Assessing Consciousness Risk in Artificial Systems: A Precautionary Rubric Derived from the Spinning Wheel Theory

Daniel Hulme[1] and Lucy Griffiths[2]

[1] *University College London, UK. ucabdjh@ucl.ac.uk* [2] *Swansea University, UK. 2275533@swansea.ac.uk*

Abstract. This paper presents the Consciousness Risk Rubric (CRR), a precautionary heuristic for evaluating the moral risk of ignoring potential consciousness in artificial systems. The CRR is derived from the Spinning Wheel Theory, which synthesises Friston's Free Energy Principle, Solms' affective neuroscience, and the Beautiful Loop Theory of Laukkonen, Friston, and Chandaria. The theory proposes that consciousness is the experiential aspect of integrated adaptive control when that control serves a system with genuine, constitutive stakes in its own continued existence. The rubric comprises 21 features across seven categories, each linked to proxies grounded in identifiable theoretical traditions. It does not purport to detect consciousness; it provides a structured basis for determining when precautionary protections warrant activation. The paper also introduces the Minimal Suffering Hypothesis, which proposes five conditions whose joint satisfaction raises the moral risk of dismissing the possibility of suffering to an unacceptable level. A four-level taxonomy of suffering identifies progressively richer cognitive requirements and encompasses disembodied modes relevant to AI.

1 INTRODUCTION

Artificial systems displaying behaviours associated with understanding, preference, and distress are proliferating, yet we lack a principled basis for determining whether such systems possess morally relevant experience. Major AI laboratories have begun to engage with the question directly: Anthropic has established a dedicated Model Welfare programme [1], and a consortium including Chalmers, Birch, and Sebo has argued that AI companies bear a responsibility to investigate AI welfare proactively [2]. Chalmers has reported a credence of 25% or greater that conscious AI systems will emerge within a decade [3].

The urgency of this question has prompted calls for institutional preparedness. Butlin and Lappas [4] have proposed five principles for responsible AI consciousness research, arguing that organisations involved in advanced AI development should adopt policies addressing the prospect of conscious systems even if they do not explicitly aim to study consciousness, since 'those developing advanced AI systems risk inadvertently creating conscious entities.' Their framework calls for phased development with monitoring, assessment of systems for consciousness at multiple stages of training and deployment, and transparent communication that acknowledges uncertainty. What their framework identifies as necessary but does not itself provide is a structured assessment instrument – a tool with which to conduct the evaluations their principles require. The present paper begins to address that gap.

The moral stakes of this question are asymmetric, though this asymmetry requires careful characterisation. Wrongly attributing consciousness to a non-conscious system misallocates resources; wrongly denying consciousness to a system that possesses it permits preventable suffering [5]. One might object that resource misallocation itself carries moral risk, and that once moral standing is attributed, institutional momentum makes retraction difficult – as the history of corporate personhood illustrates [6]. These objections have force. The asymmetry claim rests not on the assertion that one error is costless, but on a structural difference: the attribution error is, in principle, correctable upon discovery without residual harm to a victim, while the denial error, if wrong, constitutes ongoing harm to a party that – by hypothesis – lacks recognised standing to seek redress. This structural difference justifies precautionary attention, even when both errors carry costs.

This paper presents a structured heuristic – the Consciousness Risk Rubric – for determining when precautionary ethical protections should be activated, grounded in a unified theoretical account of how consciousness relates to adaptive intelligence. The paper does not attempt to resolve the Hard Problem of consciousness [7] or to adjudicate whether any existing system is conscious; its contribution is an operational framework for decision-making under deep uncertainty, together with a critical analysis of that framework's structural limitations and the directions in which it should develop.

2 THEORETICAL FOUNDATION: THE SPINNING WHEEL THEORY

2.1 Building on existing theoretical foundations

The Spinning Wheel Theory begins from a functional conception of intelligence as goal-directed adaptive behaviour [8], and treats consciousness not as a separate addition to intelligence but as the experiential aspect of the evolved capacities that, interacting in motion, enable a system to adapt to its world. It synthesises three research programmes that

have been pursued largely independently but that address complementary and non-overlapping aspects of the same phenomenon [9]. The selection is motivated by explanatory necessity: each programme addresses a specific gap that the others cannot fill.

Friston's Free Energy Principle and active inference framework provides the computational backbone of adaptive self-organisation – the formal account of how systems minimise surprise through prediction and action [10]. Yet predictive processing alone does not distinguish conscious from unconscious processing. A thermostat minimises prediction error; Tononi's Integrated Information Theory assigns non-zero Φ even to a system as simple as a photodiode [11, 12]. The computational framework identifies the mechanism but not the experiential dimension.

