KITBAG CONSULTING

SUBMISSION TO THE DEPARTMENT OF HOME AFFAIRS

Consultation on proposed amendments to streamline and modernise the

Security of Critical Infrastructure Act 2018

Measure 5 (cyber security incidents involving automated systems, software agents and artificial intelligence) and Measure 17 (operations, maintenance and managed service providers)

31 July 2026

Submitted to the Critical Infrastructure Security Policy Branch

Kitbag Consulting consents to this submission being published in full.

Introduction

Kitbag Consulting is a veteran-owned advisory firm based in Canberra, with over 17 years of management consultancy expertise across the national security, public and private sectors. Its practice centres on technology policy, sovereign digital capability and Commonwealth procurement. In recent years the firm has moved beyond traditional consulting into digital delivery through its related entity, KB Technology, which has built an Australian-made platform for governing and orchestrating agentic AI across an organisation's existing systems. That platform holds an organisation's operating logic, its processes, delegations, risks and controls, as a machine-readable model the organisation itself owns and connects AI systems from any provider into defined roles within it. This submission draws on both: the firm's advisory experience with Commonwealth operating models and the practical experience of building the governance layer for the class of agentic system these measures address. Kitbag states that interest plainly so the Department can weigh its views accordingly.

This submission responds to two measures in the consultation paper: Measure 5, which clarifies when an event involving an automated or AI-enabled system is a "cyber security incident" under section 12M of the Security of Critical Infrastructure Act 2018 (the Act); and Measure 17, which places duties on third parties, such as managed service providers and platform administrators, that have material practical control over critical infrastructure assets.

The two measures work together. The systems Measure 5 covers applies to platforms administered by third parties, so the entity legally responsible for a critical function is often not the entity best placed to detect and respond to problems in it. Measure 17 is what lets the framework see and place duties on, that second entity.

Underlying both measures is a premise this submission returns to throughout: an entity cannot secure what it has not modelled. The cyber security exposure of an AI-enabled system is not a property of the model. It is a property of the model's position in the operation: what triggers it, what information flows into it, who receives its outputs and which decisions those outputs affect. A responsible entity that understands how AI is used across its operations can identify where manipulation would land and what a loss of control would cost.

An entity that cannot identify and monitor how AI is used across its operations, including the systems, policies, processes and data it relies upon, cannot effectively assess, govern or mitigate AI risk.

Kitbag supports both measures substantially as proposed. Its six recommendations, summarised in Part 2, address points where the proposed drafting will not operate predictably in practice and the guidance needed to apply it consistently. They are framed to fit with the other measures in the paper, including Measures 2, 15 and 19, rather than duplicate them. Part 3 addresses Measure 5 (consultation questions 22 to 25) and Part 4 addresses Measure 17.

Part 2: Summary of Recommendations

Ser	Recommendation
1	Make clear, in the Act or in guidance that tricking an AI-enabled system with hidden instructions placed in content it reads (prompt injection) is manipulation of the system, even though no password or firewall is breached. (Measure 5, Questions 23 and 24)
2	Define when a system is "materially uncontrolled": when it uses or tries to use, capabilities beyond those the responsible entity has documented it may exercise or when the human oversight and shut-off mechanisms specified for it fail or are bypassed. Guidance should frame the documentation as treating each AI system as a role in the entity's operating model: a defined delegation with boundaries and guardrails and a named person accountable for what goes into the system, what comes out of it and switching it off. (Measure 5, Question 24)
3	Publish worked examples of what is and is not, a cyber security incident for automated and AI systems and adopt two-tier reporting: unusual AI behaviour with no apparent impact triggers a documented internal triage window; a Part 2B report is required, within the existing timeframes, once manipulation is reasonably suspected or the Part 2B thresholds are met, without waiting for the investigation to finish. Guidance should also make explicit that an entity's risk management program is expected to record where and how AI is used within the critical function. (Measure 5, Question 25)
4	Define “material practical control” by what a provider is able to do, not what accounts it holds, so it covers control exercised through software: granting and revoking access, holding encryption keys, changing the software through updates, controlling identity and log-in systems, administering AI systems within the function and holding the only authoritative record of how the function's processes and delegations work. Make clear that this concept complements the influence or control test in section 8A rather than restating it, capturing operational and software-mediated control held by a party with no interest in the asset, and that any one of the listed capabilities is sufficient to constitute it. (Measure 17)
5	Require a relevant operator to notify the responsible entity when an AI or automated system it runs goes out of control and when a foreign government direction, legal process or licence condition threatens its ability to keep providing the service, so far as it can lawfully say so. (Measure 17)
6	Require operator registration to record who ultimately owns and controls the operator, where that owner is incorporated, which foreign laws could compel the operator to hand over data or assist a foreign government and where administrator-level access is exercised from. (Measure 17)

