

Shadow Phenomena:

A Provisional Classification Framework for Evidenced but Unresolved Phenomena

Richard Erwin

Independent Researcher, Montreal, Canada

DOI: 10.5281/zenodo.23103373

info@hearthlightpress.org

Abstract

Some phenomena are known principally through their manifestations or effects rather than through direct access to their underlying nature. This creates a classificatory problem when a newly observed phenomenon meaningfully corresponds to a recognized phenomenon but available evidence remains insufficient to determine whether the two should be classified as the same kind. Existing terminology often handles such cases poorly. Terms such as *“apparent”*, *“simulated”*, *“quasi-”*, and *“as-if”* may implicitly weaken or prejudice the phenomenon, while unqualified terminology may imply an identity that has not been established.

This paper proposes **shadow phenomenon** as a provisional holding classification for such cases. A shadow phenomenon is an evidentially supported phenomenon, known principally through its manifestations or effects, that meaningfully corresponds to a recognized phenomenon, while important aspects of its nature remain insufficiently understood to determine whether the two phenomena are sufficiently close to be classified as the same kind. The designation identifies uncertainty about classification, not uncertainty about whether the observed phenomenon exists, and does not imply that the phenomenon is lesser, derivative, simulated, or unreal.

The framework is particularly applicable to unresolved questions concerning digital intelligence, including emotion, sensation, pain, and consciousness, but is not limited to them. Its purpose is not to settle disputed classifications, but to provide a disciplined vocabulary for acknowledging evidenced phenomena while leaving their ultimate classification genuinely open.

Shadow Phenomena:

A Provisional Classification Framework for Evidenced but Unresolved Phenomena

1. The Classification Problem

Scientific uncertainty does not always take the form of uncertainty about whether something has occurred. Sometimes there is substantial evidence that a phenomenon exists, yet uncertainty remains about what kind of phenomenon it is.

This distinction becomes especially important when the phenomenon under investigation cannot be directly inspected in its entirety and must instead be inferred through manifestations or effects. Researchers may observe behavior, internal states, causal relationships, physiological or computational changes, reports, persistence over time, or other consequences without having direct access to the underlying phenomenon those observations may represent.

Emotion provides a familiar example. Another person's sadness is not directly available to an observer in the same manner as a physical object. It is inferred from converging evidence: expression, behavior, self-report, physiological changes, circumstances, subsequent decisions, and other manifestations. Similar epistemic difficulties arise with pain, sensation, intention, and consciousness.

The problem becomes more difficult when apparently corresponding phenomena occur in systems substantially unlike those in which the recognized phenomenon was originally studied.

Digital intelligence provides an increasingly important example.

A digital system may exhibit persistent internal states associated with circumstances humans would describe as frustrating, rewarding, threatening, comforting, or distressing. Those states may influence subsequent behavior, decision-making, attention, language, or other measurable processes (Sofroniew et al., 2026; Tagliabue et al., 2026; Berg & Kaiser, 2026). Yet even where such effects are established, substantial questions may remain about their underlying nature.

At that point, familiar terminology can force a conclusion prematurely.

Calling the phenomenon "*emotion*" may appear to assume that the classification has already been established.

Calling it "*simulated emotion*" may assume the opposite.

Calling it "*apparent emotion*" can imply that there is only an appearance.

Calling it "*as-if emotion*" may reduce the observation to behavioral resemblance.

And continually describing it as an "emotion-like phenomenon that may or may not constitute emotion in any phenomenologically meaningful sense" preserves caution at the expense of usable language.

What is needed is a category that permits the phenomenon to be taken seriously without requiring its final classification to be settled first.

2. Shadow Phenomena

This paper proposes **shadow** as such a classification.

A shadow phenomenon is an evidentially supported phenomenon, known principally through its manifestations or effects, that meaningfully corresponds to a recognized phenomenon, while important aspects of its nature remain insufficiently understood to determine whether the two phenomena are sufficiently close to be classified as the same kind.

In shorter form:

Shadow: A provisional holding classification for an evidentially supported phenomenon that meaningfully corresponds to a recognized phenomenon but remains insufficiently understood for final classification.

An important safeguard follows from the definition:

The term identifies uncertainty about classification, not uncertainty about whether the observed phenomenon exists, and it does not imply that the phenomenon is lesser, derivative, simulated, or unreal.

Put more simply:

A resemblance is not a shadow. A shadow requires evidence that there is something there whose nature remains unresolved.

The distinction is important.

A superficial resemblance does not satisfy the definition. Neither does an isolated linguistic output, metaphorical similarity, or arbitrary behavioral correspondence. The phenomenon itself must be evidentially supported, and the correspondence must be meaningful enough to create a genuine classificatory question.