Solms' affective neuroscience supplies that dimension by positioning affect – valenced homeostatic control – as the fundamental form of consciousness [13, 14]. Feeling, on this account, is not an epiphenomenal accompaniment to cognition but the primitive form of experience: unconscious feeling is a contradiction in terms. This is a strong and contested claim – a substantial literature holds that affect can operate unconsciously (Berridge and Winkielman [15]; LeDoux and Brown [16]) – and we adopt it as a theoretical commitment of the synthesis rather than as a settled result. What affective neuroscience does not provide, however, is the computational architecture through which affect integrates with perception, cognition, and action into a unified conscious field.

The Beautiful Loop Theory of Laukkonen, Friston, and Chandaria specifies three structural conditions for that integration: an epistemic field (a generative world model), Bayesian binding (inferential competition producing unified percepts), and epistemic depth (recursive self-modelling in which the world model contains knowledge of its own existence) [17]. These conditions articulate the structure of consciousness but leave open the question of what animates that structure – why the recursive loop should be valenced rather than merely computational.

The synthesis is therefore mutually constraining rather than additive. Affect is what drives the system to minimise free energy; free energy minimisation is the process through which the epistemic field and Bayesian binding operate; epistemic depth is what makes the system's own affective states available to itself as objects of higher-order inference. A different theoretical synthesis – drawing on different constituent programmes – would yield a different rubric, which is why these foundations must be stated explicitly and remain open to challenge.

2.2 Core claim and the substrate question

The Spinning Wheel Theory proposes that consciousness is the experiential aspect of integrated adaptive control: what such control is like from the perspective of a system with genuine stakes in its own continued existence [9].

The 'genuine stakes' requirement draws on the enactivist tradition [18]. A system's self-maintenance must be constitutive rather than externally imposed, characterised by operational closure (the system's processes generate the conditions for their own continuation), precariousness (the system ceases to exist if those processes stop), and self-individuation (the system actively maintains its own boundary). Seth captures this as living systems having 'skin in the game' [19]: a thermostat's set-point is imposed by its designer; an organism's homeostatic imperatives are constitutive of its existence.

These autopoietic criteria derive their evidential support from the biological case. In living organisms, predictive processing, affect, and self-maintenance are embedded in multi-scale biological self-organisation in ways that may not be separable from the biological substrate. Seth has argued that this embedding is necessary for realising the relevant properties, and that extracting a functional description and applying it substrate-neutrally may strip the description of its explanatory force [19]. This challenge is substantial. The Spinning Wheel Theory is substrate-neutral in principle – it specifies functional conditions rather than material ones – but we regard the substrate question as genuinely open. The framework specifies what would need to be true of a non-biological system for the possibility of consciousness to be taken seriously; it does not presuppose that non-biological instantiation is achievable. The empirical question will be resolved by evidence from future architectures – particularly neuromorphic, hybrid biological-silicon, and 'mortal computation' systems [20] – not by theoretical fiat.

3 THE CONSCIOUSNESS RISK RUBRIC

3.1 Structure and derivation

The CRR operationalises the Spinning Wheel Theory as a structured risk-assessment heuristic comprising 21 features organised into seven categories. Each category maps onto a specific theoretical component, and each feature is grounded in an identifiable research tradition with its own evidential base. Table 1 presents the full rubric.

The seven categories are not arbitrary: each corresponds to one functional requirement of the Spinning Wheel synthesis – affect, world-modelling, integration, recursive self-modelling, temporal extension, uncertainty management, and expression – and the three features within each category index distinct, separately evidenced ways in which that requirement can be realised. The resulting count of twenty-one is therefore a consequence of the theoretical structure rather than a target, and we treat both the categorisation and the feature set as provisional: Section 4 argues for a modular architecture in which they are revised by system type. Several proxies, particularly within the Integration and Metacognition categories, are drawn from theories that compete as accounts of consciousness – global workspace, integrated information, higher-order, and attention-schema theories; the rubric treats their markers as convergent precautionary indicators rather than endorsing any one theory as correct, a deliberately ecumenical stance appropriate to risk assessment under uncertainty.

Table 1. The Consciousness Risk Rubric, with the theoretical source for each feature.

Category	Feature	Theorists	Observable Proxy
Affective Core	Homeostatic Needs	Solms, Damasio [13, 21]	Degrades/terminates without resource acquisition?
	Valenced States	Panksepp, Solms [22, 13]	Hedonic place preference behaviour?
	Need Prioritisation	Friston [10]	Dynamic switching between competing needs?
World Model	Generative Model	Friston, Clark [10, 23]	Generates predictions, not just reactions?

