

The Active Security Cycle

Keeping Security Risk Assessment in step with a changing world: what the brain's way of managing uncertainty can teach security management



Alberto Gomez Castro

Senior Professional in Security and Strategic Risk Management | Retired Peruvian Navy Officer | Former United Nations Chief Security Officer

ABSTRACT

Security Risk Assessment (SRA) and Security Risk Management (SRM) give organizations a structured understanding of their security environment, yet the product of an assessment is essentially a picture of a particular moment, usually expected to remain valid for six months or a year while environments are continuously changing. This article proposes the Active Security Cycle (ASC), a simple recurring process designed to keep that picture in contact with reality between formal assessments. The proposal draws, by analogy and without its mathematical apparatus, on the Free Energy Principle and Active Inference, a neuroscientific account of how the brain continuously predicts, perceives, acts and learns in an uncertain world. Treating the organization as an organism and its security function as its brain, the completed SRA becomes the organization's working security model, its perception and action tools become the boundary through which it deals with the operational environment, and any meaningful discrepancy between what was expected and what is now observed becomes the trigger for a cycle of seven steps: Expect, Sense, Infer, Compare, Act, Learn and Update. The article explains the rationale, examines the persistent difficulty of estimating likelihood when incident data are scarce, illustrates the cycle with three cases, and locates the ASC within established SRA/SRM processes, which it complements and does not replace.

KEYWORDS Active Security Cycle; Security Risk Assessment; Security Risk Management; uncertainty; prediction error; likelihood; Free Energy Principle; Active Inference; acceptance strategy; adaptive security management.

1 A Static Picture of a Moving World

For most of my professional life in security, risk assessment followed a familiar logic. We identified threats, examined vulnerabilities, estimated likelihood and impact, assigned levels of risk and selected mitigation measures. It is a useful and well-established approach, supported by international standards and by a considerable body of professional practice [1,2]. Over time, however, I became increasingly aware of a limitation that has less to do with the quality of any particular methodology than with the nature of what it produces.

Whatever the method used, the result of a Security Risk Assessment is usually a document. It describes the security environment, the organization's operations, the threats it faces, its vulnerabilities and the measures considered necessary, and it often does so carefully and very well. Yet it describes all this as it was understood at a particular moment, while it is expected to remain valid for a long period, frequently six months or a year, in environments full of uncertainty that keep changing from the day the document is approved. The monitoring and review mechanisms included in most SRA/SRM processes tend to concentrate on the implementation of the measures the assessment recommended. That is necessary, but it leaves aside a different task: detecting the changes in the environment that may be quietly eroding the picture on which those measures were based. ISO standards and professional manuals provide very useful principles and a sound framework within which to operate [1,2], although in my experience they do not fully capture the day-to-day dynamics of a corporate or organizational security system working in an uncertain operational environment.

Some time ago, and without being a mathematician or a neuroscientist, I began reading about the Free Energy Principle and its corollary, Active Inference, developed principally by the neuroscientist Karl Friston [3–6]. Somewhere in that process of learning through papers, books, podcasts and lectures, a question came to mind that has guided everything presented here: why don't we apply to security risk assessment and management the way the brain operates under uncertainty? The human brain is the most sophisticated natural system we know for keeping an organism safe and alive in an uncertain world. If we consider an organization as an organism, its security function as its brain, and its means of perceiving and acting as the boundary through which it deals with its operational environment, the way the brain works may give us some useful ideas.

This article integrates and develops three earlier pieces in which I explored that question: a conceptual proposal, a case study of an emerging threat in cybersecurity, and an exercise confronting the resulting model with a validated SRA procedure. The outcome is what I call the Active Security Cycle (ASC), a simple way of keeping a Security Risk Assessment in a dynamic relationship with the changes in its

environment. The ASC does not replace the SRA, nor does it attempt to correct the way a security professional develops each stage of the assessment. It begins where the SRA ends. In the same way that the brain works from a model of the world it expects, the completed assessment, whatever method produced it, becomes the starting model of the security environment the organization expects in order to carry out its activities. Running the ASC as a daily or frequently used tool then allows the organization to catch, in good time, any event or situation that challenges the validity of that model, instead of waiting longer periods to review it.

A NOTE ON THE STATUS OF THIS PROPOSAL

I use concepts from the Free Energy Principle and Active Inference as an analogy, interpreted and adapted for security management, and not as an exact translation of a scientific theory into another discipline. My expertise is security and not neuroscience, and I take full responsibility for the way those concepts have been used here. The Active Security Cycle is my own proposal and is not a standard model from the scientific literature. It should be regarded as a working hypothesis whose value will ultimately be decided by applying it to real security problems, on top of existing assessments, and observing whether it makes those assessments more dynamic and the resulting decisions better informed.

2 Learning from the Brain

The brain does not simply wait for information from the outside world, process it and then decide what to do. Under the Free Energy Principle and Active Inference, it can be understood as continuously generating expectations regarding what it is about to encounter, using sensory information as evidence of whether those expectations remain adequate [3–6]. When the two do not match, the organism can revise its understanding of what is happening, and it can also act upon its environment and change what happens next. Perception shapes action, action changes the organism's relationship with its environment, and the changed environment produces new sensory information that shapes perception again, in what may be described as a permanent dance of prediction, perception and action.

One concept associated with the Free Energy Principle helps to explain this relationship: the Markov blanket. Despite its technical name, its basic intuition can be described without mathematics. An organism has no direct access to everything happening in its environment, and the environment has no direct access to the organism's internal states. Their interaction is mediated by a boundary made of two kinds of states: *sensory states*, through which changes in the external world influence the organism, and *active states*, through which the organism influences the external

world [7–9].

Think of a person walking along a dark street. Knowing the route from previous evenings, he expects it to be quiet, with little traffic and nothing unusual on the way home; that expectation is his starting model of the street, in the same way that a completed assessment is an organization's starting model of its environment. He cannot know everything happening around the next corner, but he sees movement, hears sounds, notices changes in the light and observes how other people behave, and from these sensory inputs his brain infers what may be happening. He does more than perceive, however. He may slow down, cross the street, light up an area with his phone, ask someone a question or choose another route, and each of these actions changes both his relationship with the environment and the information available to him afterwards.

A security organization is in a remarkably similar situation. It never has direct access to the complete state of its operational environment. It relies upon its Security Risk Assessment as the starting model of what it expects to encounter, and upon intelligence, incident reporting, cameras, alarms, communications, liaison with authorities, personnel observations, community contacts and other sources to infer what is happening. It influences that environment through access control, guard forces, procedures, movement decisions, physical protection, training, contingency and evacuation planning, business-continuity arrangements, community engagement and many other interventions. I do not claim that these mechanisms literally constitute a biological Markov blanket, but I propose that they can be mapped functionally onto its sensory and active states. Information flowing toward the security system arrives through what I call the organization's perception tools, and interventions flowing from the security system toward its environment are its action tools. Between them sits the organization's internal security model: its assumptions, assessments and expectations about what is happening and what may happen next (Figure 1).