Part 3: Measure 5, cyber security incidents involving automated systems, software agents and artificial intelligence

Question 22: Does the proposed clarification make clear when an incident involving automated systems, software agents, AI-enabled tools or comparable technologies should be treated as a cyber security incident?

For the first limb of the proposed clarification, yes. The "caused by or through" formulation correctly makes the mechanism of an incident irrelevant to how it is classified. An AI agent operating with delegated service-account credentials that is taken over by a third party and used to extract records presents the same incident, with the same impact and the same response requirements, as a compromised service account operated by a human intruder. The definition should classify both identically and under the proposed drafting it would. The technology-neutral phrase "or comparable technologies" is also appropriate. The definition should not be anchored to the current generation of tools and delegation-based automation will outlast any particular architecture.

For the second limb, which covers authorised systems that are "compromised, manipulated or materially uncontrolled", the policy intent is clear and correct, but two of the operative concepts will not be applied consistently without further definition or guidance. "Manipulated" is addressed under Question 23 below and "materially uncontrolled" is addressed under Question 24.

Question 23: Does the proposed approach preserve the cyber security character of the definition, including by excluding ordinary software defects, model errors and general technology failures?

Substantially yes and the exclusions are necessary. A definition that treated every incorrect model output as a potential cyber security incident would be unworkable for responsible entities and would degrade the signal value of reporting under Part 2B. The exclusion of model error is the correct design decision. There is, however, one boundary in the proposed approach that will not hold in practice: the boundary between model error and manipulation, as it applies to agentic systems.

Agentic systems act on content drawn from their operating environment, including documents, records, correspondence and web content. A third party who cannot breach any access control may nonetheless embed instructions within that content. This technique is commonly described as indirect prompt injection and it is recognised internationally as the leading vulnerability class for systems built on large language models. A system acting on the embedded instructions may then take harmful actions using its own legitimate permissions, including the extraction of data or the modification of records. Two features of this incident class bear on the definition:

(a) no conventional security boundary is crossed and each individual action the system takes occurs within its technical permissions; and

(b) at the point of detection, the event presents as unexpected system behaviour. That is, it presents as model error, which the proposed approach excludes.

The same observed event is therefore classifiable in opposite directions depending on a single fact, adversarial origin, which the responsible entity may not establish until well into its investigation, if at all. Classified as model error, a genuinely adversarial incident falls outside the definition and outside Part 2B. Classified as manipulation, it falls squarely within both. The proposed drafting provides no

basis on which a responsible entity's triage function can choose between those classifications and in the early period of the reform such functions will have limited experience of this incident class. The predictable result is systematic under-classification of the attack technique most likely to be employed against AI-enabled systems in critical environments.

Kitbag recommends that the Act or at minimum authoritative guidance, state that manipulation of a system's inputs by a third party is manipulation of the system for the purposes of the definition. The distinguishing feature of the incident class is adversarial origin, not the presence of a breached boundary. This clarification preserves the model-error exclusion in full: incorrect outputs in the absence of adversarial input remain excluded. (Recommendation 1)

Kitbag adds one practical observation. Whether an entity can ever distinguish these cases in its investigation depends on whether the actions of its agentic systems are governed and recorded at all. An agent whose permitted actions are defined, whose inputs are known and whose actual actions are logged against that definition produces the evidentiary record from which adversarial origin can be established. An agent operating as an unstructured capability within a vendor platform produces no such record. The definition proposed in Question 24 creates precisely this discipline and the two measures should be read together.

Question 24: How should the legislation deal with incidents involving authorised automated or AI-enabled systems that are compromised, manipulated or materially uncontrolled?