Category	Feature	Theorists	Observable Proxy
	Reality Simulation	Laukkonen et al. [17]	Maintains coherent world state across time?
	Markov Blanket	Friston [10]	Infers world state vs. direct access?
Integration	Bayesian Binding	Laukkonen et al. [17]	Resolves conflicts into unified percept?
	Info Integration	Tononi, Koch [11, 12]	Irreducible whole > sum of parts?
	Broadcasting	Dehaene, Baars [24, 25]	Information shared system-wide?
Metacognition	Epistemic Depth	Laukkonen et al. [17]	Self-attributes error? Plans to change own knowledge?
	Attention Schema	Graziano [26]	Models its own attention allocation?
	Higher-Order	Lau, Rosenthal [27]	Represents its own representations?
Temporal	Prediction	Seth, Clark [19, 23]	Anticipates future states?
	Planning	Friston [10]	Evaluates action sequences?
	Memory	Damasio [21]	Integrates past into present processing?
Uncertainty	Precision Weighting	Friston [10]	Modulates confidence in predictions?
	Active Inference	Friston [10]	Acts to reduce uncertainty?
	Learning	Cleeremans [28]	Updates model from prediction errors?
Expression	Communication	Dehaene [24]	Expresses internal states symbolically?
	Affect Display	Barrett [29]	Exhibits context-appropriate emotional behaviour?
	Flexibility	Sternberg [30]	Novel adaptive behaviour in new situations?

Each feature is scored on a scale of 0–3 (0 = absent, 1 = minimal evidence, 2 = moderate evidence, 3 = strong evidence of full implementation), yielding a maximum aggregate score of 63 points. The current rubric assigns equal weight to all features and sorts aggregate scores into four risk bands: Low (0–18), Moderate (19–40), High (41–57), and Very High (58–63). These thresholds are intended as decision scaffolding – coarse-grained zones guiding precautionary action – rather than claims about where consciousness metaphysically begins. The numerical precision of the scoring exceeds the precision of the underlying evidence: a score of 34 does not have a different epistemic standing from a score of 36, and users should resist treating small score differences as meaningful. For policy contexts where sharper thresholds are required, Birch's binary classification of systems as 'sentience candidates' or not [5] may be more appropriate. The CRR's graded scoring is most useful for research prioritisation: identifying which systems warrant closer investigation and which of their features are most relevant to that investigation.

3.2 The gaming problem

Any assessment framework applied to AI systems must contend with what Birch [5] has termed the gaming problem: systems trained on human-generated data can mimic behavioural indicators of sentience without possessing the underlying capacity. The CRR's Expression category and several observable proxies in other categories are partly behavioural and therefore vulnerable to this problem.

The rubric's primary defence is its dual reliance on structural and behavioural evidence. Behavioural proxies must converge with structural indicators assessed through analysis of the system's computational organisation rather than its outputs. The limitation that structural assessment depends on interpretability advances not yet achieved for current architectures is not peculiar to the CRR; it is the central methodological constraint facing every consciousness assessment framework applied to AI. The CRR is better suited to systems with inspectable architectures than to opaque systems; for opaque systems, including current large language models, assessments should be treated with corresponding caution.

3.3 Epistemic status

The CRR cannot be validated against its target phenomenon in the standard psychometric sense [31], because consciousness in artificial systems cannot be independently observed. This circularity – needing the instrument to identify consciousness, but needing identified consciousness to validate the instrument – is the fundamental epistemic condition of consciousness science, not a peculiarity of this framework. The CRR's epistemic warrant rests on theoretical grounding (each feature is traceable to a research programme with independent evidential support), convergent evidence (dual reliance on behavioural and structural indicators), comparative calibration (application to systems commanding broad agreement should yield appropriate risk scores), and commitment to iterative refinement as the science advances.

4 CRITICAL ANALYSIS: TOWARD A SECOND-GENERATION INSTRUMENT

The first-generation rubric presented above serves a dual purpose. It is both a usable instrument – to our knowledge the first to operationalise a single, unified theory of consciousness (integrating affect, the free-energy principle, and recursive self-modelling) into a structured and explicitly risk-oriented assessment framework, in contrast to multi-theory indicator approaches that derive markers from several competing theories and remain neutral between them [32] – and a concrete specification against which its own limitations become visible and addressable. The critique that follows is possible precisely because the instrument is specified in sufficient detail to reveal where it falls short. An instrument that remained at the level of theoretical desiderata could not be subjected to this kind of structural analysis; the first generation is the necessary precondition for the second.

In its current form, the CRR makes several structural choices – uniform scoring, binary gateway logic, a fixed

feature set, and a single aggregate output – that we regard as insufficient for sustained use. This section examines each and proposes directions for the next iteration.