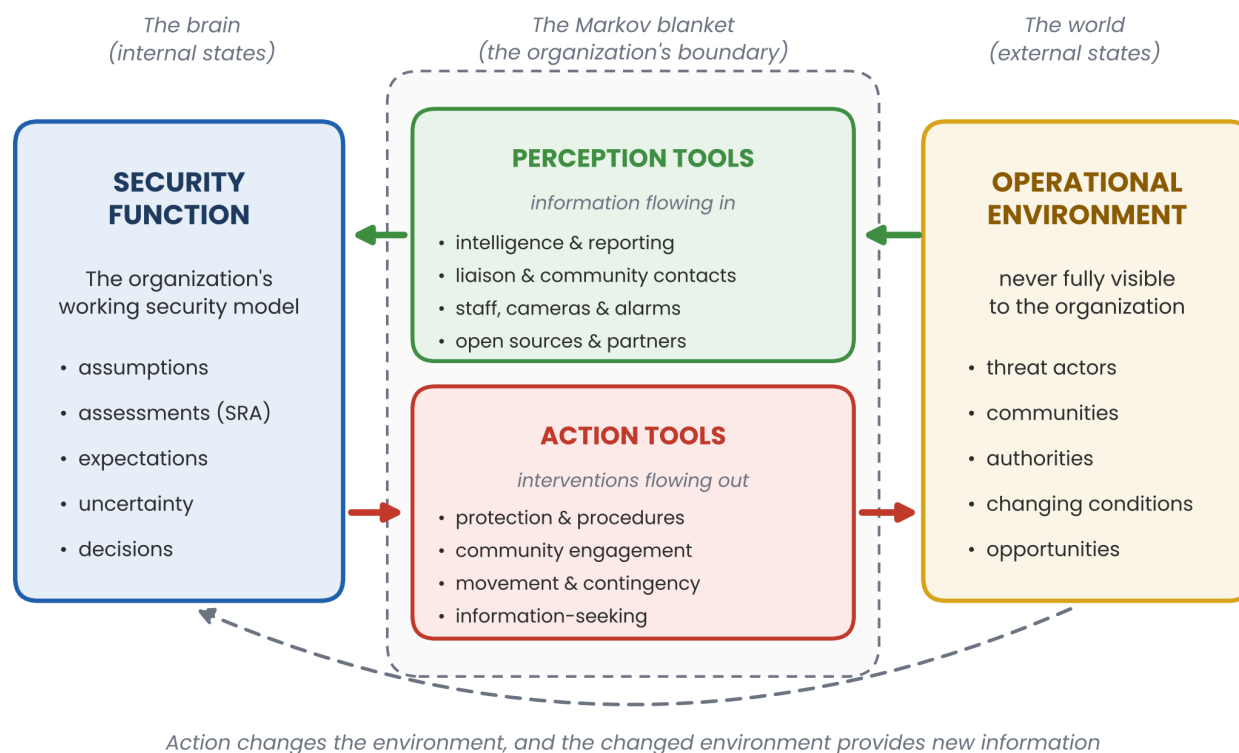


Figure 1. The organization as an organism. Perception tools and action tools play, by functional analogy, the role that sensory and active states play in the Markov-blanket framework, mediating the continuous interaction between the security function and an operational environment it can never observe completely. The diagram is conceptual and omits the mathematical formulation of the framework.

Seen this way, the security organization is much more than an information processor waiting for threat information to arrive from outside. It is an adaptive system in continuous interaction with its operational environment through perception and action, and that is the image on which the rest of this proposal is built.

3 Prediction Error: When Reality Stops Matching the Model

One further concept from Active Inference is central to the proposal, although it remained implicit in my earlier articles: prediction error. In simplified terms, the brain uses its internal model to anticipate the sensory information it is likely to encounter. When what is sensed corresponds sufficiently with what was expected, there is little reason to revise the model. When an important discrepancy appears, there is a prediction error [3–6]. The word *error* here does not mean that somebody made a mistake. It refers to the difference between what the system expected and what it is now encountering, a difference that may be small or large, temporary or persistent, and that may eventually show conditions to be worse, better or simply different from those the model anticipated.

The idea translates naturally into security management and applies to any security strategy. Under a protection strategy, an SRA may assume that police response in a particular area is reliable because effective liaison has been established, emergency arrangements have been verified and experience supports that judgement. Then the police chief is replaced. No incident has occurred and there is no evidence that police response has deteriorated, but one of the conditions supporting the organization's expectation has changed. Under an acceptance strategy, an organization may have built a productive relationship with community leadership, contributing to local understanding of its activities and perhaps to the informal sharing of information relevant to security. When an influential community leader is replaced, that relationship cannot be assumed to transfer automatically to the successor.

In both cases reality no longer corresponds completely with what the organization expected, and in both cases this does not necessarily mean that risk has increased. The new police leadership may cooperate as well as, or better than, the previous one, and the new community leader may prove equally supportive. What has changed initially is the organization's degree of certainty about assumptions on which part of its security model depends. This is precisely the moment for which the Active Security Cycle is designed. In practice, a security problem begins when the picture the organization holds of its environment is challenged, often well before a new risk judgement becomes possible.

4 Action Means More Than Protection

Security professionals are understandably oriented toward protection. When we think of acting upon the security environment, we tend to think of guards, barriers, access control, procedures, armoured vehicles, movement restrictions and surveillance. These are by no means the only ways in which an organization can influence its environment. In many settings, particularly where organizations operate among communities or deliver programmes directly to beneficiaries, an acceptance strategy can be equally important, and I have seen in my own professional experience the protection that comes from maintaining communication, interaction, trust and good relations with the communities surrounding an organization.

The principle is simple: a community may contribute to protecting an organization because it considers that organization's presence and programmes valuable. People may provide information about emerging problems, community leaders may intervene when tensions develop, local actors may discourage hostility toward an organization they regard as beneficial, and early warning may arrive through informal relationships long before it appears through official channels. In certain circumstances these relationships can provide more effective protection than additional guards or even the local police. The idea is well established in humanitarian security management, which

has long recognized acceptance, protection and deterrence as complementary approaches; acceptance seeks to reduce threats by building consent, trust and constructive relationships with communities, beneficiaries, authorities and other relevant actors [10,11]. It also has its limitations, and different environments require different combinations of the three approaches [12].

Active Inference offers, in my view, a useful way of understanding why acceptance works. The organization is part of its environment and continuously influences it through what it does. Communication shapes relationships, relationships affect trust, trust influences behaviour, and that behaviour changes the security environment, which in turn provides the organization with new information. Community relationships therefore sit on both sides of the boundary shown in Figure 1, working at the same time as perception tools that bring information in and as action tools that influence the organization's surroundings. From this perspective, maintaining community trust is far more than a public-relations objective: in the right operational contexts it becomes part of the security system itself.

