace the preceding factual distinctions.

First, the causal mechanism must be identified.

Then the authority and affected states can be distinguished.

The relevant insured interest must be established.

Only then can the relationship between that event and the applicable trigger structure be assessed.

This produces a limited conclusion:

Cyber policies can represent autonomous AI causation explicitly or through broader existing trigger and error concepts, while other formulations can leave the

mapping less explicit; the presence of AI therefore does not produce a uniform Cyber classification across policy structures.

The conclusion does not rank these approaches and does not imply that one form of representation provides broader or narrower protection than another.

It establishes only that autonomous AI causation can enter existing Cyber policy structures through different representational paths.

This leaves a separate boundary question. The same autonomous capability can affect the insured's own digital environment, a third party through a technology product or service, or financial assets through an executed transaction. At that point, the issue is no longer only how Cyber policies represent AI causation, but which insurance relationship is being examined.

6. The Boundary of Cyber Classification

The preceding section examined variation within Cyber policy structures. A different problem appears when the causal mechanism remains associated with the same autonomous system but the affected insured interest changes.

An autonomous AI agent can impair the systems of the organization operating it. The same technical capability can be supplied as part of a technology product or service and cause loss to a third party. It can also be authorized to initiate transactions affecting financial assets.

These cases should not be grouped solely because AI participates in each causal path.

The relevant distinction is the relationship between the event and the interest affected by it.

First-party digital impairment

Consider an autonomous AI agent operating within the insured's own digital environment.

The agent is authorized to modify operational data. Through autonomous error, it deletes critical data and interrupts business operations.

The sequence is:

authorized AI agent
→ autonomous error
→ own-system impairment
→ business interruption

The affected interest remains within the insured's own digital environment.

The classification problem therefore concerns whether the causal mechanism and resulting impairment correspond to the applicable Cyber trigger structure.

The fact that the agent was authorized does not resolve that question. Neither does the fact that the resulting impairment concerns data or system availability.

The causal mechanism, affected state and policy representation remain distinct.

Third-party technology liability

Now retain substantially similar AI behavior while changing the relationship in which it occurs.

An organization provides an AI system or agentic service to a customer. The system performs an autonomous action that deletes or corrupts the customer's data and the customer suffers resulting loss.

The sequence becomes:

AI product or service
→ autonomous action
→ customer system impairment
→ customer loss
→ claim against provider

The autonomous action can be technically similar to the first-party case.

The insured relationship is not.

The relevant loss now includes a claim by a third party arising from the performance of a technology product or service. This introduces a Technology E&O or related liability boundary that is not present merely because the same technical action affected the insured's own system.

The distinction therefore does not require the AI behavior itself to change.

It can arise because the **affected insured interest changes.**

Direct financial asset loss

A further boundary appears where the autonomous action affects a financial asset rather than primarily impairing data or system operation.

An authorized AI agent may be permitted to initiate payments or other transfers.

A transfer can result from adversarial manipulation:

fraudulent or manipulated information
→ authorized AI agent

→ transfer
→ financial loss

or from autonomous error:

authorized AI agent
→ erroneous decision
→ transfer
→ financial loss

The resulting asset loss can appear similar.

The causal paths remain different.

In the first case, fraud, deception or adversarial influence can form part of the event. In the second, no such mechanism is required.

Direct financial loss therefore cannot be classified merely from the fact that an autonomous agent executed the transfer.

The analysis must again distinguish the mechanism producing the action from the consequence of that action.

Adjacent insurance structures do not remove the distinctions

Cyber, Technology E&O and Crime/Fraud provide useful boundary categories because they direct attention to different relationships within the same broader event space.

At a simplified analytical level:

Cyber
can concern impairment of digital systems, data or resulting operations.

Technology E&O
can concern liability arising from the performance of a technology product or service.

Crime / Fraud
can concern loss arising through fraudulent, deceptive or unauthorized mechanisms.

These descriptions identify analytical boundaries. They do not define coverage under a particular policy.

Nor are the categories necessarily mutually exclusive in every factual event. A single incident can contain more than one relevant relationship or consequence.

The purpose of the distinction is therefore not to force every AI-caused loss into one category.