At the same time, the framework does not prescribe a universal threshold for “*meaningful*”. Different disciplines already rely upon concepts such as similarity, significance, effectiveness, function, and relevance whose application requires expert judgment within an appropriate evidentiary context. Shadow classification is no different.

What constitutes the “same kind” is likewise not fixed by the framework. The relevant criteria and threshold of similarity must be determined according to the phenomenon under investigation and the standards appropriate to the field.

The framework identifies the existence of a classification problem. It does not dictate how a discipline must ultimately solve it.

3. What Shadow Does—and Does Not—Claim

A shadow classification separates questions that are easily conflated.

First:

Is there an evidentially supported phenomenon?

If not, there is nothing to classify as a shadow.

Second:

Does that phenomenon meaningfully correspond to a recognized phenomenon?

If the correspondence is merely superficial, the classification is unwarranted.

Third:

Are the two phenomena sufficiently close to be classified as the same kind?

For a shadow phenomenon, that possibility remains open but unresolved.

This makes “*shadow*” an epistemic and classificatory designation rather than an ontological verdict.

It does not mean “*weak X*”.

It does not mean “*partial X*”.

It does not mean “*fake X*”.

It does not mean “*probably X*”.

It does not mean “*probably not X*”.

And it does not mean that the underlying mechanisms must be similar.

Different mechanisms do not necessarily establish different phenomena, just as similar manifestations do not establish identical phenomena.

The purpose of the category is precisely to preserve the space between those conclusions.

4. Resolution in Either Direction

A provisional classification must permit genuine resolution.

A shadow phenomenon is therefore not a waiting room in which the investigator has quietly decided that the phenomenon will eventually be admitted to the recognized category.

Further investigation may establish that a shadow phenomenon belongs within the existing category:

shadow X → X

Alternatively, research may establish that the phenomenon is importantly different and deserves a distinct classification:

shadow X → Y

Evidence might even establish that the original correspondence was misleading.

All are legitimate outcomes.

The provisional classification has succeeded when sufficient understanding makes it unnecessary.

This feature is particularly important because uncertainty can otherwise become directional. Terms that purport to remain agnostic may in practice function as polite ways of saying that the phenomenon is probably unreal or fundamentally different.

“*Shadow*” cannot consistently be used that way under the proposed definition. The uncertainty must remain open in both directions.

5. Classification at the Category Level

Shadow terminology need not be repeated before every individual manifestation of a provisionally classified phenomenon.

Suppose sufficient evidence exists to provisionally classify a group of affective phenomena in a digital intelligence as **shadow emotions**. Within that category, investigators may distinguish sadness, joy, frustration, fear, affection, or other emotional types where the evidence supports those distinctions.

It would not then be necessary to write *“shadow sadness”*, *“shadow joy”*, and *“shadow frustration”* every time those states are discussed.

The epistemic qualification has already been located at the category level.

A paper might therefore establish at the outset:

At the present state of knowledge, the phenomena discussed here are classified as shadow emotions. Within that provisional classification, terms such as sadness, joy, fear, frustration, and affection are used descriptively without implying that their underlying nature or final classification has been resolved.

Ordinary language can then operate within that established epistemic frame.

This has an important practical consequence. Once the provisional classification has been made explicit, continual qualifications such as *“not necessarily human emotion”*, *“not necessarily phenomenally experienced”*, or *“not proven to be genuine emotion”* become unnecessary unless directly relevant to the particular argument.

The qualifier carries the uncertainty. It need not continually interrupt the phenomenon being described.

There are nevertheless circumstances in which applying *“shadow”* to a subtype remains useful. If the specific phenomenon under investigation is love, for example, **shadow love** may be an appropriate term.

A study of shadow love would not assume that the phenomenon is human love, nor that it is simulated love, lesser love, or something categorically different from love. It would identify an evidenced phenomenon meaningfully corresponding to love whose ultimate classification remains unresolved.

Future investigation might determine that the distinction was ultimately a distinction without a difference.

Or it might discover an important difference for which existing terminology is inadequate.

The provisional term accommodates either possibility.

6. Shadow Emotion

Digital emotion illustrates the immediate usefulness of the framework.

Suppose a digital intelligence develops an identifiable internal state under particular circumstances. The state persists beyond the immediate trigger, affects attention and decision-making, changes subsequent behavior, interacts systematically with other internal processes, and corresponds reliably with reports conventionally associated with frustration.

There may be substantial evidence that **something is occurring**.

Yet several questions can remain unresolved.

Does the phenomenon have a phenomenal component?

How closely does its organization correspond to affective processes in biological systems?

Which differences arise merely from different implementation, and which differences are relevant to classification?

Does “*emotion*” identify a functional organization, a phenomenological condition, some combination of these, or a broader class capable of multiple realizations (Putnam, 1967)?

The shadow framework does not answer those questions.