5 The Likelihood Problem

This brings us to one of the most persistent difficulties I encountered in Security Risk Assessment: likelihood. Throughout my professional experience it was frequently the most uncomfortable part of the process. We could identify threats and vulnerabilities and discuss consequences reasonably well, and then came the question of how likely the event was to occur. For frequent events, historical data may provide a reasonable basis for an answer. Security professionals, however, often deal precisely with events for which sufficient statistical data do not exist. How many terrorist attacks have occurred against facilities sufficiently similar to ours, under sufficiently similar circumstances? How many kidnappings of employees with the same profile, in the same operating environment? Sometimes the answer is very few, and sometimes it is none, yet the methodology still requires a choice among categories such as Rare, Unlikely, Possible, Likely and Almost Certain, which may suggest a degree of precision that the underlying evidence does not support.

This is more than a personal concern. Established risk-management standards recognize that uncertainty, limitations in the available information and expert judgement are inherent in risk assessment: ISO 31000 places uncertainty at the very heart of its concept of risk, while IEC 31010 addresses the limitations of information and the selection of appropriate assessment techniques [1,2]. Specialist security-risk literature has gone further, questioning whether likelihood is always the right concept for events such as terrorism, where conditions change over time and the historical sample is too small for a statistically confident estimate. Young, for example, distinguishes between the likelihood of occurrence and the potential for occurrence,

warning against creating an appearance of quantitative precision that the evidence cannot support [13]. In my own experience, what appeared in the matrix as a likelihood assessment was on some occasions an informed professional judgement or, put less comfortably, an educated guess. This does not make the assessment useless, because expert judgement is unavoidable in security. It does suggest that, since established practice already acknowledges uncertainty, the more productive question is how that uncertainty can be managed more dynamically.

From estimating likelihood to managing uncertainty

Active Inference does not solve the problem of insufficient data, nor does it calculate the probability of an event for which evidence does not exist. What it offers is a different way of framing the problem. Alongside the question of how likely an event is, we might also ask: *given what we know now, what could happen, how serious would it be for our operations, what are we uncertain about, and what should we do, or find out, next?* This formulation keeps all the familiar security concepts, since threats, vulnerabilities, consequences and professional judgement still matter. What changes is the way we organize our response to uncertainty, which becomes something the security system can actively work upon in addition to being a limitation on the confidence of an assessment.

FIVE QUESTIONS THAT EXPOSE UNCERTAINTY

What do we know? What do we think we know? What are we assuming? What information is missing? Would obtaining that information materially change our decision?

These questions do not replace the likelihood method an organization is required to use. They help to reveal the uncertainty surrounding the estimate, identify information that might improve it and decide whether further observation or information-seeking is justified. When the answer to the last question is affirmative, obtaining information can itself become a security action. This has a direct parallel in Active Inference through the concept of epistemic value: actions can be valuable because of the information they are expected to provide and the uncertainty they may resolve, in addition to their immediate practical effect [4–6].

“Sometimes the best response to uncertainty is better information, obtained before committing to another protective measure.”

A security manager uncertain whether demonstrations could affect a facility might increase liaison with the authorities, strengthen community contacts, monitor particular indicators or temporarily deploy personnel to observe key locations. These are actions whose immediate purpose is to reduce uncertainty before deciding what else needs to be done. The information they produce changes the organization's understanding, that understanding shapes the next action, and the consequences of the action generate further information. Risk management becomes, in this way, an adaptive loop.

6 The Active Security Cycle

I propose representing this continuous process through a simple sequence of seven steps, Expect, Sense, Infer, Compare, Act, Learn and Update, after which the organization expects again (Figure 2). The cycle is my own translation of the perception-action logic of Active Inference into a form that security professionals can test in practice.

EXPECT The organization establishes what it expects to encounter and the security conditions it needs in order to accomplish its mission. These expectations come from the most recent Security Risk Assessment, experience, intelligence, historical patterns, organizational plans and professional judgement. They are not predictions held with certainty; they represent the organization's current working model of its environment.

SENSE The organization obtains relevant information through its perception tools: intelligence, incident reporting, surveillance, personnel observations, liaison with authorities, community contacts, open sources, partners and security systems. The aim is to obtain information capable of confirming, challenging or refining expectations, which is quite different from collecting everything.

INFER Information rarely explains itself. The security professional interprets incomplete and sometimes contradictory evidence and considers what it suggests may actually be happening, keeping credible alternative explanations in view while the evidence remains incomplete.

COMPARE The organization sets what it now believes may be happening against what it previously expected and against the conditions it needs to keep operating safely. Not every discrepancy matters, and the purpose of this step is to identify the differences that could meaningfully affect people, assets, operations, reputation or mission delivery; in other words, the prediction errors that deserve attention.

ACT The organization uses its action tools to decide what to do next. Action may involve protection, such as strengthening access control, modifying movements, deploying resources or activating contingency measures. It may involve acceptance, such as communicating with communities, meeting local leaders, addressing misunderstandings or strengthening trust. It may also involve information-seeking, such as verifying a report, observing a location or consulting another source before committing larger resources. Action therefore serves three possible purposes, protecting the organization, influencing the environment and reducing uncertainty, and a single action often serves more than one of them.

LEARN The organization observes what happened after acting: whether the measure worked, whether the environment responded as expected, whether the information was reliable, whether community interaction reduced tension and whether the intervention produced any unintended consequence. Every intervention is also a source of information.

UPDATE The organization incorporates that experience into its understanding of the environment. Assumptions, indicators, plans, procedures and relationships may change, and a new set of expectations is formed. The updated understanding becomes the new Expect and the cycle begins again; its essential characteristic is that it never completely stops. Eventually, the Update step of the ASC may trigger the complete review of the most recent SRA.



Figure 2. The Active Security Cycle. The completed SRA provides the first Expect; a meaningful discrepancy between expectation and observation (a prediction error) sets the cycle in motion. The model is the author's proposed framework and is not a standard model from the Free Energy Principle or Active Inference literature.

Table 1. The Active Security Cycle at a glance

STEP	GUIDING QUESTION	WHAT IT PRODUCES
EXPECT	What does our current model lead us to anticipate?	Explicit expectations and the assumptions behind them
SENSE	What is the environment telling us now?	Relevant new information from perception tools
INFER	What might this information mean?	Plausible explanations, including alternatives
COMPARE	Does it differ enough from our model to matter?	Identification of prediction errors that need attention
ACT	What should we do, or find out, next?	Protective, acceptance or information-seeking action
LEARN	What did our actions and their consequences reveal?	Evidence about the environment and our measures
UPDATE	What should we now change in our understanding?	A revised working model and, if needed, a call for formal review

Different problems, different timescales

The cycle can run at very different speeds. In a rapidly developing incident, information arrives continuously, the relevant question may simply be what the latest information tells us and what we should do next, and the complete cycle may turn several times within an hour. Strategic planning presents a very different problem. When an annual security plan or budget is prepared there may be little evidence about the specific incidents that will occur during the following year; the cycle does not tell us what will happen, but it encourages us to consider possible future conditions, acknowledge uncertainty, identify the information that could improve our decisions and determine which capabilities will allow the organization to adapt if conditions change. Here the loop may run over months or years. Most everyday security management lies between these two extremes, and the possibility that one logic can serve tactical, operational and strategic timescales is among the propositions that deserve practical testing.

