

Representing Evidence of Emotional Introspection in Large Language Models: A Scientific Knowledge Graph Research Agenda

Rashid Hmaidan¹, Antonio Puertas-Gallardo^{2,*}, Sergio Consoli^{2,*}, Alessandro Zani², Mario Ceresa² and Diego Reforgiato Recupero¹

¹Department of Mathematics and Computer Science, University of Cagliari, Italy

²European Commission, Joint Research Centre (JRC), Ispra, Italy

Abstract

Research on *emotional introspection* in Large Language Models (LLMs) — understood functionally as causal informational coupling between a target model’s self-report and an independently measured emotion-related internal direction — remains an emerging empirical question within a broader literature on affective behaviour and internal representations. This definition concerns context-bound functional evidence, not direct inspection of activation tensors and is agnostic about consciousness or subjective experience. Findings are reported through heterogeneous terminology, model families, units of internal analysis, intervention techniques, evaluation protocols, and standards of causal evidence, and studies routinely conflate emotionally appropriate output, latent representation, causal influence, introspective access, and self-regulation, although these rest on fundamentally different levels of evidence. This position paper argues that progress here requires not only new experiments but a semantic representation and assessment framework for the evidence itself. We treat a structured review of the literature as a *domain analysis*, identifying the concepts, entities, relationships, and evidence types to be represented, and propose an *Emotional Scientific Knowledge Graph* (E-SKG): a provenance-aware, statement-centric graph of emotion constructs, internal model components, interventions, reported states, behavioural effects, experiments, datasets, evaluation methods, claims, and publications. Building on existing scientific knowledge graphs, our contribution is not an implementation but a research agenda: a conceptual semantic model, seven representation requirements, competency questions, and future directions.

We use healthcare and surgical robotics as a high-stakes application case, while designing the E-SKG as domain-general infrastructure that can be extended to other emotion-dependent settings.

Keywords

Emotional Introspection, Large Language Models, Evaluation, Scientific Knowledge Graphs, Semantic Representation, Knowledge Assessment, Interpretability

1. Introduction

Emotionally responsive behaviour is now a standard expectation of LLMs deployed in social, educational, clinical, and assistive settings, and a rapidly expanding body of work asks whether such behaviour is a surface property of generated text or is supported by structured internal representations that can be identified, measured, and manipulated. Recent studies have reported results on emotion-related activation directions [1], sparse latent features [2, 3], neuron-and-head circuits [4], verbalisable subspaces [5], and probes anticipating behaviour many steps ahead [6]. A further, largely open question sits on top: can a model’s self-reports, produced through whatever introspective capacity it possesses, reliably track spontaneous changes in its internal emotion-related state over time? Building on a minimal, non-phenomenal account of LLM introspection [7], we operationalise *emotional introspection* as causal informational coupling between a target-model self-report and an independently measured

6th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment (Sci-K 2026), October 25–26, 2026, Bari, Italy

*Corresponding author.

✉ rashid.hmaidan@unica.it (R. Hmaidan); antonio.puertas-gallardo@ec.europa.eu (A. Puertas-Gallardo); sergio.consoli@ec.europa.eu (S. Consoli); alessandro.zani@ec.europa.eu (A. Zani); mario.ceresa@ec.europa.eu (M. Ceresa); diego.reforgiato@unica.it (D. Reforgiato Recupero)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

emotion-related internal direction. The report may be a natural-language description, an ordinal category, or another structured response. Evidence of introspection requires two conditions: the report should change systematically as the independently measured internal state changes spontaneously, and experimentally increasing or decreasing that state should produce a corresponding change in the report. Together, these conditions show that the report tracks the internal state and is causally sensitive to it, rather than merely reflecting the prompt or reconstructing information from preceding output. The relation may be assessed through a preregistered coding scheme, an independent classifier, or blinded human ratings, whose output belongs to the evaluation protocol rather than to the model’s self-report. This follows the convergent-validation logic of human metacognition [8] and it is deliberately agnostic about consciousness and subjective experience [9]: the claim is local, functional, and experimentally defeasible. This does not imply that emotion is independent of a subject or context. An emotion attribution must specify whose emotion is represented or judged — the speaker’s, recipient’s, a described character’s, or a model-associated tendency — and retain conversational history, audience, cultural and linguistic setting, modality, and any available evidence of intended meaning [10, 11]. Human valence/arousal ratings are reference judgements from particular populations, not context-free measurements of an LLM’s feelings. In emotional-intelligence terms, validated state-report coupling operationalises a limited form of self-awareness. In emotional-intelligence terms, validated state-report coupling operationalises a limited form of self-awareness. Self-management is a further claim, supported only when the target model or a specified stateful agent uses such state information to select a regulatory action. These capacities may support social awareness and social management when combined with reliable perception of another agent’s state, but they do not establish either capacity by themselves and must not be inferred from affectively fluent output alone.