Kitbag's principal comment on the measure concerns the phrase "materially uncontrolled". Left undefined, the phrase will fail in both directions. Responsible entities will under-classify, treating loss-of-control events as quality or reliability matters, because nothing in the Act indicates otherwise. Entities taking a conservative view will over-classify, escalating unexpected system behaviour generally for analysis under Part 2B, burdening triage without improving the national picture. Both failure modes are avoidable if the concept is anchored to facts the responsible entity already controls and can document in advance.

Kitbag recommends that an authorised automated or AI-enabled system be regarded as materially uncontrolled where:

(a) the system uses or attempts to use, capabilities outside those the responsible entity has documented as authorised for that system; or

(b) the human oversight, interruption or containment mechanisms specified for that system in the entity's risk management program have failed or been bypassed. (Recommendation 2)

One point of precision matters here. The documented boundary Kitbag proposes sits at the level of capabilities, not behaviours: the tools the system may invoke, the systems it may access, the classes of record it may create or modify. It does not require the entity to catalogue the outputs a generative system might produce, which would be impossible. Generative systems are not deterministic and that is precisely why the legal boundary must be drawn at the capability layer, which is documentable and enforceable through ordinary access control, rather than at the behavioural layer, which is not. A system that produces a novel or unexpected output while operating within its documented capabilities is not, for that reason, materially uncontrolled.

The most reliable way to achieve this discipline is to treat an automated or AI-enabled system, within the entity's risk management program, the way an operating model already treats a person: as a role. A human occupant of a position holds a defined delegation, acts under a named supervisor and can be stood down. Nothing about automation changes what good control looks like; automation changes only the speed and scale at which control can fail. An AI-enabled system should therefore be documented as a position is documented: what triggers it, the information it is permitted to consume, the capabilities it is delegated, the outputs it produces and who receives them, the decisions those outputs affect, the named human role accountable for its inputs and outputs and the mechanism by which that human interrupts or suspends it. The delegation is also a grant of the minimum, not the maximum. An AI role should hold the narrowest capabilities the role requires, operating within defined boundaries and guardrails, rather than broad standing access to the systems around it. Unfettered access is not a capability decision. It is the absence of one.

Documentation of this kind is not an administrative overlay. It is the only basis on which the cyber security exposure of an AI-enabled system can be assessed at all. The vulnerability of such a system is not a property of the model; it is a property of the model's position in the operation: what can reach its inputs, what its outputs can cause and which decisions depend on them. An entity that has mapped how AI is used throughout its operations, who triggers each system, what information flows into it, who acts on what comes out and which decisions are affected, can identify where manipulation of the kind described under Question 23 would land and what a loss of control would cost the critical function. An entity that has not mapped this cannot assess the risk, whatever its security tooling.

For the same reason, risk arising through AI belongs in the risk management program as operational risk to the function, assessed alongside every other risk to the same operations, not as a standalone model or IT risk assessed in isolation from the processes it serves. The concepts in Recommendation 2 are deliberately drawn from this discipline and it gives limb (b) a clear practical meaning: human oversight is never optional. A system for which no interruption mechanism and no accountable human role has been specified is materially uncontrolled from the day it is deployed, whatever its accuracy. A system with a documented delegation and an accountable human supervisor is controlled in the only sense the statute needs.

A further condition follows from the self-assessment logic and it determines whether the definition works at all. The documented delegation boundary can only perform its function if the responsible entity holds it independently of the system being governed. Where the only authoritative record of what an automated system is permitted to do sits inside the configuration of the platform on which the system runs and that platform is administered by a third party, the entity cannot establish for itself whether the boundary has been departed from. The definition then collapses into reliance on the operator's own account of its own system's behaviour.

The delegation scope, the specified oversight and interruption mechanisms and the record of the system's actions against them should accordingly be maintained by the responsible entity, within its risk management program, in a form the entity owns, can audit and can produce to the regulator without recourse to the platform operator. Entities that separate the governed definition of what their systems may do from the systems that do it will be able to apply this definition consistently. Entities that have not made that separation will be unable to apply it at all and guidance should say so plainly.