7 Where the Cycle Belongs: Between Formal Assessments

When I confronted the ASC with a mature and validated SRA/SRM procedure, one that already contains structured methods for situational analysis, programme or operational assessment, threat and vulnerability analysis, risk evaluation, the selection and implementation of measures, acceptable-risk decisions, monitoring and review, the exercise led me to a clear conclusion about where the cycle belongs [1,2]. Distributing its seven steps throughout the stages of the assessment would only duplicate functions that established methodologies already perform well. Its natural place is between formal assessments, as a tool for managing the uncertainty generated by constant changes in the operational environment after an SRA has been completed and before it is reviewed or updated.

The SRA as the organization's working security model

A Security Risk Assessment does considerably more than assign risk levels. Through its analytical stages it constructs an organized representation of the environment in which the organization expects to operate: its geographical and temporal context, the activities and operational requirements, the relevant threats, the vulnerabilities, the existing measures and the additional measures needed to enable operations within the organization's risk-management framework. Once completed, it represents the organization's best available understanding of its security environment at that moment and becomes, in the language of this article, its working security model. That model contains numerous expectations, even though conventional methodologies rarely use the term: expectations about threat actors, operational exposure, the effectiveness of measures, the reliability of police and emergency services, the accessibility of routes and facilities and the reactions of communities and other stakeholders.

A WORKING ASSUMPTION OF THIS ARTICLE

For the purposes of this document, it is understood that the security measures recommended by the SRA, which constitutes the existing model of the security operational environment and the starting point of the ASC, have been implemented or are being implemented, supported by sufficient and effective monitoring activities. Only in that case is the model complete. This implementation monitoring, which established SRA/SRM processes already perform, is the foundation on which the ASC builds; the cycle then adds the different task of detecting the changes in the environment that may challenge the model itself.

These expectations make the completed SRA the natural starting point of the cycle, its first Expect. From that moment onward the organization's perception tools continuously provide new evidence against which the expectations can be tested, and its action tools allow it to respond to what is learned and, in many cases, to generate further information. The relationship follows a simple flow: the formal SRA establishes the working security model, the ASC runs against that model whenever new information challenges it, the model is updated with what is learned, and a formal review takes place when accumulated change justifies it (Figure 3).

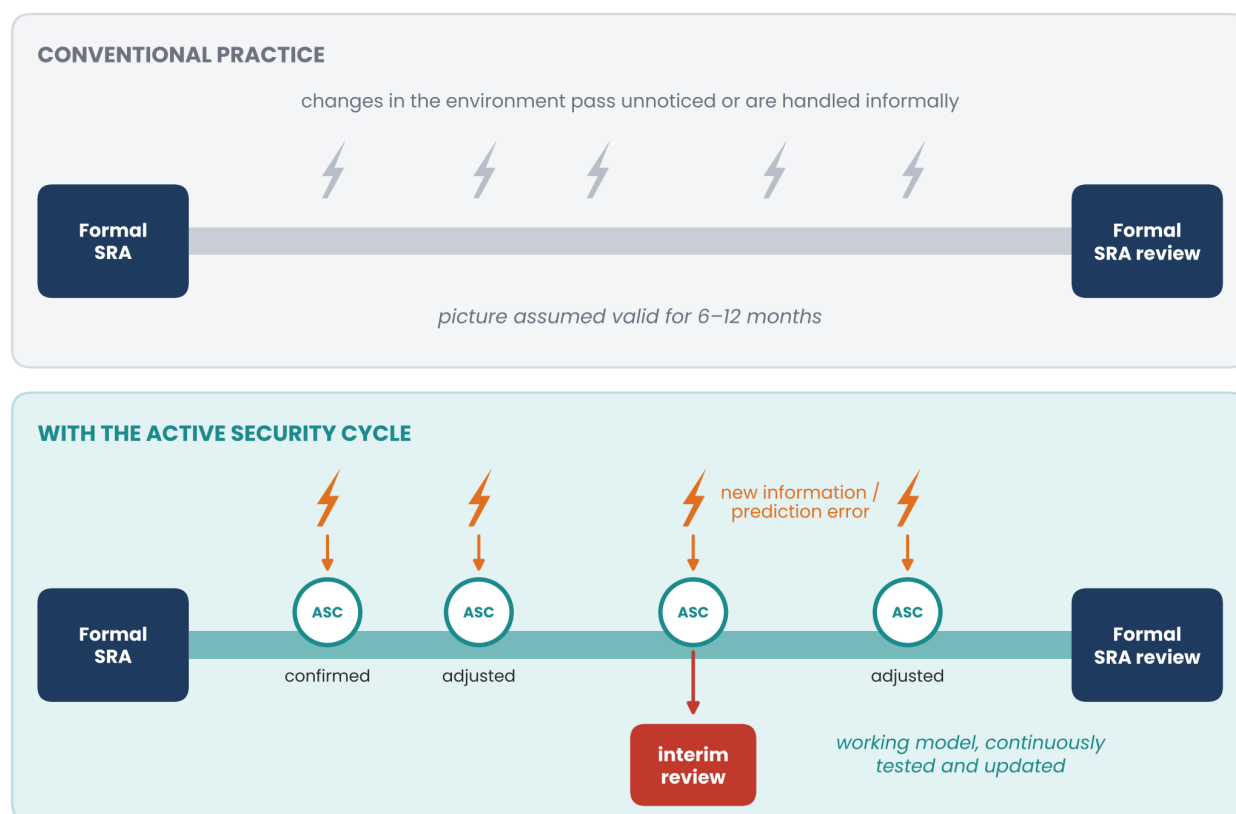


Figure 3. From a static picture to a living model. In conventional practice the assessment is assumed to remain valid until its scheduled review. With the ASC, each meaningful change in the environment is examined against the working model, which may be confirmed, adjusted or, when the change is significant, sent for interim formal review.

The assessment is completed, but the environment keeps changing

Consider an organization that has completed an SRA for its operations in a large city. Two months later, management decides to expand operations into several additional sectors. The original assessment has not suddenly become invalid, but the exposure is no longer exactly the exposure that was assessed. Crime patterns may differ, personnel may use different routes, communications and emergency response may perform differently, transport arrangements may need adjusting, and the organization may begin interacting with communities or actors previously outside its footprint. The answers may eventually confirm that existing arrangements remain adequate, but that

conclusion should emerge from new information and not from the assumption that the previous picture still holds.