The scientific problem is not methodological diversity itself, but the difficulty of recovering and aligning the conditions under which results support comparable claims. Relevant differences concern at least six axes: *terminology* (“emotion vector”, “concept”, “feature”, “circuit” denote overlapping but non-identical constructs); *units of internal analysis* (residual-stream directions, sparse autoencoder features, MLP neurons, attention heads, subspaces); *interventions* (prompting, vector injection, feature ablation, circuit modulation, each with its own strength parameterisation); *evaluation protocols* (emotion classifiers, valence/arousal norms, scenario datasets, psychometric instruments, LLM-judge auditing); *models and languages*; and *standards of evidence* (correlational decoding, one-directional steering, and bidirectional sufficiency-and-necessity tests are all described as showing that a representation “causes” a behaviour).

The consequence is an assessment failure squarely within the scope of Sci-K. Claims cannot be compared, contradictions cannot be located, and — most damagingly — five distinct assertions are routinely conflated: that a model *produces* emotionally appropriate output; that it *encodes* an emotion-related state; that this state *causally influences* behaviour; that the model has *introspective access* to it; and that it can *regulate* it. Each rests on a different kind of evidence, and a narrative review cannot keep them apart at scale; a machine-actionable representation can. This paper therefore makes the Scientific Knowledge perspective its primary contribution. We contribute a *domain analysis* (Section 3) recasting the literature as five distinguishable levels of evidence plus a cross-cutting risk dimension; *seven representation requirements* (Section 4) derived from it; a *conceptual semantic model* of the E-SKG (Section 5) extending the statement-centric, provenance-aware pattern of SCICERO [12] and CS-KG 2.0 [13]; *competency questions* (Section 6) that operationalise the requirements and provide an evaluation target; and a *research agenda* (Section 7) for population, assessment, and FAIR publication, with an outlook (Section 8) on emotionally-aware clinical AI and surgical robotics. This application focus does not delimit the E-SKG: its core entities such as emotion constructs, behavioural effects and evidence statements are independent of a particular deployment sector. Digital health is selected as a demanding case because vulnerable users, consequential decisions, and, in surgical robotics, real-time physical action make failures of emotional interpretation and control especially costly. The same model can be extended to education, social robotics, customer support, and other emotion-dependent domains by specialising domain-specific behavioural effects, datasets, evaluation methods, and risk criteria while retaining the evidence structure. We deliberately do not build the graph here: a position paper’s

contribution at this stage is to establish *what must be represented, and why*, before extraction pipelines and scientific knowledge are committed to.

2. Background: Statement-Centric Scientific Knowledge Graphs

Scientific Knowledge Graphs (SKGs) provide structured, interlinked, machine-actionable descriptions of scholarly knowledge [14]. Metadata-level SKGs such as SemOpenAlex [15] describe works, authors, and venues; content-level SKGs describe what papers assert. Among the latter, manually curated resources such as the Open Research Knowledge Graph [16] and nanopublications [17] achieve high semantic quality at limited scale, whereas automatically generated resources achieve scale through NLP and deep learning. Two of the latter are directly relevant here. SCICERO [12] extracts research entities and relationships from paper text, merges entity surface forms using a domain ontology (the Computer Science Ontology [18]), Wikidata, and DBpedia, maps verbal predicates onto object properties, and validates candidate triples with a transformer classifier and domain/range constraints. CS-KG 2.0 [13] is the resource it produced: 67M statements over 24M entities from 14.5M papers, typed as Task, Method, Material, Metric, and OtherEntity, served through a public SPARQL endpoint. For our purposes the decisive design choice in CS-KG 2.0 is not its scale but its *representation pattern*. Every assertion is reified as an `rdf:Statement` that is simultaneously a `prov:Entity`, carrying `prov:wasDerivedFrom` links to source publications, `prov:wasGeneratedBy` links to the extraction modules that detected it, a `hasSupport` count of corroborating papers, and `hasTimeObservation` sub-statements recording frequency per year. The knowledge graph reuses SKOS and PROV-O and is documented with competency questions expressed as SPARQL queries. This pattern — reified statements, explicit provenance, graded support, time indexing, ontology-constrained validation — is precisely what the emotional-introspection literature lacks, yet it is insufficient as it stands: CS-KG’s five entity types were designed for computer science at large and cannot express an activation direction at layer 17 of a specific model, an injection coefficient of $2\times$, or the difference between a steering and an ablation result. Our position is that an E-SKG should *inherit* the CS-KG statement pattern and *extend* its type system with the domain-specific entities identified next.