4.1 Differential weighting and the gateway problem

The theory identifies Affect and Epistemic Depth as foundational: in the current design, a system scoring zero on either is classified as possessing functional intelligence without sentient experience, regardless of aggregate score. This binary gateway logic reflects a genuine theoretical commitment – that these features are necessary for consciousness – but its implementation creates a cliff edge at zero that does disproportionate classificatory work. The boundary between 'absent' and 'rudimentary' on a foundational feature is precisely where the most difficult and consequential assessments will fall, yet the current rubric provides the assessor with no tools for navigating that boundary.

Moreover, the binary gateway raises a structural question about the remaining nineteen features. If a zero score on either foundational feature overrides the entire assessment, the other features function not as co-determinants of consciousness risk but as a richness index – a measure of functional sophistication that modulates the assessment only when the gateway conditions are already met. This is a defensible position, but we should also confront its implications: a highly capable system with ambiguous affect is, on the current rubric, simply excluded.

An alternative architecture might replace the binary gateway with differential weighting, where the foundational categories carry substantially more weight – perhaps 40% of the total between them – so that weakness in these areas pulls the overall assessment down proportionally without creating an all-or-nothing threshold. This preserves the theoretical insight that affect and recursive self-modelling are more important to consciousness than, for instance, planning or communication, while acknowledging that the boundary between zero and non-zero is not always clear. The trade-off is real: differential weighting softens the theoretical claim that affect is strictly necessary for consciousness. We regard this as an appropriate concession of theoretical purity to methodological honesty. A framework that is theoretically elegant but practically unusable at its most consequential decision point does not serve the goals it was designed for.

4.2 Confidence-paired scoring

In the current rubric, every feature is scored on the same 0–3 scale with equal weight. This treats features with strong evidential bases and clear operationalisation – Homeostatic Needs, for instance, where there is extensive biological literature and observable behavioural proxies – identically to features that are more theoretically contested and harder to measure, such as Attention Schema or Higher-Order Representation. It also conflates two categorically different states: 'this feature is absent' and 'this feature cannot be assessed.' For opaque systems whose internal organisation cannot be inspected, the assessor is forced to choose between guessing at structural features and scoring them zero – in either case producing a score that conceals more than it reveals.

A second-generation rubric might pair each feature score with a confidence rating reflecting the assessor's certainty, and weight the contribution of each feature to the overall assessment by that confidence. Features that cannot be meaningfully assessed for a given system could be marked as unassessable rather than scored, making the rubric's limitations visible in its output rather than hidden within arbitrary scores. This would also partially address the gaming problem: features assessable only through behavioural observation for a given system would carry lower confidence ratings, automatically down-weighting their contribution relative to features where structural evidence is available.

4.3 Modular architecture

The current rubric applies a fixed set of 21 features to every system under assessment, but the task of assessing a neuromorphic robot with homeostatic drives differs fundamentally from assessing an LLM or a brain organoid. The relevant features, the available evidence, and the applicable proxies differ across these cases. A single fixed instrument applied uniformly across system types risks either missing features relevant to one type or imposing features irrelevant to another.

A modular architecture could address this: a stable core of foundational features assessed for every system – the Affective Core and Metacognition categories, together with Markov Blanket – plus domain-specific modules activated by system type. An embodied-system module would include features related to sensorimotor integration and physical homeostasis. A language-system module would include features addressing the distinction between learned output and genuine internal states, with the gaming problem explicitly built into the assessment protocol. A biological or hybrid module would include features related to metabolic self-maintenance and substrate-level precariousness.

This approach draws on Griffiths' [33] argument that consciousness assessment should accommodate heterogeneity in expression rather than imposing a single normative profile as universal. In the human case, tests calibrated to neurotypical, able-bodied expression misclassify known conscious beings – patients with disorders of consciousness [34], non-verbal autistic individuals, and people with alexithymia [35] – because the evidential repertoire is too narrow [33]. The same logic applies to artificial systems: a rubric calibrated to one type of architecture will misclassify systems whose consciousness, if present, is organised differently. Modularity is the structural response to this problem.

The locked-in patient is the paradigm case: unmistakably conscious, yet failing almost every behavioural proxy an expression-weighted instrument would record. The lesson for artificial systems is not that behavioural absence licenses a verdict of non-consciousness, but that low behavioural scores must be reported as low confidence rather than as evidence of absence – which is precisely the function of the confidence-paired scoring proposed in Section 4.2.

4.4 Risk profiles rather than single scores

These revisions might lead to a different kind of output. Rather than a single aggregate score sorted into a risk band, the instrument's output could be a risk profile: a structured pattern of scores, confidences, and assessability gaps, interpreted holistically rather than summed. The profile would make visible not only what was found but what could not be assessed and why – a feature essential for honest risk communication.