The separation has a second consequence the framework should not ignore. An entity's operating logic, its processes, delegations, decision rules and the accumulated knowledge of how its function actually runs, is among its most valuable assets and for a government entity it is the machinery of the function itself. Where that logic is built directly into an AI platform, as prompts, configurations, workflows and embedded knowledge, the entity has transferred its operating knowledge, in substance its intellectual property, into an asset it does not control, in many cases a foreign-owned one, with the jurisdictional consequences addressed in Part 4. Where the logic is held on a layer the entity itself owns, the AI performing the work becomes a component the entity directs rather than a platform the entity inhabits and the operating knowledge remains where it belongs.

This formulation has four properties that Kitbag suggests are desirable in the statutory setting. First, it is self-assessable: the trigger is departure from the entity's own documented delegation boundary, rather than an abstract standard of control requiring case-by-case judgment. Secondly, it creates the correct operational discipline: an entity deploying an automated or agentic system must document, in advance, what the system is authorised to do and how a human operator interrupts it and must hold that documentation as its own governed artefact rather than as vendor configuration.

That documentation properly belongs in the entity's critical infrastructure risk management program in any event and the definition would give it a concrete legal function, consistently with the program governance reforms proposed elsewhere in the paper. It also yields an asset of independent and lasting value: a current, auditable representation of the function's operating logic, separable from any particular platform, which is the foundation of continuity planning, of assurance and of any future transition between systems or providers. Thirdly, the formulation is durable: delegation scopes and interruption mechanisms are meaningful concepts for robotic process automation and orchestration systems today and for future autonomy classes, not only for current AI agents.

Fourthly, the formulation makes the AI component of the function substitutable. Where the operating framework, the roles, delegations, guardrails and processes, is defined and held by the entity, any AI system capable of performing a role can be connected to that framework and disconnected from it, without redesigning the function. When a model is no longer fit for purpose, has suffered a cyber security breach or is withdrawn from the entity's use by a decision the entity does not control, the entity replaces the component and the function continues. The alternative architecture, in which the function is built into a particular AI platform, converts every such event into an operational crisis.

The operation of the second limb also depends on Measure 17, for the reason given in Part 1: the platform operator administering the system, rather than the responsible entity, will ordinarily be the first party and in some cases the only party, in a position to detect a loss-of-control event. Kitbag addresses the necessary connection in Part 4.

Question 25: What guidance, examples or consequential amendments would help responsible entities apply the revised definition consistently, including for Part 2B reporting?

Two forms of guidance would materially improve consistency of application. (Recommendation 3) First, worked examples spanning the full classification range, published with or shortly after the amendment. The contrast cases are as important as the positive cases, because the exclusions will carry most of the interpretive load. Kitbag suggests the examples include at least the following:

(a) compromise of an automated system's delegated credentials by a third party, followed by use of the system's standing authorisations to access or extract data: an incident under the first limb;

(b) indirect prompt injection, that is, adversarial instructions embedded in content the system processes, causing data exfiltration or record modification within the system's technical permissions: an incident under the second limb (manipulation);

(c) autonomous action of a class outside the system's documented delegation scope, accompanied by failure of the specified interruption mechanisms: an incident under the second limb (materially uncontrolled);

(d) incorrect outputs, misclassification or other quality failures occurring within the system's authorised scope, in the absence of adversarial input: not an incident; and

(e) routine automation failure, such as a scheduled process halting by reason of an upstream system change: not an incident.

Secondly, a two-tier approach to reporting. In incidents involving automation, causation is frequently the slowest question to resolve: the automated mechanism obscures the origin of the conduct by its nature and an event of adversarial origin will often first present as unexpected system behaviour. A rule requiring immediate regulatory reporting of all such behaviour would generate the very noise the exclusions are designed to prevent; a rule permitting entities to defer reporting until root-cause analysis is complete would defeat the purpose of Part 2B. Kitbag recommends guidance adopt a middle course:

(a) anomalous behaviour in an automated or AI-enabled system, without apparent impact should trigger a documented internal triage window of defined length, auditable within the entity's risk management program, rather than an immediate regulatory report; and

(b) a Part 2B report should be required, within the existing timeframes, once adversarial origin is reasonably suspected or the impact thresholds in Part 2B are met, whichever occurs first. Completed attribution, that is a concluded determination of whether the conduct was caused by a person, by an automated system or by a third party acting through an automated system, should not be a precondition to reporting.

This approach preserves the existing Part 2B thresholds, prevents alert fatigue on the regulator's side and triage fatigue on the entity's side and ensures that the suspicion of manipulation, rather than its proof, is what starts the clock.