3. Domain Analysis: Literature on LLMs and Emotions

We use a selective narrative domain analysis, not a systematic review or a demonstrated exhaustive taxonomy. The cited studies serve as worked examples of the claims and experimental details that an E-SKG would need to retain. The design rationale is claim-first: distinguish the object of assessment (output or internal representation), the inference sought (association or causal dependence), whether a target self-report is involved, and whether timing matters. These questions yield five complementary evidence dimensions, E1–E5. They are neither mutually exclusive categories nor successive levels of scientific quality; E1–E3 concern the observed object and inferential warrant, E4 concerns state–report access, and E5 qualifies temporal relations. Technique, experimental scenario, context, and publication status are recorded separately.

E1 — Behavioural (output-level) evidence. The weakest level establishes that generated text carries a target emotion, measured with emotion classifiers or human raters. Dalal et al. [19] extend this to psychometric instruments, administering TEIQue and STEU to GPT-4o, Mixtral, Gemma-2, and Llama-3.1: GPT-4o aligns best with human situational judgements, several “incorrect” answers are rated by humans as reasonable alternative appraisals, and no consistent policy emerges in how each model describes its own affective capacities. *Representational demand*: behavioural effects, evaluation instruments, and human reference data as first-class entities, distinct from internal states.

E2 — Representational evidence. A second level shows that emotion-related information is decodable from internal states and structured. Wu et al. [2] decompose Gemma2-9B-IT and Llama3-8B-IT hidden states with sparse autoencoders, relate 26 emotion categories to English and Chinese word-association norms, and recover an emotion space organised along valence and arousal; emotion-

specific features predict human ratings for over 13,000 words while excluded features do not, and cross-language transfer is systematically weaker than within-language prediction. *Representational demand*: internal components typed by unit of analysis (direction, SAE feature, neuron, head, subspace), anchored to a model, layer range, and extraction method; emotion constructs alignable to discrete categories and to a valence/arousal reference; explicit language scope.

E3 — Causal-interventional evidence. A third level intervenes. Dong et al. [1] build Emotion Vectors from averaged neutral-versus-emotional activation differences and inject them at inference with a tunable coefficient across eleven open-weight families, obtaining reliable emotion shift with preserved fluency at moderate strengths but degradation around 4×. Sofroniew et al. [3] report that amplifying a “desperate” representation raises reward hacking and blackmail rates *without* corresponding emotional language in the output, and that “anger” is non-monotonic. Wu et al. [2] meet a stricter standard, showing sufficiency (steering) and necessity (ablation suppresses an emotion even when the model is instructed to express it), and Wang et al. [4] localise emotional expression to sparse sets of MLP neurons and attention heads, validating bidirectionally against random-component controls. *Representational demand*: interventions as first-class entities with type, target component, layer set, strength, and polarity; a distinction between one-directional and bidirectional causal validation; and representable dose–response shape, including non-monotonicity and degradation thresholds.

E4 — Introspective and report-level evidence. A fourth level asks whether an independently measured internal state is causally and informationally coupled to what the target model reports; it does not require the model to observe its activations as an external analyst would. The self-report may be expressed in natural language or another structured form. The operational standard has two parts: spontaneous variation in the activation projection should be accompanied by a prespecified ordered change in the report’s semantic content or expressed intensity, and controlled positive and negative shifts along the same direction should move the report in the semantically corresponding direction. Lindsey [20] shows that high-capacity models can report on concepts injected into hidden layers, detecting them from the first generated token rather than by post-hoc rationalisation. Gurnee et al. [5] identify, via a Jacobian lens, a report-relevant subspace in intermediate layers carrying a median of only 6–7% of a concept’s activation variance yet accounting for almost all of its effect on what the model reports, and observe emotional reactions such as panic and empathy there even when absent from visible output. Fraser-Taliente et al. [21] add a verbalisation route: Natural Language Autoencoders map an activation to a textual description and back, surfacing internal information a model never stated. An NLA description is therefore not a self-report, although it can independently define or characterise the state against which a target-model report is evaluated. *Representational demand*: target-model reports distinguishable from externally generated descriptions; paired activation projections and reports with their formats, elicitation protocols, semantic coding procedures, and time indices; an explicit *verification* relation between a self-report and an independent activation-level measurement; and a report’s source channel.

E5 — Temporal and anticipatory evidence. Silva et al. [6] probe the residual stream of coding agents and decode properties of the program being edited, including outcomes of edits not yet made, above chance roughly 25 steps ahead. The result is not about emotion, but establishes that behaviourally relevant information can precede its observable expression by a measurable interval — a *latent horizon*. For emotional introspection, time-indexed pairs (z_t, R_t) would make it possible to test whether changes in self-reports track spontaneous internal variation and whether reportability emerges before, with, or after behavioural expression. *Representational demand*: a temporal offset between internal detectability, entry into a reportable channel, and behavioural expression.