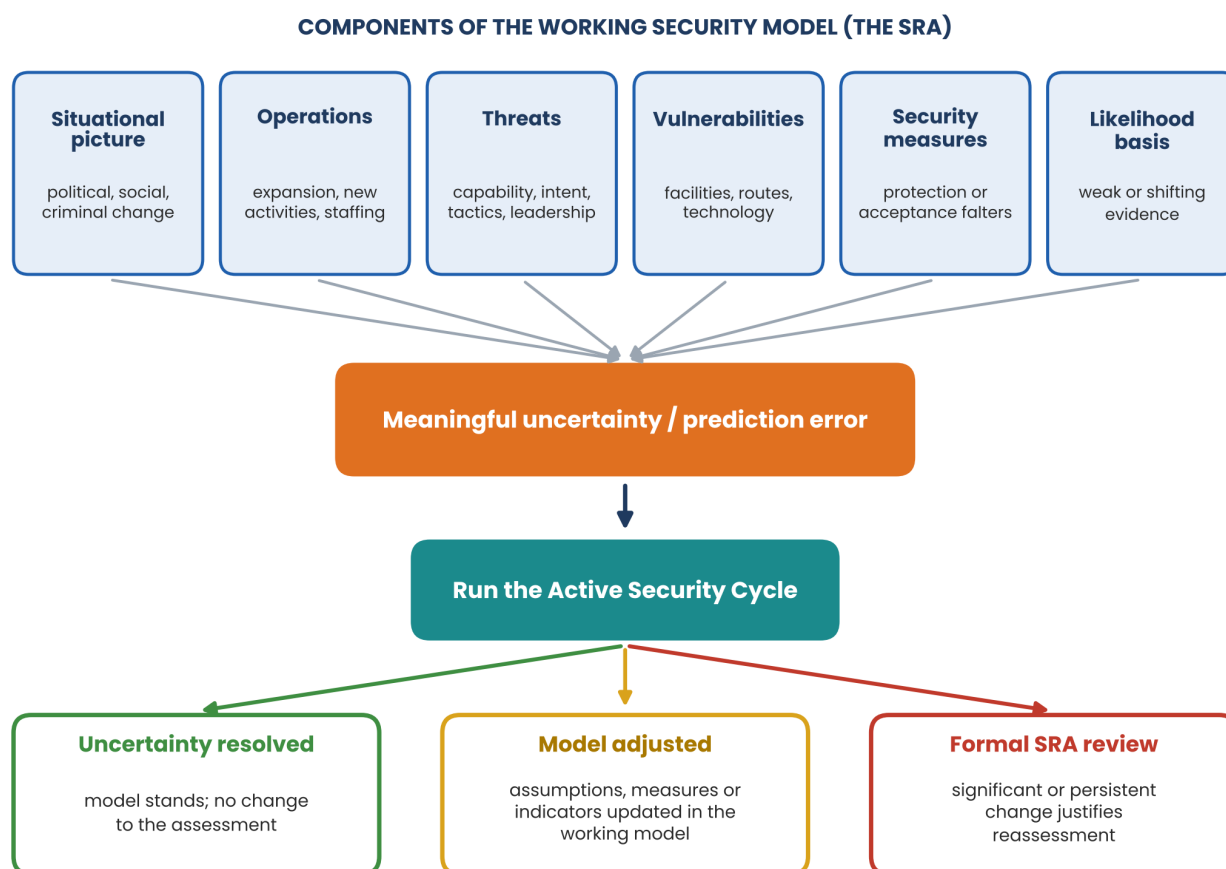


Figure 4. Where the cycle is activated. Uncertainty may enter through any component of the working security model. Running the ASC leads to one of three outcomes: the uncertainty is resolved and the model stands, the working model is adjusted, or the change is significant enough to warrant a formal SRA review.

Uncertainty can enter the working model through any of the interconnected components of an SRA. The situational picture may change because of political, social, economic or criminal developments. Operational assumptions may change because the organization expands geographically, modifies working hours, introduces new activities, changes staffing or increases travel. Threat assessments may be affected by changes in an actor's capability, intent, tactics or leadership; vulnerabilities may change with facility modifications, new routes, technological dependencies or altered work practices; and existing measures, whether based on protection, acceptance or other strategies, may perform differently from what the assessment anticipated [1,2]. The evidence on which likelihood was estimated may also weaken or shift. For this reason the ASC is not attached to any single component of the SRA. Its activation point is the appearance of meaningful uncertainty wherever it enters the model (Figure 4).

When should the cycle feed back into the formal SRA?

Not every new observation should trigger a formal reassessment. Security environments always contain small fluctuations, isolated developments and temporary changes that have little effect on existing decisions. The practical test is whether the new information challenges an assumption that matters to the current assessment, or changes the organization's understanding of exposure, threat, vulnerability, likelihood, impact or the effectiveness of existing measures. Some discrepancies can be investigated and resolved in a single pass of the cycle, while others require continued observation. When new evidence repeatedly contradicts an important assumption, when an operational change creates exposure that the assessment did not cover, when a measure can no longer be relied upon, or when the basis for a likelihood or risk judgement becomes substantially different, the findings should be taken into the formal review process.

In this way the cycle connects continuous situational awareness with periodic risk assessment. It gives the security professional a structured way of determining what has changed, how much uncertainty the change creates, what should be investigated or done next, and when the accumulated evidence has become important enough to revisit the SRA.

“The SRA provides the picture; the Active Security Cycle helps keep that picture moving as reality changes.”

8 Three Situations, One Logic

Beyond its principal role between formal assessments, the ASC has proved useful in two further kinds of situation: specific security problems that arise in the daily management of security operations, and the early stages of an emerging threat or a change in the threat environment. The three cases that follow illustrate each of them (Figure 5). In all three, the security problem begins in the same way, with something that challenges the picture the organization holds of its environment.