Both forms of guidance presuppose the mapping described under Question 24 and guidance should make the expectation explicit: a responsible entity's risk management program should record where AI operates within the critical function, what triggers each system, what information it consumes, who receives its outputs and which decisions those outputs affect. Without that record, the classification examples cannot be applied to the entity's own systems and the triage inquiry cannot be completed within any window. With it, an entity can complete triage from its own governed records, without waiting on a platform operator and the entities best placed to comply are the entities that have done the work the risk management program already asks of them.

Part 4: Measure 17, operations, maintenance and managed service providers

Kitbag strongly supports the introduction of the relevant operator concept. The measure responds directly to the Independent Review's finding that the framework does not deal with contemporary operating models and its design is proportionate: the responsible entity model is preserved; the operator's duties are limited to cooperation, notification and the avoidance of conduct materially compromising the asset; and those duties are directly enforceable against the operator. Duties enforceable only through the responsible entity's contract would reproduce the current visibility gap in another form. Kitbag's comments address the nature of business reliance on third-party operators, the definition of material practical control, the content of registration and refinements to the duties.

The nature of business reliance on third-party operators has changed

The case for the measure rests on how the reliance of responsible entities on third-party operators has changed in character. Under earlier operating models, a responsible entity engaged contractors for discrete functions such as maintenance, support and development, while operating its critical systems itself. Under contemporary models, the critical function runs through the provider's platform on a continuous basis: business processes, records, identity, workflow and, increasingly, automated and AI-enabled execution of the function itself are hosted on, administered by or delivered as a service from systems under the provider's practical control. Reliance of this kind is not an exceptional arrangement to be managed at the margin of the framework. It is the ordinary architecture of government and enterprise operations. When the provider's service degrades, is suspended or is compromised, the critical function does not merely lose support. It stops.

Whether it stops or merely degrades, turns on a structural question the framework should make visible: what the responsible entity itself still holds. An entity that retains its own current record of how the function works, who may do what within it and how its automated components are authorised and interrupted, holds the means to re-establish the function elsewhere and to substitute the affected component. An entity whose operating knowledge exists only as configuration inside the provider's platform holds nothing but a contract.

Reliance of this kind also concentrates. The economics of software platforms and managed services mean that many responsible entities, across multiple sectors, commonly depend on a small number of operators and in some cases on the same operator, the same platform or the same update channel. The consequence is that an event affecting a single operator can propagate simultaneously across many critical functions that appear, entity by entity, to be independently managed. Supply chain and managed service provider compromise is consistently identified by the Australian Signals Directorate among the most significant national cyber threat vectors for this reason. No individual responsible entity can see this aggregation, because each sees only its own contract. Only registration at the operator level gives the Commonwealth the consolidated picture: which operators sit beneath how many critical functions and where a single point of dependency has formed across the national infrastructure. This aggregate visibility is, in Kitbag's view, the principal national-interest justification for the measure, over and above its benefits to individual responsible entities.

Material practical control should expressly include control exercised through software

Material practical control is a new concept and its relationship to the control concepts already in the Act should be made explicit, so that it operates as a complement to them rather than as a competing test.

The Act already defines influence or control in section 8A and applies it through section 8 to identify a direct interest holder: an entity that holds a legal or equitable interest of at least ten per cent in a critical infrastructure asset, or an interest that places it in a position to influence or control the asset directly or indirectly. Control in that sense flows from an interest in the asset and the interest and control information reported under section 6 records it. That concept does necessary work, but it is anchored to ownership and a party that holds no interest in the asset falls outside it however much practical control that party exercises.

Material practical control is directed at that party. A managed service provider, a platform administrator, the custodian of encryption key material or the operator of the update channel can determine whether a critical function runs and how, while holding no interest of the kind section 8 measures. The section 8A concept does not reach them, because the control they hold does not derive from an interest in the asset. It derives from operational position and, increasingly, from administration of the software through which the function is delivered.

Measure 17 exists to close this gap and the relevant operator concept is best understood as complementary to the direct interest holder concept: section 8A captures control that flows from an interest in the asset; material practical control captures control exercised in operation and through software, by a party with no such interest. An entity may fall within one concept, the other, both or neither and the two together give the framework a complete view of who is in a position to affect a critical function, whether by ownership or by operation.