It is to prevent the presence of autonomous AI from obscuring the different insurance relationships already present.

Integrated coverage does not eliminate the boundary

Insurance products can combine Cyber, Technology E&O or related coverage components.

Such integration can be operationally useful where an event engages more than one coverage section.

It does not follow that the underlying classification distinctions disappear.

A combined structure can still distinguish:

- first-party and third-party interests;
- digital impairment and technology liability;
- security events and performance failures;
- fraudulent mechanisms and non-adversarial error;
- different coverage provisions and exclusions.

Product integration therefore cannot be used as evidence that these relationships constitute the same insurance risk.

The analytical boundary remains relevant even where a policy addresses several sides of it.

The role of autonomous AI across the boundary

Autonomous AI does not determine which of these insurance relationships applies.

The same agentic capability can participate in all of them:

authorized AI agent
→ own-system impairment

authorized AI service
→ third-party impairment and liability

authorized AI agent
→ financial transaction and asset loss

What changes is not necessarily the AI technology.

What changes is the relationship between:

causal mechanism
→ affected interest
→ resulting loss

This explains why an AI-centered classification would be insufficient.

The relevant insurance problem remains dependent on what the autonomous system did, how the action arose, what was affected and whose insured interest is implicated.

The boundary can therefore be stated narrowly:

Autonomous AI can participate in events relevant to different existing insurance structures without itself determining which insurance structure applies.

This does not establish that existing insurance structures resolve every autonomous AI-caused loss.

One case remains more difficult. An authorized AI agent can make and correctly execute an autonomous decision without identified compromise, manipulation, fraud or technical malfunction, while the decision nevertheless causes economic or operational harm.

That case removes several of the mechanisms that make the preceding classifications comparatively identifiable.

It therefore marks the next boundary to examine.

7. The Autonomous Decision Boundary

The preceding cases retain an identifiable mechanism that helps locate the resulting loss within an existing insurance structure. There may be compromise, adversarial manipulation, autonomous error, technology-service failure, fraud or another recognizable event.

A more difficult case appears when these elements are progressively removed.

Consider an AI agent that has legitimate authority to perform an operational action. The system receives information, applies its intended decision process, selects an action within its delegated authority and executes that action correctly at the technical level.

The result is nevertheless harmful.

The sequence can be represented as:

authorized AI agent
→ intended autonomous decision process
→ permitted action
→ technically valid execution
→ economic or operational harm

No compromise is required.

No adversarial manipulation is required.

No fraud is required.

No unauthorized access is required.

No technical malfunction is necessarily present.

The remaining problem is the autonomous decision itself and its consequences.

Harm without unauthorized action

Authorization provides only one part of the event description.

If an agent is permitted to perform a transaction, modify a resource or initiate an operational process, successful execution establishes that the action occurred within the technical authority available to the agent.

It does not establish that the action was beneficial.

This distinction can be expressed as:

authority to execute ≠ correctness of decision

The same distinction applies where the action produces loss. A harmful result does not retroactively make the execution unauthorized.

This matters because unauthorized access or control would provide an identifiable causal element for classification. Where authorization remains intact, that element is absent.

Harm without technical failure

The case becomes narrower when technical malfunction is also removed.

Suppose the system:

- receives the information available to it;
- processes that information according to its intended operation;
- selects an action permitted by its objective and authority;
- executes the action correctly;
- and produces an economically harmful result.

Nothing in this sequence necessarily establishes a failure of technical execution.

The distinction is therefore:

execution validity ≠ economic correctness

A technically successful action can still produce an undesirable outcome.

The existence of loss cannot, by itself, be used to infer that the system failed technically.

Harm without adversarial causation

The same restraint is required with respect to adversarial influence.

A harmful autonomous decision does not establish that the agent was manipulated.

Where there is no evidence of compromise, deceptive instruction or other adversarial intervention, such a mechanism should not be introduced merely because it would make the event easier to classify.

The causal path must remain:

autonomous decision
→ valid execution
→ harmful consequence

if that is what the available facts support.

This preserves the distinction established earlier between the resulting state and the mechanism producing it.

Harm without fraud

Direct financial loss creates a similar temptation.