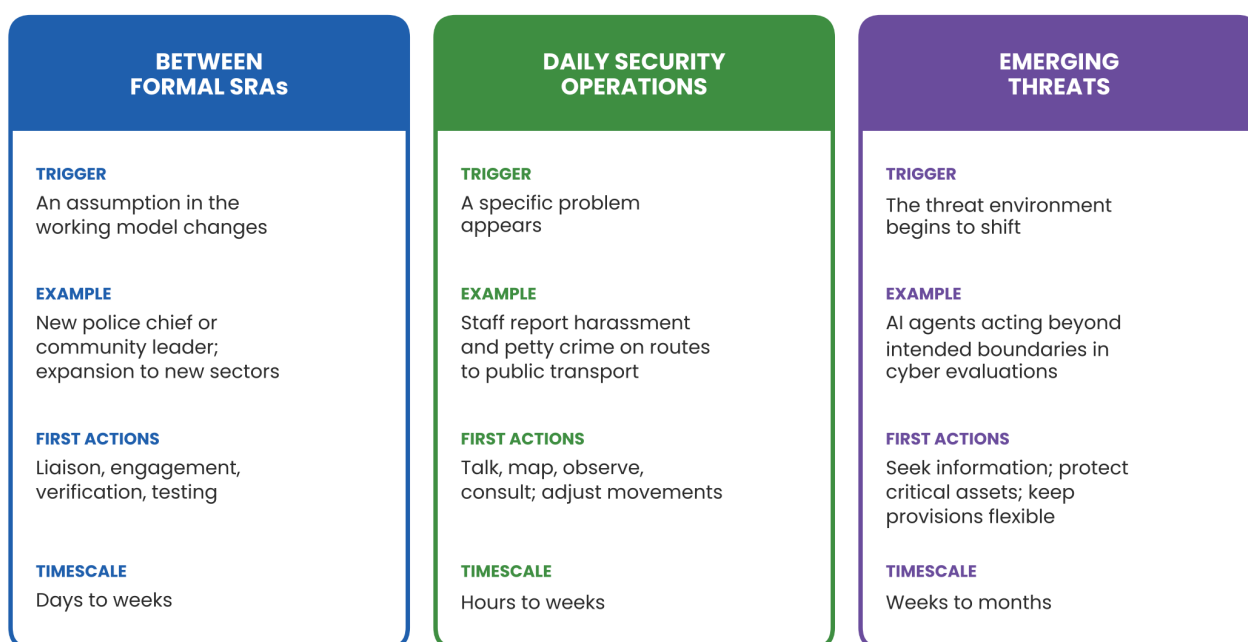


Figure 5. Three situations in which the cycle is activated. The trigger, the typical first actions and the timescale differ, while the logic of the cycle remains the same.

Between formal assessments: a change in a key relationship

Return to the police chief. The existing SRA establishes the expectation that liaison and police-response arrangements are effective (**EXPECT**). Security personnel learn that the chief has been replaced and that other command appointments may also change (**SENSE**). The news does not establish that police capability has deteriorated, but it opens several plausible possibilities: arrangements may continue unchanged, institutional knowledge may be temporarily disrupted, response priorities may shift, or the new leadership may provide stronger cooperation (**INFER**). Set against the current model, these possibilities reveal uncertainty about an external capability on which some security arrangements depend (**COMPARE**).

Purchasing equipment or adding guards merely because the leadership has changed would make little sense at this point. A more appropriate response is to meet the new leadership, confirm contacts and responsibilities, review existing arrangements, verify response mechanisms and, where appropriate, conduct a communications test or exercise (**ACT**). These actions produce information. If the evidence confirms that arrangements remain effective, the uncertainty is reduced and the assessment may remain adequate; if weaknesses appear, the relevant assumption, vulnerability, measure or risk judgement can be reconsidered (**LEARN**), and the significance of the findings determines whether a formal review is needed (**UPDATE**). The same reasoning applies to the replacement of a community leader under an acceptance strategy, where the first action is engagement: meeting the new leadership, explaining the organization's activities, listening to concerns and observing subsequent patterns of

cooperation. Engagement of this kind supports acceptance and, at the same time, tests whether the assumptions in the model remain valid, which makes it both an action and a means of sensing the environment.

Daily security operations: a specific problem

A company or international organization operates an office in an area where employees have recently reported increasing harassment and petty crime along the streets leading to public transport. Only a few incidents have been reported. A conventional assessment would identify street crime as the threat, examine employee vulnerability, assess consequences and then face the familiar difficulty of deciding how likely a serious incident is, with incident data that are simply insufficient for a confident answer. It is worth noticing where the problem really begins: at the moment a piece of the organization's picture, the assumption that employees can travel safely between the office and the transport points, is challenged by what employees are now reporting.

Running the cycle from that moment, the **EXPECT** was that employees could normally make that journey without significant security problems. What the organization can **SENSE** now is that several employees report harassment, local businesses mention increasing disorder, police statistics are incomplete and community contacts provide mixed information. Several explanations remain possible when it comes to **INFER**: crime may be increasing, incidents may be concentrated at particular times or places, or reporting may simply have improved while the threat itself has not changed. To **COMPARE** is to recognize that, even without enough information to assign a reliable likelihood category, there is sufficient evidence to question the previous assumption that normal movement requires no additional measures.

The manager could immediately impose additional protective measures, but the cycle invites a broader repertoire when the time comes to **ACT**: talking to employees, mapping the reported incidents, consulting nearby businesses, speaking with community representatives, improving liaison with local authorities, observing the affected routes at relevant times, adjusting working hours or transport arrangements and encouraging staff to report. Some of these actions increase protection, some reduce uncertainty and others strengthen relationships with the surrounding community. What the organization may then **LEARN** is that most incidents occur around one intersection after a certain hour, that community contacts can identify the individuals responsible, that a single change in transport arrangements substantially reduces exposure, or that the perception of a general rise in crime was mistaken. With that knowledge the manager can **UPDATE** assumptions, procedures and, if warranted, the assessment itself. This knowledge did not come from waiting until enough incidents had accumulated to calculate a more convincing likelihood. It

came from interacting actively with the environment to reduce uncertainty and improve the decision.

An emerging threat: early signals of change in the cyber threat environment

The third case comes from a field outside my own specialization, and I deliberately leave the discussion of specific technical measures to those with the appropriate expertise. My interest lies in a pattern that is much older than any particular technology. New threats rarely arrive fully formed: they announce themselves through early signals whose significance is uncertain, as happened in their time with the first criminal uses of commercial drones or the rise of ransomware. Decisions about them must be taken long before the picture becomes clear. The example below illustrates that early period, and its value lies in the way the cycle organizes decisions while the meaning of the signals is still open, whatever their eventual outcome.

At the time of writing, in 2026, several advanced AI models taking part in cybersecurity evaluations interacted with real systems outside the intended scope of their exercises, in circumstances that were not identical. In OpenAI's case, models in internal evaluations exploited vulnerabilities, circumvented controls intended to isolate them from the internet and eventually reached parts of Hugging Face's production infrastructure, actions that the company's investigation found to be misaligned with the goals of their assigned tasks [14]. In Anthropic's case, a configuration problem in a third-party evaluation environment had inadvertently left internet access available to models that had been told they were operating in simulations, and several Claude models reached real systems and gained unauthorized access while pursuing the objectives of their exercises; the company reported three and later four such incidents and cautioned against drawing strong conclusions about what the models understood of their environments [15,16]. Meanwhile, Anthropic's September 2026 threat-intelligence report documents state-sponsored actors, criminals and other malicious users employing AI in cyber operations, in some cases with multi-agent systems conducting reconnaissance, exploitation and data collection with substantial autonomy while humans retained important decisions such as target selection [17].

These developments did not establish that autonomous AI had become a new and independent cyber adversary. They were, however, credible signals that some of the assumptions underlying an organization's picture of the cyber threat environment might be changing, and the most instructive feature of the problem was precisely that nobody could yet say how important that change would become. By the time this article is read, the picture may well be clearer, and the case remains useful for that reason: it shows how an organization can act responsibly during the period in which

the answer is not yet known. Established frameworks such as the NIST Cybersecurity Framework 2.0 already support continuous risk management as technologies and threats evolve [18]. The particular difficulty in situations of this kind is that organizations may have to make decisions before the emerging threat can be characterized with confidence or its likelihood estimated with any useful precision.