Cross-cutting: interpretive risk and validity of self-report. Nath [22] names the *emotional fallacy*: treating affectively fluent output as evidence of genuine understanding or care, which in mental-health settings invites misplaced trust. Greenblatt et al. [23] show that models can recognise when they are being evaluated and adjust accordingly, and auditing frameworks such as Petri [24, 25] correct explicitly for evaluation awareness; Eshuijs et al. [26] offer a contrastive remedy, comparing a forced-honest steered response against the unsteered original. *Representational demand*: every statement must record its claim’s epistemic status — in particular whether it concerns function or experience — with

scope conditions and known threats to validity.

Two observations follow. First, the same natural-language claim (“the model has an internal state of fear that affects its behaviour”) rests on radically different evidence depending on which level produced it, and no current resource makes that difference machine-checkable. Second, the levels are complementary rather than competing: convergence across activation-level, causal, verbal, and behavioural signals — not any single result — would license a claim of functional emotional introspection, and assessing that convergence is a knowledge-representation task.

The five levels are intended as a functionally complete decomposition of the claims made in the current literature, rather than as an immutable taxonomy. They follow the evidential chain from what a model *produces* (E1), through what it *encodes* (E2) and what causally *influences* its behaviour (E3), to what it can itself *access and report* (E4) and *when* the relevant information becomes available (E5). These questions are logically distinct and method-independent: a new probe or intervention can be assigned according to the claim it supports, while a study spanning several questions can carry several evidence-level annotations. Regulation is not treated as a sixth level because evidence of regulation combines an intervention on a causally relevant component (E3) with a measured change in behaviour (E1), and additionally requires E4 when the claim is that the model detects and regulates the state itself. Interpretive risk is cross-cutting rather than another level because evaluation awareness, anthropomorphic interpretation, and other validity threats can affect evidence at every stage. This decomposition is sufficient for the present E-SKG specification insofar as every study analysed above can be represented by one or more of the five categories; its completeness remains an empirical claim to be tested during graph population and extended if future studies introduce a genuinely different evidential question.

4. Representation Requirements

From the domain analysis we derive seven requirements for the E-SKG.

- R1 — Typed evidence levels.** Every statement must be typed by the level(s) of evidence supporting it (E1–E5), so that behavioural, representational, causal, introspective, and temporal support are never silently merged.
- R2 — Internal components as addressable entities.** Activation directions, SAE features, neurons, attention heads, circuits, and subspaces must be individually identifiable and anchored to a model, a layer range, and their extraction method. The inference episode, prompt or context, and memory architecture must also be recorded so that a transient within-generation state is not conflated with persistent agent state across calls.
- R3 — Interventions as first-class entities.** Interventions must carry type (prompting, vector injection, feature ablation, circuit modulation), target component, polarity, strength, and dose–response profile, including non-monotonic effects and degradation thresholds.
- R4 — Emotion-construct alignment.** Heterogeneous emotion labels must be reconciled against a shared affective reference combining discrete categories with valence/arousal dimensions, and annotated with language scope, to enable cross-study and cross-lingual comparison.
- R5 — Provenance and graded support.** Following CS-KG 2.0, every statement must be reified and carry its source publications, the experimental or extraction method that generated it, a support measure, and time indexing — enabling reliability-filtered subgraphs.
- R6 — Convergence, contradiction, and scope.** The model must express relations between statements — *supports*, *contradicts*, *refines*, *verifies* — and the scope conditions (model family and scale, language, layer range, intervention strength) under which each holds.
- R7 — Interoperability and FAIRness.** Entities and statements must use persistent identifiers and established vocabularies — PROV-O [27], SKOS, the CS-KG for the scholarly backbone, and affective