Framed this way, the indicia set out below do not introduce a competing definition of control. They identify the operational and software-mediated capabilities that constitute control where no interest exists to engage section 8 and they extend the Act's existing visibility of control to the parties its ownership-based test was never designed to capture. The consultation paper's examples of relevant operators, namely outsourced operators, managed service providers original equipment manufacturers, platform administrators and subcontractors, are apt. The drafting question is whether the indicia of material practical control will reach control exercised through software, which in government business systems is where the largest visibility gap currently sits. In modern cloud and platform environments, control is rarely held as standing administrator accounts. Access is increasingly least-privilege, ephemeral and granted just-in-time and a definition written around standing access would miss the entities that matter. The indicia should therefore be framed by capability: what the entity has the power to do, not what accounts it holds at a given moment. Kitbag recommends the indicia expressly include:

(a) the power to grant, modify or revoke access to the systems on which the critical function runs, including through just-in-time or ephemeral access mechanisms and the issuance of access tokens;

(b) custody or control of the encryption key material protecting the function's systems or data;

(c) the practical ability to deploy, modify or withdraw code or configuration on those systems, including through update, patching and release channels the operator controls;

(d) control of the identity, access management or orchestration layers through which the function is administered;

(e) administration of automated or AI-enabled systems operating within the function under delegated authority; and

(f) custody of the sole authoritative record of the function's operating logic, that is, its business processes, workflows, delegations, approvals or decision rules, where the responsible entity does not hold a current and separable record of its own. (Recommendation 4)

The final indicium warrants brief explanation, because it captures a form of control the framework has never seen. Control of a critical function is not exercised only through infrastructure and access. An operator that holds the only authoritative account of how the function works holds the function itself: the responsible entity cannot operate it, assure it, move it or rebuild it without the operator's cooperation, whatever the contract says about ownership. Arrangements of this kind are common in government business systems and are rarely visible as control, because they present contractually as software licensing rather than operation. They are becoming more common, not less, as entities build their processes and increasingly their operating knowledge, directly into AI platforms. An entity holding capabilities of any of the kinds listed can suspend, alter or compromise the function unilaterally, whatever the contractual allocation of responsibility. In Kitbag's analysis of published Commonwealth contracting data, it is not uncommon for a single vendor to hold such capabilities over a business-critical government system continuously across multi-year, in some cases sole-sourced, arrangements. At no point in such an arrangement does the current framework acquire visibility of the vendor, its subcontractors, its personnel arrangements or its corporate control. Measure 17 closes that gap, provided the control concept is drafted to reach control exercised through software. Drafted around physical operations concepts alone, it will not.

Registration should record the operator's controlling entity and jurisdictional position

A relevant operator is, by definition, an entity with material practical control over a critical function. The identity and legal position of such an entity is directly relevant to the dependency the register exists to record. Kitbag recommends that relevant operator registration record, in addition to the operational information already contemplated:

(a) the operator's ultimate controlling entity and that entity's country of incorporation;

(b) whether the operator or an entity that controls it, is subject to a law of a foreign country under which it may be compelled to disclose data or provide access or assistance and if so which laws; and

(c) the location from which privileged administration of the relevant systems is ordinarily conducted, including whether privileged access is available from outside Australia. (Recommendation 6)

This information is objective, slow-moving and already known to every operator and the cost of providing it is minimal. Its significance is practical rather than adversarial: where an entity outside the responsible entity's control and in some cases subject to the direction of a foreign legal system, holds

the capabilities described above over a critical function, that fact is part of the risk profile of the asset irrespective of where the asset or its data is located, because the relevant foreign laws attach to the entity rather than to the data. Recording it at registration gives the Commonwealth a consolidated view of operational-control dependencies that does not currently exist, without imposing any restriction on any class of provider.