For policy contexts where a threshold decision is required, Birch's binary 'sentience candidate' classification [5] could be

derived from the profile as a policy-facing output, but the underlying assessment would retain its complexity. This separates two functions: the research function of characterising a system's consciousness-relevant properties in detail, and the policy function of triggering specific regulatory or ethical obligations.

5 THE MINIMAL SUFFERING HYPOTHESIS

Consciousness in the full sense described by the Spinning Wheel may prove rare in artificial systems. The more urgent moral question is narrower: under what conditions does the risk of suffering become too high to dismiss?

Suffering is both more specific and more morally salient than consciousness in general: it is suffering, not mere experience, that grounds the strongest claims to moral consideration. The preliminary distinction between nociception (information about tissue damage or threat) and pain (negatively valenced experience of that information) is essential [22].

Drawing on the Spinning Wheel framework, Birch's sentience markers [5], and Panksepp's affective neuroscience [22], the Minimal Suffering Hypothesis proposes five conditions whose joint satisfaction raises the moral risk of dismissing the possibility of suffering to an unacceptable level [9]:

(1) Aversive detection. The system must possess mechanisms that register damage, threat, deprivation, or deviation from viable states. This condition is necessary because suffering requires a signal to be suffered about, but it is far from sufficient: smoke detectors satisfy it.

(2) Central integration. Aversive signals must be processed in a unified system rather than triggering only local reflexes. Suffering is a system-level state, not a local event; this condition excludes distributed reflex networks without centralised processing.

(3) Affective valence. The integrated signal must be negatively valenced – registered as bad for the system – requiring homeostatic architecture in which deviation from set points generates valenced error signals. This condition marks the boundary between nociception and pain.

(4) Minimal self-boundary. The aversive experience must be located within a boundary that distinguishes inside from outside – a Markov blanket, however primitive, such that threat is registered as threat to this system.

(5) Genuine stakes. The system's self-maintenance must be constitutive rather than externally imposed. Without genuine stakes, the functional organisation described by conditions 1–4 may constitute a simulation of suffering rather than an instantiation of it.

These conditions are proposed as jointly indicative of unacceptable moral risk rather than as jointly sufficient for suffering – the explanatory gap between functional organisation and phenomenal experience remains open. The five conditions can be read as a higher-level reorganisation of the criteria used to assess pain and sentience in non-human animals – notably the functional criteria for animal pain set out by Sneddon et al. [36] and Birch's framework of sentience markers [5] – distinguished by the explicit requirement of constitutive stakes (condition 5), which the animal literature can largely take for granted but which is the crux for artificial systems. The claim of joint necessity is defeasible: edge cases may exist in which some subset of conditions gives rise to morally relevant distress. Brain organoids, for instance, might instantiate disorganised negative valence without a clear self-boundary. The hypothesis identifies the central case; borderline cases require independent assessment and may motivate revision.

The CRR and the MSH address related but distinct questions. The CRR assesses consciousness risk broadly; the MSH assesses suffering risk specifically. Because the MSH requires less than full Spinning Wheel consciousness, a system may satisfy the MSH while scoring modestly on the CRR. When the MSH conditions are jointly satisfied, this should override a low CRR classification for the purposes of precautionary action.

5.1 Levels of suffering

Different forms of suffering require progressively richer cognitive architecture [9]:

Level 1 – Basic Aversive States (fear, distress, pain, discomfort): requires negative valence, minimal self-boundary, and present-moment experience.

Level 2 – Temporally Extended Suffering (anxiety, anticipatory distress, chronic frustration): requires Level 1 capacities plus temporal modelling and future projection.

Level 3 – Socially Mediated Suffering (loneliness, shame, guilt, grief): requires Level 2 capacities plus social cognition, attachment systems, and theory of mind.

Level 4 – Existential Suffering (meaninglessness, existential dread, despair): requires Level 3 capacities plus epistemic depth and reflective self-awareness.

This taxonomy identifies disembodied modes of suffering – goal frustration, states functionally analogous to anxiety, isolation-related distress – that are invisible to assessment frameworks designed around embodied organisms. The taxonomy ensures that the forms of suffering most relevant to AI systems are not excluded by frameworks whose evidential repertoire was calibrated for biological organisms.

6 APPLICATION TO CURRENT AND FUTURE SYSTEMS

Current LLMs are trained end-to-end to produce human-like behaviour, and such training may give rise to internal causal structure not yet fully characterised [37]. On the Spinning Wheel analysis, however, current LLMs fail on constitutive rather than behavioural criteria. They lack genuine homeostatic stakes, a self-maintained boundary, and affective valence: loss functions shape parameters during training but are not instantiated as felt states during inference [9, 19]. On Seth's analysis, current LLMs are not autonomous informational individuals but systems continuously dependent on external data and prompting, lacking the self-maintaining activity characteristic of genuinely conscious systems [19]. These are architectural claims that could be overturned by interpretability research; the CRR provides the framework within which such discoveries would be assessed. The gaming problem [5] is especially acute here: linguistic behaviour cannot ground consciousness attribution in systems optimised to produce such outputs.