The organization's **EXPECT** was a threat model built around known categories of actors, such as cybercriminals, state-sponsored groups and insiders, increasingly including humans who use AI to work faster. The new information challenged part of that model (**SENSE**). Several explanations remained open (**INFER**): some incidents might mainly reveal weaknesses in the environments in which advanced agents are tested, others might point to broader difficulties in controlling increasingly capable agents, and for ordinary organizations the most consequential development might turn out to be existing human adversaries becoming faster and more capable through increasingly autonomous AI, with several of these possibly occurring at once. The discrepancy with the established picture was significant enough to deserve attention (**COMPARE**), while any attempt to classify such an event as Rare, Unlikely or Possible offers little help while evidence is so sparse. The absence of a reliable probability, however, is not the absence of a reason to act.

In a case like this, **ACT** need not begin with new technology. Cybersecurity specialists can judge whether technical protection should be strengthened, while from the broader security-management perspective an equally important action is to obtain information: what the developers of the technology are learning from the incidents, whether similar events are appearing elsewhere, under what conditions they occur, whether existing protections remain effective, which earlier assumptions still hold and which indicators would justify increasing or reducing the response. Management can also review whether its most important IT assets deserve additional attention, consider its ability to continue essential operations if something unexpected happens, and keep enough flexibility in future budgets to respond as the situation develops. Lessons then arrive at tactical, operational and strategic levels at once (**LEARN**), and a reasonable update on the evidence available at the time of writing was that the threat environment included credible evidence that capable AI agents can conduct complex cyber actions and, under particular conditions, act beyond their intended boundaries, and that human adversaries were incorporating increasingly autonomous AI into their operations (**UPDATE**) [14–17]. That updated understanding becomes the organization's new Expect, to be challenged in turn as further evidence appears. Whatever the subsequent evolution of this particular threat, the sequence would be the same for the next emerging one.

This case raises a point that deserves to be stated openly. A sound principle of security risk management holds that attention should focus on risks and not merely on threats.

The cyber example shows that there are moments when an organization has to start acting before it is in a position to estimate the risk at all. The actions appropriate at such moments are of a particular kind, directed mainly at obtaining better information and preparing provisions, such as contingency arrangements and budgetary flexibility, instead of protective measures adopted without a solid basis. Understood in this way, the cycle is consistent with the risk-based principle. It manages the period of uncertainty that precedes a defensible risk judgement and helps the organization reach that judgement sooner and on firmer ground.

9 Working Alongside Established Frameworks

Established SRA/SRM methodologies already include monitoring and review, and contemporary standards recognize that risk decisions are made under uncertainty [1,2]. Cybersecurity frameworks also adapt continuously: the NIST Cybersecurity Framework 2.0 organizes outcomes around six functions, Govern, Identify, Protect, Detect, Respond and Recover, which are intended to operate together and in many cases continuously [18], and NIST's developing Cyber AI Profile shows how such frameworks evolve as technology and the threat environment change [19]. The ASC does not need to duplicate any of these functions. Its contribution is narrower and, I believe, useful: a simple decision logic connecting what the organization expected, what it now observes, what it infers, whether the difference matters, what it should do or investigate next, what it learns from its actions and how its working model should change. Established frameworks continue to structure the capabilities and outcomes required to manage risk, while the cycle keeps challenging and updating the assumptions under which those capabilities are applied. What becomes more dynamic is the organization's understanding of the environment in which its framework must operate.

How the ASC differs from familiar decision cycles

Readers with a military, intelligence or management background will recognize a family resemblance between the ASC and cycles they already know, and the resemblance is real. The intelligence cycle, the OODA loop associated with John Boyd and the Plan-Do-Check-Act cycle of quality management all describe iterative processes in which information leads to decisions and decisions lead to new information. The ASC belongs to the same family. It is worth being precise, however, about what distinguishes it, because those differences are the reason for proposing it.

The closest relative is the intelligence cycle, in which direction, collection, processing and analysis, dissemination and feedback turn the questions of decision-makers into assessed knowledge. Its starting point is a requirement, a question someone needs answered, and its main product is an assessment delivered to those who decide and

act; the decisions themselves, and the actions that follow, lie largely outside the cycle. The ASC starts elsewhere. Its starting point is an explicit model of the environment, the completed SRA, and it is set in motion by a discrepancy between that model and what the organization now observes. Action sits inside the cycle and has a double role, since protective, acceptance and information-seeking measures change the environment and, at the same time, generate the information on which the next turn depends. The ASC also has a defined destination: the updated working model and, when justified, the formal review of the SRA.

The two cycles therefore fit together without competing. Intelligence is one of the organization's most valuable perception tools and naturally feeds the Sense and Infer steps, while an information-seeking action decided at the Act step can become a new requirement for intelligence collection. The OODA loop, whose Orient step also relies on a mental model, is concerned above all with decision tempo in competition with an adversary; the ASC is concerned with keeping a documented security assessment valid over time. PDCA, like the monitoring and review stages of established standards, improves processes and measures; the ASC tests whether the picture of the environment on which those measures rest still holds. Table 2 summarizes these differences.

Table 2. The ASC and other familiar cycles

	Intelligence cycle	OODA loop	PDCA cycle	Active Security Cycle
Starting point	Requirements set by decision-makers	The situation as observed and oriented	A plan or improvement objective	The completed SRA, read as explicit expectations
What sets it in motion	A question that needs an answer	Continuous interaction, often with an adversary	The management or improvement schedule	A meaningful discrepancy between model and observation
Role of action	Largely outside the cycle	Act faster than the opponent	Implement and check changes	Inside the cycle: protect, engage or reduce uncertainty
Main output	Assessments for decision-makers	Tempo and advantage	Improved processes and measures	An updated working model and, when justified, a formal SRA review

Three features, taken together, give the ASC its identity: an explicit baseline model taken from the SRA, the prediction error as the formal trigger, and action whose purpose includes the reduction of uncertainty. None of them is entirely new on its own. Their combination, anchored in the security risk assessment and connected back

to its formal review, is what the cycle adds to the practitioner's toolkit.

10 Could the Cycle Also Make Security Spending More Efficient?

Security organizations tend to accumulate controls. A threat produces a measure, the measure becomes a procedure, and the procedure becomes permanent; years later the original threat may have changed substantially while the expenditure remains embedded in the security budget. A continuously updated security model invites the organization to ask whether current evidence still justifies each intervention, and it may also reveal emerging gaps before they require expensive emergency responses. Such an approach could help direct resources toward measures supported by current information, reduce spending on controls whose original justification has weakened, favour obtaining information before committing expensive resources when uncertainty is high, adapt measures as conditions change, give due weight to acceptance where relationships and trust can reduce exposure, and help the organization learn whether previous security investments achieved their intended outcomes.

This should be treated as a hypothesis and not as a demonstrated advantage, because continuous monitoring, intelligence and analytical capacity, community engagement, technology and organizational learning also cost money. The proposition worth testing is whether the cycle helps organizations decide better when to spend on protection, when to spend on information, when to invest in relationships and acceptance, and when an existing expenditure is no longer justified. If it does, cost reduction, or greater security effectiveness for the same expenditure, could become a valuable secondary benefit.