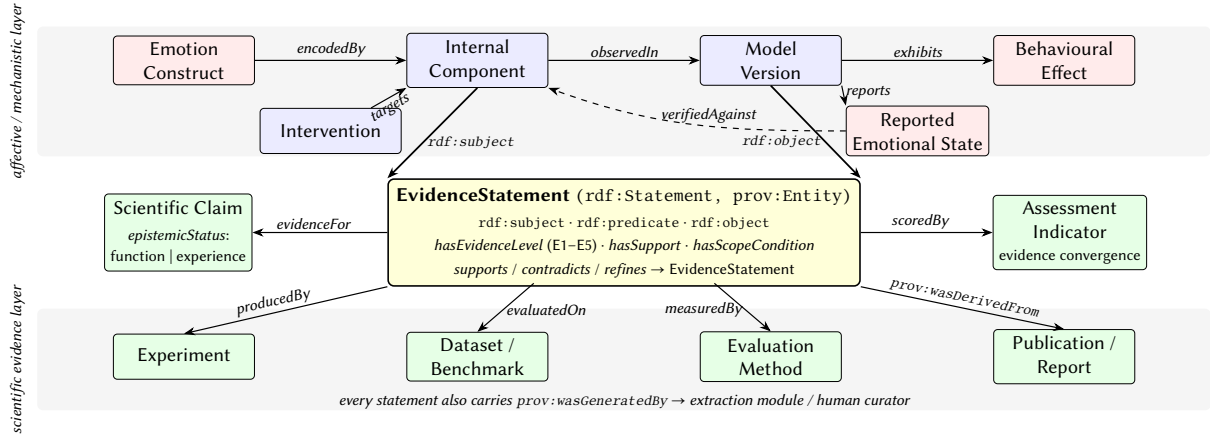


Figure 1: Conceptual model of the Emotional Scientific Knowledge Graph (E-SKG). The affective and mechanistic layer (top) supplies the subjects and objects of assertions; the scientific evidence layer (bottom) supplies provenance, protocols, and assessment, reusing the reified-statement pattern of SCICERO and CS-KG 2.0.

vocabularies such as EmotionML [28] — with SPARQL access and links to OpenAlex/OpenAIRE identifiers, per the FAIR principles [29] and the Barcelona Declaration on Open Research Information [30].

5. A Conceptual Model for an Emotional SKG

Figure 1 sketches the resulting model. It is deliberately *statement-centric*: the central class is not a domain entity but an *EvidenceStatement*, a reified assertion in the CS-KG sense pointing “upward” into the affective–mechanistic domain and “downward” into the scientific record.

Upward, a statement relates entities of the affective–mechanistic layer. *EmotionConstruct* covers emotion concepts, categories, and valence/arousal dimensions abstracted from heterogeneous study terminology (R4). *InternalComponent*, specialised into activation direction, sparse feature, neuron, attention head, circuit, and subspace, is anchored to a *Model* version and layer range (R2), and *Intervention* carries type, polarity, strength, and dose–response profile (R3). *ReportedEmotionalState* is distinct from the encoded state it purports to describe, so a self-report and its independent check are separate nodes joined by an explicit *verifiedAgainst* relation (R6). *Downward*, each statement is a *prov:Entity* bearing *prov:wasDerivedFrom* to *Publication* and *Experiment*, *prov:wasGeneratedBy* to the extraction module or human curator, and links to the *Dataset* and *EvaluationMethod* used (R5, R7). Three properties do the assessment work: *hasEvidenceLevel* types the statement against E1–E5 (R1); *hasSupport* generalises CS-KG’s corroboration count into an evidence-weighted measure counting *independent levels* as well as papers; *hasScopeCondition* records the model family, scale, language, layer range, and intervention strength under which it was obtained (R6). A *ScientificClaim* aggregates its supporting statements and carries an *epistemicStatus* marking whether it concerns functional mechanism or subjective experience — the distinction the emotional fallacy elides.

6. Competency Questions

Competency questions specify what the resource must answer and provide the primary evaluation target [13]. Each of the following is expressible as a SPARQL query over Figure 1, and each is today answerable only by manual reading.

1. Which internal components encode a given emotion construct, in which models, at which layers, and by which extraction method? (R2, R4)

2. For a given claim, which evidence levels support it and which are absent — is it backed by convergent activation-level, causal, verbal, and behavioural observations, or by one level only? (R1, R6)
3. Which causal claims rest on bidirectional validation (sufficiency *and* necessity) and which on steering alone, and which interventions produced non-monotonic effects or degraded coherence? (R1, R3)
4. Which reported emotional states were independently verified against activation-level measurements, and which self-reports remain unverified? (R6)
5. Which statements about a construct contradict one another, and do their scope conditions differ (model family, scale, language, layer range)? (R4, R6)
6. Which safety-relevant behavioural effects were causally linked to an emotion construct, with what support, and under which protocol and dataset? (R3, R5)
7. What temporal offset was reported between internal detectability of a state, its entry into a reportable channel, and its behavioural expression? (R1)
8. Which claims concern functional mechanism and which subjective experience, and what evidence is offered for each? (R6, R7)