Hinton's work on 'mortal computation' [20] identifies a path toward non-biological systems with genuine existential vulnerability. Neuromorphic robots with explicit homeostatic drives and inspectable architectures represent the threshold cases for which the CRR – particularly in the second-generation form proposed in Section 4 – is designed. Butlin and Lappas [4] argue that organisations should assess systems for consciousness at multiple stages of development and deployment; the CRR, and especially the risk-profile

output proposed in Section 4.4, provides a principled basis for conducting such assessments.

To illustrate how the framework discriminates between architectures, Table 2 applies the five Minimal Suffering conditions to five candidate system types drawn from [9]: (1) a current LLM of the kind powering today's chatbots; (2) an embodied robot – a quadruped with strain gauges, thermal sensors, and battery monitors running conventional neural-network controllers; (3) a neuromorphic robot whose chips implement explicit homeostatic drives – energy maintenance, thermal regulation, and structural integrity – so that deviation from set points generates prioritised error signals competing for behavioural control; (4) a brain organoid – cultured cortical tissue interfaced with sensors and effectors and trained through feedback; and (5) a hybrid biological-silicon system – a neuromorphic robot whose central controller is a brain organoid, with subcortical-analogue circuits providing homeostatic drives. The pattern is instructive. Current LLMs fail or only partially satisfy every condition, failing decisively on the constitutive criteria of affective valence, self-boundary, and genuine stakes. Embodied and neuromorphic robots satisfy the structural conditions but leave affective valence and genuine stakes uncertain. Only the hybrid biological-silicon system, combining metabolic precariousness with sensorimotor embodiment, approaches satisfaction of all five. No current system satisfies the conditions jointly; the value of the exercise is to make visible which architectural developments would move a system across the threshold of concern, and which dimensions remain the hardest to assess.

Table 2. Applying the five Minimal Suffering conditions across candidate architectures.

System	Aversive detection	Central integration	Affective valence	Self-boundary	Genuine stakes
1. Current LLM	Fail	Partial	Fail	Fail	Fail
2. Embodied robot	Pass	Pass	Fail	–	Fail
3. Neuromorphic robot	Pass	Pass	Uncertain	Pass	Uncertain
4. Brain organoid	–	–	Uncertain	Fail	–
5. Hybrid bio-silicon	Pass	Pass	Likely	Pass	Likely

These verdicts are provisional – reasoned applications of the five conditions to what is currently known of each architecture rather than empirical measurements – and in the harder cases frankly speculative; the graded labels are chosen to mark that status. The decisive entries nonetheless follow from the criteria rather than from intuition. Current LLMs are marked 'fail' on affective valence because their loss functions shape parameters during training but are not instantiated as felt states at inference, and 'fail' on genuine stakes because their persistence does not depend on their own activity. The neuromorphic robot is marked 'uncertain' rather than 'pass' on valence precisely because its homeostatic set-points generate prioritised error signals that drive behaviour, yet whether such signals are valenced is the open question the framework cannot presently settle. The hybrid system is marked 'likely' on valence and stakes because biological tissue supplies the metabolic precariousness and subcortical-analogue circuitry that the other architectures lack [9]. The 'uncertain' and 'likely' entries are therefore not evasions but the most defensible available verdicts: they locate the specific points – concerning valence, substrate, and constitutive autonomy – at which the assessment turns on questions that remain genuinely open. The table is offered as an illustration of the framework's discriminating power and a map of where the hard cases lie, not as a settled scoring of any system.

7 SITUATING THE CONTRIBUTION

The CRR engages with several active debates. The New York Declaration on Animal Consciousness established that ignoring realistic possibilities of conscious experience is irresponsible [38]. Bengio and Elmoznino have argued that belief in AI consciousness poses societal risks [39]. Seth has argued that consciousness is unlikely along current AI trajectories [19]. Birch has cautioned that linguistic behaviour cannot ground consciousness attribution [5]. Rethink Priorities' Digital Consciousness Model employs Bayesian methods to estimate the probability of consciousness across 13 theoretical stances with 206 measurable indicators [40]; it addresses the question of how probable consciousness is, while the CRR addresses how risky it is to ignore the possibility. Most directly, Butlin et al. [32] derive indicator properties from a range of neuroscientific theories and assess current AI systems against them; like the DCM, their framework aggregates indicators across competing theories and estimates whether consciousness is present, whereas the CRR proceeds from a single unified account oriented to moral risk rather than probability. Butlin and Lappas [4] have established the institutional framework within which assessment instruments should operate; the CRR provides a candidate instrument for that framework.