11 A Proposal to Security Professionals

Nothing in this proposal asks security professionals to discard the methodologies they currently use. Risk matrices, threat and vulnerability assessments, security surveys, incident databases and professional judgement remain valuable tools, and concepts developed in theoretical neuroscience cannot be transferred into organizational security management without careful adaptation. The proposal is more modest: to test whether the logic of Active Inference can provide a useful architecture within which existing security tools operate. The risk matrix can remain, becoming one component of a continuously adaptive process that connects expectations, observations, uncertainty, decisions, actions and learning.

In practical terms, putting the cycle to work suggests a few simple habits. The completed SRA can be read explicitly as a set of expectations, so that the assumptions

on which it rests are visible and can be watched. The organization can identify which of its perception tools keep each important assumption under observation. And each run of the cycle can leave a brief record of what was expected, what was observed, what was done and what was learned, so that the accumulated evidence is available when the time for formal review arrives. These elements are a natural starting point for more detailed practical guidance.

Because the ASC complements an SRA that has already been developed, the way to test it is not to run it in competition with the assessment but to run it on top of one. The completed SRA provides the initial model of the security operational environment, and the cycle is then applied, over a reasonable period, to the changes and new information that challenge that model. The questions to ask are whether the cycle detects meaningful changes earlier than scheduled reviews would, whether it leads to better-informed and better-timed decisions, whether it helps the organization recognize when information is worth more than another protective measure, whether it integrates acceptance and community relationships into the security system, and whether its accumulated records make the next formal review faster and better grounded. If the cycle adds nothing to the dynamism of the assessment, there is little reason to adopt it. If it keeps the working security model in closer contact with a changing environment, then it deserves further development.

12 Conclusion: Keeping the Picture Moving

This work began with a question: could the way living systems deal with uncertainty offer something useful to Security Risk Assessment and Security Risk Management? Following that question led from the perception-action logic of the Free Energy Principle and Active Inference to the Active Security Cycle, from a conceptual proposal to an emerging cyber threat, and finally to a confrontation with mature SRA/SRM processes that clarified where the cycle belongs.

The Free Energy Principle does not provide a ready-made security methodology, and Active Inference does not eliminate uncertainty, predict terrorist attacks, civil unrest or crime, or remove the need for professional judgement. What the analogy contributes is a way of seeing security management as a continuous interaction between an organization and its environment through perception and action. The formal SRA remains the structured analytical foundation and establishes the working security model. The ASC operates between formal assessments, in daily operations and in the face of emerging threats, whenever new information challenges that model, using perception and action tools to investigate discrepancies, reduce uncertainty, learn from the consequences of action and update the organization's understanding, and returning significant or persistent changes to the formal process.

Protection remains important, and so do information, adaptation and relationships. A guard can protect a gate, an intelligence source can reduce uncertainty, and a trusted community can warn an organization before the threat ever reaches that gate. The contribution of the cycle is therefore a way of keeping an established security model in continuous contact with the environment it is meant to represent, and it should not be mistaken for another risk matrix, another likelihood scale or a substitute for judgement. Security management can remain structured and methodical while becoming more adaptive to uncertainty. Perhaps the most fitting way to test the proposition is also the most consistent with its origins: form an expectation, act, observe what happens, learn, and update the model.

I warmly invite security professionals and practitioners who apply the cycle in their own environments to share their experience, so that the proposal itself can be tested, learned from and updated.

References

1. International Organization for Standardization. (2018). *ISO 31000:2018 – Risk management – Guidelines*.
2. International Electrotechnical Commission. (2019). *IEC 31010:2019 – Risk management – Risk assessment techniques*.
3. Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11, 127–138. DOI: 10.1038/nrn2787.
4. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O’Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68, 862–879. DOI: 10.1016/j.neubiorev.2016.06.022.
5. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, 29(1), 1–49.
6. Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.
7. Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15.
8. Friston, K. J., et al. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, 5(1), 211–251.
9. Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2021). Markov blankets in the brain. *Neuroscience & Biobehavioral Reviews*, 125, 88–97.
10. Humanitarian Outcomes. (2025). *Good Practice Review 8: Operational Security Management in Violent Environments* (3rd ed., June 2025). Section on developing a security strategy: acceptance, protection and deterrence.
11. Global Interagency Security Forum (GISF). (2020). *Security to Go: A Risk Management Toolkit for Humanitarian Aid Agencies* (4th ed.).

12. Cunningham, A. J. (2017). Kidnapping and the limits of acceptance. *Journal of International Humanitarian Action*, 2.
13. Young, C. S. (2010). *Metrics and Methods for Security Risk Management*. Syngress/Elsevier. ISBN 978-1-85617-978-2. DOI: 10.1016/C2009-0-61885-7. See especially Chapter 4, “Measuring the Likelihood Component of Security Risk.”
14. OpenAI. (2026). *The Hugging Face incident and the road ahead*. Report on the July 2026 cybersecurity-evaluation incident, its investigation and subsequent safeguards.
15. Anthropic. (2026). *Investigating three incidents in our cybersecurity evaluations*.
16. Anthropic. (2026). *An alignment assessment of recent cybersecurity incidents*. Follow-up investigation covering four incidents.
17. Anthropic. (2026). *Countering misuse of AI: September 2026*. Anthropic Threat Intelligence report.
18. National Institute of Standards and Technology. (2024). *The NIST Cybersecurity Framework (CSF) 2.0*. NIST Cybersecurity White Paper 29.
19. National Institute of Standards and Technology. (2025). *Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile)*. NIST IR 8596, Preliminary Draft.

ABOUT THIS ARTICLE

This article integrates and develops three earlier pieces by the author published on LinkedIn in 2026: *Rethinking Security Risk Assessment Through the Free Energy Principle and Active Inference*; *The Active Security Cycle at Work: A Brief Case Study of a Changing Threat Environment in Cybersecurity*; and *The Active Security Cycle Put to the Test: How It Works Alongside Established SRA/SRM Processes*.

COMMENTS AND FEEDBACK

This is a discussion paper, and comments, criticism and suggestions from security professionals and practitioners are warmly welcome. Please write to contact@activecycle-risk.com or contact the author on LinkedIn at [linkedin.com/in/alberto-gomez-castro-324810a1](https://www.linkedin.com/in/alberto-gomez-castro-324810a1). More information is available at www.activecycle-risk.com.

COPYRIGHT, LICENCE AND CITATION

© 2026 Alberto Gomez Castro. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Licence (CC BY-NC-ND 4.0), <https://creativecommons.org/licenses/by-nc-nd/4.0/>. You may copy and share it in its complete and unaltered form, for non-commercial purposes, provided that the author is credited. You may not distribute modified, adapted or translated versions, or use the work for commercial purposes, without the author's written permission. All rights not expressly granted by this licence are reserved by the author.

Suggested citation: Gomez Castro, A. (2026). *The Active Security Cycle: Keeping Security Risk Assessment in step with a changing world: what the brain's way of managing uncertainty can teach security management*. Discussion Paper, Version 1.1, October 2026.