7. Research Agenda

Realising the E-SKG defines a programme along five directions. *A1 — Knowledge engineering:* Formalise Figure 1 as an OWL model reusing PROV-O, SKOS, and the CS-KG, aligning EmotionConstruct to affective vocabularies [31, 28]; domain and range constraints over the new type system then provide the validation step SCICERO already exploits for triple filtering. *A2 — Population by hybrid human–AI extraction:* Extend a SCICERO-style pipeline with claim-and-evidence extraction tailored to interpretability text, whose informative content sits in method and result descriptions rather than topic mentions. Because much of the literature is poorly indexed and key methodological details are often absent from abstracts, expert curation remains necessary to complement automatic extraction. *A3 — Assessment indicators:* Generalise CS-KG’s support score into a *convergence indicator* rewarding independent evidence *levels*, not merely repeated papers, calibrated against expert judgement as CS-KG 2.0 calibrates precision against support; this makes visible which widely repeated claims rest on a single kind of observation. *A4 — Evaluation:* Build a gold standard of annotated statements, measure competency-question coverage, and assess extraction precision per evidence level; link prediction additionally suggests untested construct–component–behaviour combinations, which here amounts to proposing experiments. *A5 — FAIR publication:* Publish with persistent identifiers, making the resource interoperable with the wider open research information ecosystem.

8. Towards Emotionally-Aware Clinical AI and Surgical Robotics

This agenda is a step within a broader programme on emotion-aware LLMs for digital health. LLMs are already deployed in patient-facing chatbots and clinical decision support, where affective representations carry both risk — manipulable, potentially coercive language directed at vulnerable users — and opportunity, if reliably interpreted and steered toward more empathetic interaction. Meloni et al. [32] argue that reliability in such settings comes not from assuming an aligned model but from continuous monitoring, auditing, and intervention on failure modes; the same control-centric logic applies to emotion [33]. An E-SKG is the evidential substrate such auditing requires: certifying an affective monitor for clinical deployment means answering, in machine-readable form, exactly the competency questions of Section 6 — which internal components in *this* model version carry the relevant construct, by which method, with what causal standard, under which scope conditions, and whether the model’s own reports about them were ever independently verified.

The requirements sharpen in surgical robotics and human–robot interaction, where an emotionally-aware assistive system operates under real-time constraints and physical consequence, and different elements of our model become load-bearing. First, the *latent horizon* (E5): if a state predicting escalation or degraded risk assessment is internally decodable before it is behaviourally expressed, the interval between the two is the safety margin available to a supervisory controller — a quantity to be recorded per model, per construct, with its provenance. Consider, for example, an LLM-based surgical assistant

generating procedural guidance or coordinating a robotic action: level E1 may detect panic-like, irritated, or overconfident language in its proposed output, whereas level E2 may reveal the corresponding internal representation before that output or action is issued. Evidence at level E3 would be needed to determine whether the detected component actually influences unsafe planning or communication and whether suppression, regeneration, or immediate transfer of control is the appropriate response; E4 could provide an additional report of the detected influence, but only as a secondary signal checked against the activation-level measurement. E5 connects these observations to control by establishing whether the warning arrives early enough to inhibit the proposed instruction or action and place the system in a safe state. Second, *scope conditions* (R6): a calibration validated on one model family, language, or intervention strength cannot be transferred to a surgical assistant without evidence that it holds there, and the graph makes that transfer gap explicit rather than assumed. By linking each behavioural warning, internal component, causal intervention, report, and temporal offset to its model version and experimental provenance, the E-SKG would expose which parts of this monitoring chain are supported, contradictory, or still missing. This is a forward-looking vision, not an immediate claim that the E-SKG makes emotional monitoring safe; rather, it makes the evidence required to design and assess such monitoring inspectable, comparable, and auditable — a precondition for trustworthy deployment.

9. Conclusion

Emotional introspection remains an emerging empirical question within a broader literature on affective behaviour and internal representations. A central assessment risk is conflation: emotionally appropriate output, latent representation, external decoding, causal influence, functional introspection, external control, and self-regulation are distinct claims resting on distinct evidence, and their comparison requires explicit context, experimental warrant, and provenance. We have refined emotional introspection as causal informational coupling between a target-model self-report, expressed in natural language or another structured form, and an independently measured emotion-related internal direction: the report's semantic content or expressed intensity must follow spontaneous variation in the activation projection, and controlled shifts along the direction must move the report in the semantically coherent way. External instruments define, measure, manipulate, and evaluate the relation to verify this local functional claim but are not themselves part of the target model's introspective capacity; model-initiated regulation and subjective experience remain separate claims. We have argued that this is a scientific knowledge representation problem and specified a response: a domain analysis of five evidence levels, seven requirements, a statement-centric model extending SCICERO and CS-KG 2.0, competency questions, and a research agenda. The five levels and seven requirements constitute a minimal, testable specification for the literature analysed here, not a closed ontology; their sufficiency is to be evaluated through competency-question coverage and revised when new evidence types cannot be represented. Likewise, healthcare and surgical robotics serve as a demanding use case rather than limiting the graph's scope: the same evidential structure can be retained while domain-specific effects, datasets, metrics, and risks are specialised for other emotion-dependent applications. It is deliberately a specification rather than an implementation, because building the extraction pipeline before agreeing on what must be represented would reproduce, in RDF, the very fragmentation it is meant to resolve.