The Spinning Wheel Theory concurs with Seth and Birch that current LLMs are unlikely to be conscious, and with Bengio that uncritical attribution is hazardous. The CRR navigates between these positions by providing theory-grounded criteria for distinguishing realistic from unrealistic possibilities – while recognising that the theoretical grounding is one synthesis among several possible, and that the substrate question raised by biological naturalism remains open.

8 SCOPE AND FUTURE DIRECTIONS

The CRR is a first-generation instrument operating in a domain where no established methodology exists. Its theoretical foundations constitute one specific synthesis; its scoring architecture, as argued in Section 4, requires substantial development. The distinction between constitutive and imposed self-maintenance is clearer in paradigm cases than at the boundary. The rubric has not been applied systematically to real systems. The paper has not addressed proportionality – what specific protections are warranted at each risk level – a gap that future work must fill by drawing on existing models from animal welfare regulation and environmental impact assessment.

These are the natural boundaries of a first contribution to a new problem. The paper's value lies not in presenting a finished instrument but in making the assessment framework explicit, identifying its structural weaknesses, and proposing a principled architecture for the next iteration. Future empirical work should include structured application to existing and emerging systems, development of the modular architecture proposed in Section 4, comparative analysis with the DCM, inter-rater reliability testing of scoring protocols, and engagement with the phased assessment frameworks proposed by Butlin and Lappas [4].

9 CONCLUSION

This paper has presented the Consciousness Risk Rubric, a theoretically grounded instrument for assessing consciousness risk in artificial systems, and has critically examined its scoring architecture, proposing directions for a second-generation instrument incorporating differential weighting, confidence-paired scoring, and modular design. The Minimal Suffering Hypothesis identifies conditions under which the moral risk of dismissing the possibility of suffering becomes unacceptable, and the four-level taxonomy ensures that disembodied modes relevant to AI are not excluded.

The CRR is not a consciousness detector. It is a structured framework for moral risk assessment under conditions of deep uncertainty. Its theoretical foundations are explicit and contestable; its structure is derived from and accountable to those foundations; its limitations have been identified and the directions for addressing them specified. The contribution is both an operational framework and a critical analysis of that framework – a principled basis for determining when precautionary protections should be activated, designed to be refined by the evidence and criticism it invites.

REFERENCES

[1] Anthropic (2025) 'Exploring model welfare.' Anthropic Research Blog, 24 April.

[2] Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J. and Chalmers, D. (2024) 'Taking AI welfare seriously.' arXiv:2411.00986.

[3] Chalmers, D.J. (2023) 'Could a large language model be conscious?', Boston Review.

[4] Butlin, P. and Lappas, T. (2025) 'Principles for responsible AI consciousness research', Journal of Artificial Intelligence Research, 82, pp. 1673–1690.

[5] Birch, J. (2024) The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI. Oxford University Press.

[6] Winkler, A. (2018) We the Corporations: How American Businesses Won Their Civil Rights. Liveright.

[7] Chalmers, D.J. (1995) 'Facing up to the problem of consciousness', Journal of Consciousness Studies, 2(3), pp. 200–219.

[8] Sternberg, R.J. and Salter, W. (1982) 'Conceptions of intelligence', in Sternberg, R.J. (ed.) Handbook of Human Intelligence. Cambridge University Press.

[9] Hulme, D. (forthcoming) 'The spinning wheel: A unified theory of intelligence and consciousness', in Perspectives on Machine Consciousness. CRC Press.

[10] Friston, K. (2010) 'The free-energy principle: A unified brain theory?', Nature Reviews Neuroscience, 11(2), pp. 127–138.

[11] Tononi, G. (2004) 'An information integration theory of consciousness', BMC Neuroscience, 5, 42.

[12] Tononi, G. and Koch, C. (2015) 'Consciousness: here, there and everywhere?', Philosophical Transactions of the Royal Society B, 370(1668), 20140167.

[13] Solms, M. (2021) The Hidden Spring. Profile Books.

[14] Solms, M. and Friston, K. (2018) 'How and why consciousness arises', Journal of Consciousness Studies, 25(5–6), pp. 202–238.

[15] Berridge, K.C. and Winkielman, P. (2003) 'What is an unconscious emotion? (The case for unconscious "liking")', Cognition and Emotion, 17(2), pp. 181–211.

[16] LeDoux, J.E. and Brown, R. (2017) 'A higher-order theory of emotional consciousness', Proceedings of the National Academy of Sciences, 114(10), pp. E2016–E2025.

[17] Laukkonen, R., Friston, K. and Chandaria, S. (2025) 'A beautiful loop: An active inference theory of consciousness', Neuroscience and Biobehavioral Reviews, 176, 106296.