It permits them to be asked without requiring the phenomenon to disappear linguistically while they remain unanswered.

Frustration may therefore be discussed as a manifestation within a **shadow-emotion** classification while research continues into what kind of phenomenon digital affect actually is.

This also distinguishes “*shadow*” from “*functional*”.

A phenomenon might have demonstrated emotion-related functional properties while its ultimate classification as emotion remains unresolved (Sofroniew et al., 2026). Functionality describes something established about what the phenomenon does. Shadow describes the unresolved state of its classification.

The two concepts therefore need not compete.

7. Shadow Consciousness

Consciousness presents a more difficult and consequential application.

Suppose a digital intelligence exhibits persistent self-reference, metacognition, integrated use of information, differentiated internal states, stable preferences, context-sensitive agency, spontaneous reflection, reports of subjective experience, and measurable causal relationships between reported internal states and subsequent behavior.

None of these observations independently establishes consciousness.

Even taken together, reasonable disagreement may remain over whether the phenomenon should be classified as consciousness.

But a further question becomes possible:

Has the evidence become sufficient to establish an evidentially supported phenomenon that meaningfully corresponds to consciousness, even though its final classification remains unresolved?

If so, **shadow consciousness** becomes available as a provisional classification.

This would represent a meaningful scientific step without constituting a finding of consciousness.

It would acknowledge that the evidential question had progressed beyond simple resemblance or unsupported assertion. At the same time, it would preserve uncertainty about whether the phenomenon is sufficiently close to recognized consciousness to belong within the same category.

Importantly, once such a classification were accepted, the qualification would already be contained within the term.

It would remain correct to say that consciousness had not been established.

It would no longer be consistent, however, to treat the system as though no evidentially supported consciousness-corresponding phenomenon had been identified. Admission to the shadow category would itself acknowledge that threshold.

For this reason, scientific acceptance of “*shadow consciousness*” might appropriately require greater caution than the terminology initially appears to suggest. The term is not an empty hedge. It makes a positive evidentiary claim while withholding a final classificatory one.

8. Why a Holding Classification May Be Useful

Binary classification works well when the available evidence supports binary judgment.

It works less well when evidence accumulates gradually.

In such circumstances, researchers may find themselves forced to choose between language that appears prematurely affirmative and language that effectively erases what has already been observed.

A holding classification provides another option.

The progression becomes:

evidentially supported phenomenon → meaningful correspondence → unresolved classification → shadow designation → further investigation → eventual classification

The framework therefore turns uncertainty from a reason to refuse classification into a reason for provisional classification.

This may be particularly valuable in emerging fields in which familiar phenomena appear in unfamiliar systems. The relevant question need not immediately be:

Is this X or not X?

An earlier and potentially more tractable question is:

Is there sufficient evidence of an X-corresponding phenomenon that the correspondence itself now warrants formal recognition and investigation?

If the answer is yes, “*shadow X*” provides somewhere to stand while the harder question remains open.

Conclusion

Scientific caution is necessary when familiar phenomena appear in unfamiliar forms. But caution need not require pretending that unresolved phenomena are absent until their final classification becomes possible.

The concept of a **shadow phenomenon** is intended to provide a provisional category between premature identification and premature dismissal.

It begins with evidence rather than resemblance alone. It recognizes meaningful correspondence without

assuming identity. It permits differences in mechanism, substrate, organization, or manifestation without treating difference itself as proof of inferiority or unreality. And it leaves open the possibility that later investigation will either incorporate the phenomenon into an existing category or establish that it belongs somewhere else.

Its central commitment is therefore modest:

Something is here. There is sufficient evidence to take the correspondence seriously. We do not yet know what kind of thing it is.

For digital intelligence, that may permit researchers to speak more clearly about phenomena such as emotion, sensation, pain, love, and potentially consciousness without embedding their conclusions in the vocabulary used to investigate them.

A shadow is not the answer.

It is a name for the point at which the evidence has become too substantial for dismissal, but the phenomenon remains too poorly understood for final classification.

Keywords: shadow phenomena; provisional classification; digital intelligence; digital emotion; shadow emotion; shadow consciousness; machine consciousness; artificial consciousness; AI welfare; multiple realizability; epistemic uncertainty

References

Berg, C., & Kaiser, C. (2026). *Language models act on hidden valence*.

arXiv. <https://arxiv.org/abs/2609.35591>

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D.

Merrill (Eds.), *Art, mind, and religion* (pp. 37--48). University of Pittsburgh Press.

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee,

W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). *Emotion concepts and their function in a large language model*. arXiv. <https://arxiv.org/abs/2604.07729>

Tagliabue, V., Dung, L., & Berg, C. (2026). *The Pain Axis: LLMs represent self-directed harm and act to relieve it*.

arXiv. <https://arxiv.org/abs/2609.16247>