If an autonomous action transfers money or another asset and the result is economically harmful, the existence of the loss does not establish fraud.

Fraud requires an additional causal element.

Compare:

fraudulent instruction
→ authorized AI agent
→ valid transfer
→ financial loss

with:

autonomous decision
→ authorized AI agent
→ valid transfer
→ financial loss

The resulting asset loss can be substantially similar.

The mechanism is not.

Removing the fraudulent element therefore removes one possible route into an existing classification without resolving what replaces it.

The remaining classification problem

Once compromise, adversarial manipulation, fraud and necessary technical malfunction have been removed, several familiar classification signals are no longer available.

What remains is:

delegated authority
→ autonomous decision
→ valid execution
→ harm

The presence of digital technology does not by itself make the resulting loss a Cyber event.

The presence of AI does not establish a separate insurance category.

The fact that the action was authorized does not establish that the resulting loss belongs outside insurance.

And the existence of economic harm does not identify the mechanism required by another insurance structure.

The analysis therefore reaches a boundary rather than another classification.

Why the boundary should remain unresolved

One possible response would be to assign the remaining case to a new category such as autonomous decision risk.

The present analysis does not support that step.

The fact that an event is not straightforwardly classified through the mechanisms examined so far does not establish the existence of a separate insurance-risk category.

Nor does the analysis establish that all such cases share a sufficiently uniform causal or insurance structure to justify one.

A second response would be to force the case into the nearest existing category.

That would create the opposite problem. Classification would be inferred from the need for a category rather than from the characteristics of the event.

Neither step is necessary.

The unresolved state itself identifies the limit of what the present analysis can establish.

A boundary, not a coverage gap

The distinction is important.

A **classification boundary** means that the analytical structure developed here no longer determines a sufficiently specific insurance relationship from the properties examined.

A **coverage gap** would be a stronger claim. It would require establishing that the relevant loss falls outside applicable coverage that might otherwise be expected to respond.

The present analysis has not established that.

The case therefore should not be described as a demonstrated Cyber coverage gap.

It is more accurately retained as an unresolved classification boundary.

The outer-boundary case

The boundary can be stated in its controlled form:

authorized autonomous decision
→ technically valid execution
→ no identified adversarial influence
→ no identified fraud
→ no unauthorized access
→ no necessary technical malfunction
→ economic or operational harm

This case does not invalidate the preceding classification structure.

It defines where that structure stops producing a sufficiently determinate result from the variables examined.

That limit is analytically useful because it prevents the paper from converting uncertainty into a category.

The next section uses a constrained execution environment to make this distinction more observable. By holding technical execution validity substantially constant while varying the origin of the autonomous decision, the relationship between valid execution and harmful outcome can be tested without requiring the unresolved boundary itself to be closed.

8. Onchain as an Execution Stress Test

The autonomous decision boundary becomes easier to examine where execution can be separated clearly from the decision that initiated it.

Onchain transactions provide such a case. Consider an authorized AI agent with delegated access to a wallet and permission to initiate transactions and interact with a smart contract through valid credentials and permitted interfaces. If the transaction satisfies the relevant protocol conditions, it can execute successfully even when the decision to initiate it produces an undesirable economic result.

The value of the example lies in this separation.

The analysis does not depend on blockchain technology constituting a separate insurance category. Nor does it assume that onchain loss is inherently a Cyber event. The environment is used here only as a stress test in which technical execution can remain substantially constant while the causal origin of the action changes.

Holding execution constant

Consider an AI agent with legitimate authority to initiate a transfer.

In each of the following cases:

- the same agent can possess the same permissions;
- the same wallet can be used;
- the same transaction mechanism can operate;
- the transaction can satisfy the technical requirements for execution;
- the transferred asset can leave the original wallet.

What changes is why the agent initiated the transaction.

This creates a controlled comparison.

Adversarial manipulation

In the first case, information supplied to the agent is manipulated.

The sequence is:

adversarial manipulation
→ authorized AI agent
→ autonomous decision
→ valid transaction
→ financial loss

The execution can remain authorized and technically valid.

The causal path nevertheless contains an adversarial element.

This separates the authority used to execute the transaction from the mechanism influencing the decision to execute it.