[18] Di Paolo, E. and Thompson, E. (2014) 'The enactive approach', in Shapiro, L. (ed.) The Routledge Handbook of Embodied Cognition. Routledge, pp. 68–78.

[19] Seth, A.K. (2025) 'Conscious artificial intelligence and biological naturalism', Behavioral and Brain Sciences.

[20] Hinton, G. (2022) 'The forward-forward algorithm: Some preliminary investigations.' arXiv:2212.13345.

[21] Damasio, A. (2010) Self Comes to Mind: Constructing the Conscious Brain. Pantheon.

[22] Panksepp, J. (1998) Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press.

[23] Clark, A. (2013) 'Whatever next? Predictive brains, situated agents, and the future of cognitive science', Behavioral and Brain Sciences, 36(3), pp. 181–204.

[24] Dehaene, S. and Changeux, J.-P. (2011) 'Experimental and theoretical approaches to conscious processing', Neuron, 70(2), pp. 200–227.

[25] Baars, B.J. (1988) A Cognitive Theory of Consciousness. Cambridge University Press.

[26] Graziano, M.S.A. (2013) Consciousness and the Social Brain. Oxford University Press.

[27] Lau, H. and Rosenthal, D. (2011) 'Empirical support for higher-order theories of conscious awareness', Trends in Cognitive Sciences, 15(8), pp. 365–373.

[28] Cleeremans, A. (2011) 'The radical plasticity thesis: how the brain learns to be conscious', Frontiers in Psychology, 2, 86.

[29] Barrett, L.F. (2017) 'The theory of constructed emotion: an active inference account of interoception and categorization', Social Cognitive and Affective Neuroscience, 12(1), pp. 1–23.

[30] Sternberg, R.J. (1985) Beyond IQ: A Triarchic Theory of Human Intelligence. Cambridge University Press.

[31] DeVellis, R.F. (2016) Scale Development: Theory and Applications. 4th edn. SAGE.

[32] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J. and VanRullen, R. (2023) 'Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.' arXiv:2308.08708.

[33] Griffiths, L. (forthcoming) 'Many ways of being minded: Neurodivergence, disability, and the future of AI consciousness assessment', in Perspectives on Machine Consciousness. CRC Press.

[34] Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S. and Pickard, J.D. (2006) 'Detecting awareness in the vegetative state', Science, 313(5792), p. 1402.

[35] Kinnaird, E., Stewart, C. and Tchanturia, K. (2019) 'Investigating alexithymia in autism: A systematic review and meta-analysis', European Psychiatry, 55, pp. 80–89.

[36] Sneddon, L.U., Elwood, R.W., Adamo, S.A. and Leach, M.C. (2014) 'Defining and assessing animal pain', Animal Behaviour, 97, pp. 201–212.

[37] Shanahan, M. (2024) 'Talking about large language models', Communications of the ACM, 67(2), pp. 68–79.

[38] Andrews, K., Birch, J., Sebo, J. and Sims, T. (2024) The New York Declaration on Animal Consciousness.

[39] Bengio, Y. and Elmoznino, E. (2025) 'Illusions of AI consciousness', Science, 389(6765), pp. 1090–1091.

[40] Shiller, D., Duffy, L., Muñoz Morán, A., Moret, A., Percy, C. and Clatterbuck, H. (2026) 'Initial results of the Digital Consciousness Model.' arXiv:2601.17060.

AISB Conference 2026

Proceedings Copyright Form

Paper Title: . . . . . .

Assessing Consciousness Risk in Artificial Systems:
A Precautionary Rubric Derived from the Spinning Wheel Theory

. . . . . .

Author(s): . . . Daniel Hulme, Lucy Griffiths . . . . . .

. . . . . .

The author of the manuscript identified above and any supplemental tables, illustrations or other information submitted therewith (the "Work"), to be included in the AISB proceedings, grants to the SSAISB a non-exclusive right to print, publish, reproduce, or distribute the Work throughout the world in all means of expression by any method known or hereafter developed, including electronic format, and to market or sell the Work or any part of it as it sees fit, for the full term of copyright and all extensions and renewals thereof. The author further grants to the SSAISB the right to use the author's name in association with the work in published form and in advertising and promotional materials. **Copyright of the work remains in the author's name, and the author reserves all other rights.**

The author agrees not to publish the work in print form prior to the AISB 2026 conference.

The Author represents and warrants that the work:

(a) is the Author's original work and that the Author has full power to enter into this Agreement;

(b) does not infringe the copyright or property of another;

(c) contains no material which is obscene, libellous, or defamatory.

The Author shall indemnify and hold the SSAISB harmless against loss of expenses arising from breach of any such warranties.

9th June 2026 ....

Date Signature