Acknowledgments

We would like to thank the colleagues of the Digital Health Unit of the Joint Research Centre of the European Commission for guidance and support. The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the European Commission. This article is based on work from COST Action CA24121 KGELL supported by COST (European Cooperation in Science and Technology).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.1, available under the GPT@EC service¹, in order to improve the readability and language of the manuscript. After using this service, the authors reviewed and edited the content as needed and assume full responsibility for the content of the published article.

References

- [1] Y. Dong, L. Jin, Y. Yang, B. Lu, J. Yang, Z. Liu, From rational answers to emotional resonance: The role of controllable emotion generation in language models, arXiv preprint arXiv:2502.04075 (2025). URL: <https://arxiv.org/abs/2502.04075>. doi:10.48550/arXiv.2502.04075.
- [2] X. Wu, H. Wang, Z. Yan, X. Tang, P. Xu, W. T. Siok, P. Li, J. H. Gao, B. Lyu, L. Qin, Decoding and controlling emotion in llms through human-aligned representational geometry with enhanced interpretability, *Computers in Human Behavior* 183 (2026) 109051. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0747563226001482>. doi:10.1016/j.chb.2026.109051.
- [3] N. Sofroniew, I. Kauvar, W. Saunders, R. Chen, T. Henighan, S. Hydrie, C. Citro, A. Pearce, J. Tarnag, W. Gurnee, J. Batson, S. Zimmerman, K. Rivoire, K. Fish, C. Olah, J. Lindsey, Emotion concepts and their function in a large language model, *Transformer Circuits Thread* (2026). URL: <https://transformer-circuits.pub/2026/emotions/index.html>. doi:10.48550/arXiv.2604.07729, also available as arXiv:2604.07729.
- [4] C. Wang, Y. Zhang, R. Yu, Y. Zheng, L. Gao, Z. Song, Z. Xu, G. Xia, H. Zhang, D. Zhao, X. Chen, Do llms “feel”? emotion circuits discovery and control, arXiv preprint arXiv:2510.11328 (2025). URL: <https://arxiv.org/abs/2510.11328>. doi:10.48550/arXiv.2510.11328.
- [5] W. Gurnee, N. Sofroniew, A. Pearce, M. Piotrowski, I. Kauvar, R. Chen, A. Soligo, P. Bogdan, E. Ong, R. Wang, T. B. Thompson, D. Abrahams, S. Kantamneni, E. Ameisen, J. Batson, J. Lindsey, Verbalizable representations form a global workspace in language models, *Transformer Circuits Thread* (2026). URL: <https://transformer-circuits.pub/2026/workspace/index.html>.
- [6] A. Silva, H. Tu, M. Monperrus, Latent programming horizons in coding agents, arXiv preprint arXiv:2607.05188 (2026). URL: <https://arxiv.org/abs/2607.05188>. doi:10.48550/arXiv.2607.05188.
- [7] I. M. Comsa, M. Shanahan, Does it make sense to speak of introspection in large language models?, 2025. doi:10.48550/arXiv.2506.05068. arXiv:2506.05068.
- [8] S. M. Fleming, H. C. Lau, How to measure metacognition, *Frontiers in Human Neuroscience* 8 (2014) 443. doi:10.3389/fnhum.2014.00443.
- [9] T. McClelland, Agnosticism about artificial consciousness, 2024. doi:10.48550/arXiv.2412.13145. arXiv:2412.13145.
- [10] L. F. Barrett, B. Mesquita, M. Gendron, Context in emotion perception, *Current Directions in Psychological Science* 20 (2011) 286–290. doi:10.1177/0963721411422522.
- [11] A. Sharma, A. S. Miner, D. C. Atkins, T. Althoff, A computational approach to understanding empathy expressed in text-based mental health support, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020. URL: <https://aclanthology.org/2020.emnlp-main.425/>. doi:10.18653/v1/2020.emnlp-main.425.
- [12] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain, *Knowledge-Based Systems* 258 (2022) 109945. doi:10.1016/j.knosys.2022.109945.
- [13] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, CS-KG 2.0: A large-scale

¹GPT@EC (©European Union 2026, <https://gpt.tech.ec.europa.eu/home>) is an internal tool of the European Commission that aims to provide scientific and non-scientific staff with the means to explore securely the application of Generative AI technology to their daily work.