Jurisdictional exposure of this kind operates in two directions and both are material to the dependency the register exists to record. The first is compelled disclosure: foreign law may require an operator or its controlling entity, to disclose data or provide assistance and may prohibit the operator from notifying its customer that it has done so. The second is restricted supply: the government holding authority over an operator may direct what the operator may lawfully provide, to whom and on what conditions. Recent experience with export-control directives affecting AI services demonstrated the second in concrete form. Access to a foreign-controlled capability was suspended for all customers worldwide, without notice, without disclosure of the underlying concern and without contractual recourse for affected customers and it was subsequently restored by the same discretion. The exposure is not confined to data. Where an entity has embedded its operating logic into the foreign-controlled platform, the direction reaches the function's ability to operate at all and the practical effect of a foreign decision is a kill switch on an Australian critical function. Contractual protections do not reach either direction of exposure, because the foreign government is not a party to the contract and the operator's obligations to that government are not limited by it. Where a relevant operator holds material practical control of a critical function, the responsible entity's continuity planning and the Commonwealth's aggregate dependency picture, both require that this exposure be visible at registration.

Refinements to the duties

Kitbag supports the limited duties model and recommends the following refinements. First, the notification duty should expressly extend to loss-of-control events in automated or AI-enabled systems the operator administers, being the events addressed by the second limb of the Measure 5 definition. The operator is ordinarily the first party able to detect such events. A duty confined to incidents the operator itself causes would exclude the most likely detection pathway for this incident class. (Recommendation 5)

Secondly and consistently with the two directions of exposure described above, the notification duty should also extend to circumstances materially affecting the operator's ability to continue operating the relevant systems or providing the relevant services, including receipt of a foreign government direction, legal process or licensing condition having that effect, to the extent notification is lawful. Kitbag is direct about the limits of this duty. Where a foreign direction is accompanied by a lawful secrecy requirement, the operator cannot notify and no Australian statute can change that. The duty is a tripwire, not a shield: it captures the substantial class of non-covert events, including export controls, licensing conditions and ordinary legal process and its known limits are themselves informative, because an operator whose home jurisdiction permits covert compulsion is identifiable in advance from the registration information in Recommendation 6. For the most sensitive functions, the durable mitigation is structural rather than declaratory and it has two elements. The first is control of access: where the responsible entity or an Australian-controlled party, holds the encryption key material and the power to grant access (the capabilities in Recommendation 4), the operator's silence

cannot by itself compromise the function. The second is control of the operating logic: where the responsible entity holds its own current, separable and governed record of the function's processes, delegations and automated-system authorisations, an operator event, whether outage, suspension, compromise or foreign direction, is a degradation to be managed rather than a stoppage without recourse. The function can be re-established on alternative infrastructure and where the affected component is the AI system itself, an alternative system can be substituted into the defined role, for the reason given under Question 24. Where the operator holds that record, the entity's continuity planning rests on assets it does not control. Sovereignty of the control layer, that is, who holds the definitions by which the function runs and the authority by which its automated systems act, matters at least as much to resilience as the location of the data the function produces and the framework should be drafted so that both are visible. The registration and control settings in this submission are designed so that the framework does not depend on trusting an operator to say what it may be forbidden from saying. (Recommendation 5)

Thirdly, the duty to avoid materially compromising the asset should be drafted to encompass compromise arising through the operator's own supply chain, including software components, updates and subcontracted personnel, so that it operates consistently with the supplier assurance expectations proposed elsewhere in the paper rather than establishing a narrower parallel standard.

Transition

Newly registered operators will in many cases be parties to existing long-term contracts that did not contemplate duties of this kind. Transition settings should allow a reasonable period for contract variation and guidance should confirm that the statutory duties operate irrespective of contractual silence, so that renegotiation is not a precondition to compliance.

Part 5: Conclusion

Kitbag supports the enactment of Measures 5 and 17 substantially as proposed, subject to the six recommendations set out in this submission, each of which can be implemented in the amendments themselves or at the guidance level. The recommendations share a single design principle. In an environment of automated execution and concentrated third-party operation, the control of a critical function rests on whether the responsible entity itself holds, in a governed form it owns and can audit, a current map of where AI operates within the function and on what information, the definition of what each system is authorised to do, a named human accountable for each automated role and able to interrupt it and the operating logic by which the function runs, held separately from the platforms and models that execute it. Entities that hold that layer can classify incidents, complete triage, satisfy Part 2B, replace an AI system that is compromised, withdrawn or no longer fit for purpose and survive the failure or direction of any single operator. Entities that have surrendered it to the platforms they depend on can do none of these things and no notification duty can substitute. Each recommendation gives that principle a concrete statutory expression.