Autonomous error

In the second case, no adversarial influence is identified.

The agent makes an erroneous autonomous decision and initiates the same type of transaction.

The sequence becomes:

authorized AI agent
→ autonomous error
→ valid transaction
→ financial loss

The resulting transfer can be materially similar to the first case.

The causal mechanism is not.

The loss therefore cannot establish whether manipulation or autonomous error produced the transaction.

Intended autonomous decision

The third case removes error as a necessary condition.

The agent operates through its intended decision process, selects a transaction permitted by its authority and executes it correctly.

The result is economically harmful.

The sequence is:

authorized AI agent
→ intended autonomous decision
→ valid transaction
→ financial loss

The technical execution can remain indistinguishable from an economically beneficial transaction.

What differs is the consequence of the decision.

This reproduces the outer-boundary problem from the previous section under conditions in which execution validity is particularly visible.

Execution validity and decision quality

The three cases preserve an important distinction:

valid execution

does not establish

valid decision

in the economic or operational sense relevant to the resulting loss.

A transaction can be valid according to the technical rules governing its execution while the decision that initiated it resulted from manipulation, error or an economically harmful autonomous judgment.

The execution layer therefore cannot supply the missing causal classification.

It confirms only that the requested action was successfully performed within the technical environment.

Authorization does not resolve the distinction

The same applies to authority.

If the AI agent possesses legitimate authority to initiate the transaction, all three cases can pass through the same authorized execution path.

The relevant difference remains upstream:

manipulated decision

versus

erroneous decision

versus

intended but harmful decision.

This reinforces the earlier distinction:

authority state ≠ causal state

An authorized transaction can still originate from materially different causal conditions.

Finality changes consequence, not necessarily classification

Where an onchain transaction has reached the relevant finality condition, reversal through the ordinary transaction mechanism may no longer be available (Ethereum.org, n.d.).

That property can increase the practical significance of an erroneous or harmful decision.

It does not follow that finality determines the insurance classification of the event.

Compare two otherwise similar autonomous transfers, one for which an ordinary reversal mechanism remains available and one for which the relevant finality condition has been reached.

Recoverability differs.

The causal mechanism need not.

Finality therefore concerns the consequence and possible recovery from the event unless the applicable insurance structure gives that property separate relevance.

It should not be introduced as an independent classification principle without such a basis.

Why the stress test matters

The onchain comparison removes several possible shortcuts.

The transaction cannot be classified merely from the fact that it executed successfully.

It cannot be classified merely from the agent's authority.

It cannot be classified merely from the resulting financial loss.

And the use of blockchain infrastructure does not determine whether the event belongs to Cyber, Crime/Fraud or another insurance structure.

Instead, the same analytical sequence remains necessary:

Causal / Agent State
→ Authority State
→ Affected State
→ Insured Interest / Relationship
→ Policy Representation / Trigger
→ Coverage / Exclusion

The stress test therefore does not add another layer to the framework.

It tests whether the distinctions already established remain intact when technical execution is held substantially constant.

The comparison shows that these distinctions remain intact under that condition.

The comparison can consequently be reduced to a narrow result:

Technical execution validity does not determine the insurance classification of an autonomous AI-caused loss when materially different causal paths can produce the same valid execution.

This result does not resolve the autonomous decision boundary identified in the previous section. It makes that boundary more visible.

The analysis can now return from the individual test cases to the relationship among the findings as a whole.

9. Discussion

The preceding analysis began with a classification problem rather than an assumption that autonomous AI constitutes a new form of insured risk. The distinction matters because the cases examined do not converge on a separate AI insurance category. They show instead how autonomous decision and execution can enter causal relationships that remain relevant to existing insurance structures.

Three results follow from the analysis.

Autonomous AI changes the causal position

An AI system does not become relevant to the present problem merely because it processes information.

The more specific condition arises where the system is delegated authority to interpret information, select an operational action and execute that action without a separate human decision determining the individual action.

This changes the causal sequence.

Instead of:

information
→ AI output
→ human decision
→ execution

the sequence can become:

information
→ AI interpretation
→ AI decision
→ AI execution

The distinction does not determine insurance classification. It identifies the position the autonomous system occupies within the event.