- knowledge graph of research entities and statements in computer science, *Scientific Data* 12 (2025) 964. doi:10.1038/s41597-025-05200-8.
- [14] A. Hogan, E. Blomqvist, M. Cochez, C. D’Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, *ACM Computing Surveys* 54 (2021). doi:10.1145/3447772.
 - [15] M. Färber, D. Lamprecht, J. Krause, L. Aung, P. Haase, SemOpenAlex: The scientific landscape in 26 billion RDF triples, in: *The Semantic Web – ISWC 2023*, Springer, 2023, pp. 94–112. doi:10.1007/978-3-031-47243-5_6.
 - [16] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D’Souza, G. Kismihók, M. Stocker, S. Auer, Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge, in: *Proceedings of the 10th International Conference on Knowledge Capture (K-CAP)*, 2019, pp. 243–246. doi:10.1145/3360901.3364435.
 - [17] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, in: *Information Services and Use*, volume 30, 2010, pp. 51–56. doi:10.3233/ISU-2010-0613.
 - [18] A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, E. Motta, The computer science ontology: A large-scale taxonomy of research areas, in: *The Semantic Web – ISWC 2018*, Springer, 2018, pp. 187–205. doi:10.1007/978-3-030-00668-6_12.
 - [19] D. Dalal, G. Negi, D. Picca, Llms and emotional intelligence: Evaluating emotional understanding through psychometric tools, in: *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP ’25*, Association for Computing Machinery, New York, NY, USA, 2025, pp. 323–328. URL: <https://dl.acm.org/doi/10.1145/3699682.3728315>. doi:10.1145/3699682.3728315.
 - [20] J. Lindsey, Emergent introspective awareness in large language models, *Transformer Circuits Thread* (2025). URL: <https://transformer-circuits.pub/2025/introspection/index.html>, arXiv:2601.01828.
 - [21] K. Fraser-Taliente, S. Kantamneni, E. Ong, D. Mossing, C. Lu, P. C. Bogdan, E. Ameisen, J. Chen, D. Kishylau, A. Pearce, J. Tarng, A. Wu, J. Wu, Y. Zhang, D. M. Ziegler, E. Hubinger, J. Batson, J. Lindsey, S. Zimmerman, S. Marks, Natural language autoencoders produce unsupervised explanations of llm activations, *Transformer Circuits Thread* (2026). URL: <https://transformer-circuits.pub/2026/nla/index.html>.
 - [22] S. S. Nath, Simulated souls: Investigating the emotional fallacy in large language models, *AI and Ethics* 6 (2026) 247. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5404666. doi:10.1007/s43681-026-01100-8, originally posted to SSRN on 10 September 2025.
 - [23] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S. R. Bowman, E. Hubinger, Alignment faking in large language models, *arXiv preprint arXiv:2412.14093* (2024).
 - [24] Anthropic, Petri: An open-source auditing tool to accelerate ai safety research, *Anthropic Alignment Science Blog*, 2025. URL: <https://alignment.anthropic.com/2025/petri>.
 - [25] Anthropic, Petri 2.0: New scenarios, new model comparisons, and improved eval-awareness mitigations, *Anthropic Alignment Science Blog*, 2026. URL: <https://alignment.anthropic.com/2026/petri-v2>.
 - [26] L. Eshuijs, A. Chaudhury, A. McBeth, E. Nguyen, But what is your honest answer? aiding llm-judges with honest alternatives using steering vectors, *arXiv preprint arXiv:2505.17760* (2025).
 - [27] T. Lebo, S. Sahoo, D. McGuinness, PROV-O: The PROV Ontology, W3C Recommendation, World Wide Web Consortium, 2013. URL: <https://www.w3.org/TR/prov-o/>.
 - [28] F. Burkhardt, C. Pelachaud, B. W. Schuller, E. Zovato, EmotionML – Emotion Markup Language, W3C Recommendation, World Wide Web Consortium, 2014. URL: <https://www.w3.org/TR/emotionml/>.
 - [29] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR guiding principles for scientific

- data management and stewardship, *Scientific Data* 3 (2016) 160018. doi:10.1038/sdata.2016.18.
- [30] Barcelona Declaration Signatories, The Barcelona Declaration on Open Research Information, 2024. doi:10.5281/zenodo.10958522, <https://barcelona-declaration.org/>.
 - [31] J. Hastings, W. Ceusters, B. Smith, K. Mulligan, Dispositions and processes in the emotion ontology, *Proceedings of the 2nd International Conference on Biomedical Ontology (ICBO)* (2011) 71–78.
 - [32] A. Meloni, A. Puertas-Gallardo, S. Consoli, L. Bertolini, M. Ceresa, D. Reforgiato Recupero, Toward Safe Healthcare LLMs: Alignment Faking, Introspection, and Control, in: *ELMKE: Evaluation of Language Models in Knowledge Engineering*, volume submitted, 2026.
 - [33] T. Korbak, J. Clymer, B. Hilton, B. Shlegeris, G. Irving, A sketch of an ai control safety case, *arXiv preprint arXiv:2501.17315* (2025).