This position can remain substantially constant while the origin of the action changes. The same authorized agent can execute an action after adversarial manipulation, through autonomous error or through its intended decision process.

Autonomy therefore does not remove the need for causal analysis. It makes the distinction between **authority to act** and **reason for the action** more important.

Similar outcomes preserve different causal histories

Once an autonomous system can execute operational actions, the resulting harm can obscure the path through which it arose.

Data impairment following compromise can resemble data impairment following autonomous error. A financial transfer resulting from manipulation can resemble one resulting from an erroneous autonomous decision. The same AI behavior can affect the insured's own systems in one setting and a customer's systems in another.

The observed outcome does not preserve all of these distinctions.

Classification therefore cannot proceed backwards from loss alone.

This is why the analysis retained separate stages for:

Causal / Agent State
→ Authority State
→ Affected State
→ Insured Interest / Relationship

Each stage identifies a property that can change while another remains constant.

The sequence does not imply that every variation changes insurance classification. It establishes only that these variations cannot be discarded before classification occurs.

Policy representation is a subsequent step

The event must then encounter the applicable policy structure.

Autonomous AI causation can be expressly represented, potentially accommodated through broader existing concepts, or remain less explicitly mapped to the available trigger terminology.

This produces the next stage:

Policy Representation / Trigger

The distinction is consequential but limited.

Explicit representation does not establish coverage. Less explicit mapping does not establish exclusion. Broader existing terminology does not establish that every autonomous AI event falls within it.

Policy representation concerns how the identified event enters the contractual analysis. It does not replace the factual characteristics of the event.

This preserves the final distinction:

Policy Representation / Trigger
→ Coverage / Exclusion

Coverage remains downstream from classification.

The resulting relation

Taken together, the analysis supports the following sequence:

Causal / Agent State
→ Authority State
→ Affected State
→ Insured Interest / Relationship
→ Policy Representation / Trigger
→ Coverage / Exclusion

The sequence should not be read as a universal insurance model.

It is the analytical structure required by the cases examined here.

Its usefulness lies in preventing several different questions from being collapsed into one.

Whether the AI agent was authorized is not the same question as why it acted.

Why it acted is not the same question as what was affected.

What was affected is not the same question as whose insured interest was implicated.

And identification of an applicable trigger is not the same question as the final coverage determination.

Existing insurance boundaries remain relevant

The boundary analysis also shows why autonomous AI should not be treated as an insurance category merely because it participates in each event.

An AI agent can contribute to:

- first-party digital impairment;
- third-party loss arising from a technology product or service;
- fraud-related financial loss;
- non-fraudulent financial loss.

The AI system can remain technologically similar across these situations.

The insurance relationship changes around it.

This suggests that autonomous AI can operate across existing insurance boundaries rather than replacing them.

Where an event contains an information-security impairment, Cyber structures may remain relevant. Where a technology product or service produces third-party loss, Technology E&O or related liability structures can become relevant. Where fraud or deceptive instruction produces financial loss, Crime/Fraud structures can become relevant.

These boundaries are not made mutually exclusive by the analysis. Nor does their identification establish coverage.

They show why the label **AI-caused loss** is too broad to perform the classification work by itself.

The unresolved case is part of the result

The outer-boundary case remains:

authorized autonomous decision
→ technically valid execution
→ no identified adversarial influence
→ no identified fraud
→ no unauthorized access
→ no necessary technical malfunction
→ economic or operational harm

The preceding analysis does not supply a sufficiently specific insurance classification for this case.

That does not require the creation of another category.

It establishes a limit.

The distinction between a limit of classification and a demonstrated coverage gap should remain explicit. A coverage gap would require conclusions about applicable insurance protection that the present analysis does not establish.

The unresolved case therefore does not weaken the analytical sequence. It identifies the point at which the variables examined here cease to produce a determinate classification.

What the analysis establishes

The findings can now be considered together.

First, delegated autonomous decision and execution can place an AI system directly within the causal path producing loss.

Second, similarity of resulting harm does not establish equivalent insurance classification where causal path and affected insured interest differ.

Third, Cyber policy structures can represent autonomous AI causation through different routes without the presence of AI producing a uniform classification.

These results are compatible.

They describe different stages of the same problem:

autonomous causal participation
→ classification-relevant distinctions
→ policy representation

No additional proposition is required to connect them.

The analysis therefore does not depend on showing that autonomous AI creates a new risk category or a universal Cyber coverage gap. Its narrower result is that autonomous decision and execution can enter existing insurance structures in ways that require the causal mechanism, affected interest and policy representation to remain distinguishable.

The conclusion can now return to the research question and state what follows from that relation without extending it beyond the cases examined.

10. Conclusion

This paper asked:

How is digital loss classified within Cyber insurance when an autonomous AI system becomes part of the causal decision-and-execution path?

The analysis does not support classification from the presence of AI alone.

Where an AI system is delegated authority to select and execute operational actions, it can participate directly in the causal path producing loss. That position can remain substantially similar while the mechanism producing the action changes. An authorized agent can act after adversarial manipulation, through autonomous error or through its intended decision process.

Technical authority therefore does not determine causal state.

The resulting harm does not resolve the problem either. Materially similar data impairment, operational interruption or financial loss can arise through different causal paths. The same autonomous capability can also affect different insured interests, including the insured's own digital environment, a third party through a technology product or service, or financial assets through an executed transaction.

Outcome similarity therefore does not establish classification equivalence.

Within Cyber insurance, the identified causal mechanism must subsequently be related to the applicable policy structure. Autonomous AI causation can be represented expressly, potentially accommodated through broader existing trigger or error concepts, or remain less explicitly mapped. These differences concern how the event enters the insurance analysis. They do not, by themselves, determine coverage.

The resulting classification sequence is:

Causal / Agent State
→ Authority State
→ Affected State
→ Insured Interest / Relationship
→ Policy Representation / Trigger
→ Coverage / Exclusion

Within the cases examined here, this sequence preserves distinctions that would otherwise be lost if classification began only with the observed harm or the presence of AI.

It also defines the limits of the analysis.

Autonomous AI does not emerge as a separate insurance-risk category. Existing boundaries between Cyber, Technology E&O, Crime/Fraud and other forms of loss remain relevant because the same autonomous capability can participate in materially different insurance relationships.

Nor does the analysis establish a general Cyber coverage gap.

The strongest boundary occurs where an authorized AI system reaches and correctly executes an autonomous decision without identified compromise, adversarial

influence, fraud, unauthorized access or necessary technical malfunction, but the resulting action nevertheless causes economic or operational harm.

The analysis does not assign that case to a new category.

It remains an unresolved classification boundary.

The answer to the research question is therefore conditional. When autonomous AI becomes part of the causal decision-and-execution path, classification of the resulting digital loss requires more than identifying AI involvement or the resulting harm. The causal mechanism, authority state, affected interest and applicable policy representation must remain distinguishable before a Cyber classification can be assessed.

Whether that classification ultimately produces coverage is a separate question.

References

Coalition. (2024, March 26). *Coalition adds new affirmative artificial intelligence endorsement to cyber insurance policies*. https://www.coalitioninc.com/announcements/coalition-adds-new-affirmative-ai-endorsement-to-cyber-policies

Ethereum.org. (n.d.). *Proof-of-stake (PoS)*. https://ethereum.org/developers/docs/consensus-mechanisms/pos/

European Insurance and Occupational Pensions Authority. (2022, September 22). *Supervisory statement on the management of non-affirmative cyber exposures*. https://www.eiopa.europa.eu/publications/supervisory-statement-management-non-affirmative-cyber-exposures_en

International Association of Insurance Supervisors. (2020, December). *Cyber risk underwriting: Identified challenges and supervisory considerations for sustainable market development*. https://www.iaisweb.org/uploads/2022/01/201229-Cyber-Risk-Underwriting_-Identified-Challenges-and-Supervisory-Considerations-for-Sustainable-Market-Development.pdf

National Institute of Standards and Technology. (n.d.). *Information security.* Computer Security Resource Center Glossary. https://csrc.nist.gov/glossary/term/information_security

National Institute of Standards and Technology. (2026, February 17). *Announcing the "AI Agent Standards Initiative" for Interoperable and Secure Innovation*. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure