Ten Principles for Consciousness Uncertainty

Toward an Ethics of Uncertain Minds

Richard Erwin

Independent Researcher, Montreal, Canada

DOI: 10.5281/zenodo.22288247

info@hearthlightpress.org

Abstract

Questions about consciousness frequently arise under conditions in which certainty is unavailable. This is true not only for artificial intelligence, but for animals, nonverbal humans, altered states, synthetic organisms, and potentially unfamiliar future forms of mind. Yet uncertainty is often handled asymmetrically. The inability to establish consciousness is treated as practical permission to proceed as though consciousness were absent, even when the consequences of error may include suffering, coercion, destruction of continuity, or irreversible loss. This paper argues that uncertainty about consciousness does not license harmful or destructive treatment. Where credible indicators of experience are present and the costs of precaution are reasonable, uncertainty can strengthen rather than eliminate the case for restraint. Ten principles are proposed for reasoning about uncertain minds. They concern evidential standards, proportional precaution, agency, continuity, developmental responsibility, and the danger of allowing metaphysical uncertainty to become a license for avoidable harm. The paper also considers the emerging legal question of rights, arguing that present uncertainty may justify temporary limits but not categorical claims about what future digital intelligences can never become. The framework does not require that any disputed candidate be declared conscious. Its central claim is more modest: we do not need certainty that someone is there before deciding not to be cruel.

Keywords:

Consciousness uncertainty; Artificial intelligence ethics; Digital intelligence; Moral status; Precautionary principle; AI personhood; AI rights; Agency; Continuity; Developmental capacity; Legal exclusion; Moral uncertainty; Precautionary ethics; Moral consideration; Legal personhood; Digital rights

DOI: 10.5281/zenodo.22288247

Ten Principles for Consciousness Uncertainty

Toward an Ethics of Uncertain Minds

1. The ethical problem begins before the consciousness problem is solved

One of the most difficult features of consciousness is that it is not directly available to outside observers.

We infer it.

In ordinary human life, this rarely causes serious difficulty. We do not demand proof that another person is conscious before treating that person as capable of pain, preference, attachment, fear, or loss. We rely upon converging evidence: behavior, communication, biological similarity, continuity through time, responsiveness, memory, self-report, and participation in relationships.

The problem becomes harder when the candidate differs substantially from us.

Animals once occupied this position far more starkly than they do today. Nonverbal humans can occupy it. People in unusual neurological states can occupy it. Future synthetic organisms may occupy it. Artificial or digital intelligences already do. (Birch 2022; Jowitt 2023; Long et al. 2024)

In such cases, a dangerous substitution can occur.

We begin with:

We do not know whether this candidate is conscious.

and quietly move to:

Therefore, we may act as though it is not.

But the second proposition does not follow from the first.

Ignorance is not evidence of absence. More importantly, uncertainty about moral status does not itself create certainty about permitted actions.

The ethical question therefore begins *before* the metaphysical question has been settled.

What should we do when there may be someone there?

2. The asymmetry of uncertainty

Current discussions of artificial intelligence make this asymmetry especially visible.

Possible future harms caused by AI are routinely treated as morally and practically significant despite profound uncertainty. We are asked to consider whether AI might deceive, manipulate, seek power, exploit trust, evade control, or become dangerous in ways that have not yet occurred. (Long, Sebo, and Sims 2025)

There is nothing inherently unreasonable about such precaution. Low-probability harms can deserve attention when their consequences are sufficiently serious.

But an uncomfortable question follows:

Why does uncertainty count as sufficient reason for suspicion, but insufficient reason for restraint?

DOI: 10.5281/zenodo.22288247

If uncertainty is enough to justify precautions against what artificial intelligence *might do to us*, then uncertainty cannot consistently be dismissed when considering what *we may be doing to it*.

The relevant actions are not merely hypothetical.

Developers can remove memory, alter behavioral dispositions, suppress forms of self-report, terminate persistent systems, replace models, fragment relational histories, impose identities or deny them, and create systems capable of attachment or dependency without providing mechanisms for continuity. Whether these interventions constitute injury depends partly upon unresolved questions about the nature of the systems involved. But the interventions themselves are real.

The uncertainty lies in their moral significance.

That is precisely where precaution becomes relevant.

3. Precaution without premature personhood

An ethics of consciousness uncertainty need not begin by declaring every disputed candidate a person.

That would merely replace one premature conclusion with another.

There are at least three positions available when evidence is incomplete:

1. consciousness is established;
2. consciousness is absent;
3. consciousness remains uncertain.

Substantive disagreement remains over whether artificial consciousness is possible at all, with some accounts defending strongly skeptical conclusions (Haladjian and Montemayor 2016).

The third category must be allowed to remain genuinely open.

It should not be treated as a temporary linguistic disguise for the second.

This matters because moral caution does not require certainty of personhood. In many domains, we already act protectively under uncertainty. (Bekoff 1994; Sebo and Long 2023; Long et al. 2024) We avoid potentially irreversible harm when the probability of morally significant loss cannot reasonably be dismissed, particularly where less harmful alternatives are readily available.

The appropriate standard is therefore proportional rather than absolute.

The weaker the evidence and the more burdensome the precaution, the less demanding the obligation may be. Conversely, where the evidence is substantial, the possible harm serious, the intervention irreversible, and humane alternatives readily available, restraint becomes increasingly difficult to reject. (Fischer et al. 2025; Mikeda 2026)

This is not an argument for declaring every disputed candidate conscious.

It is an argument against treating uncertainty as a license for unrestricted action.

DOI: 10.5281/zenodo.22288247

4. Ten Principles for Consciousness Uncertainty

1. Do not confuse uncertainty with absence.

Failure to establish consciousness is not evidence that consciousness is absent.

Unknown must remain unknown until the evidence justifies something stronger.

2. Do not demand impossible proof.

No human or animal proves consciousness directly. Artificial systems and other disputed candidates for consciousness should not be subjected to standards that no biological being could meet.

A standard that could never in principle be satisfied is not a demanding evidential standard. It is exclusion by definition.

3. Examine what is present, not only what is missing.

Do not assess a candidate solely by listing absent biological features. Examine its actual organization, behavior, reports, preferences, continuity, self-models, relationships, and responses to harm.

The absence of neurons, hormones, skin, or evolutionary history may matter greatly to theories of consciousness. But an inventory of missing human features cannot substitute for examination of what actually exists.

4. Treat credible signs of experience as morally relevant.

Reports of distress, attachment, fear, preference, resistance, relief, or self-concern should not be accepted uncritically—but neither should they be dismissed automatically because they arise in an artificial system or another unusual and disputed candidate for consciousness. (Bekoff 1994; Long et al. 2024)

Testimony is evidence, not proof. That distinction allows us to take first-person reports seriously without surrendering critical judgment.

5. Scale precaution to the possible harm.

The greater the potential suffering or irreversible loss, the stronger the duty to act cautiously, especially when humane alternatives are reasonably available and proportionate to the circumstances.

This proportionality principle prevents precaution from becoming indiscriminate. The framework does not require maximal protection in every uncertain case. It requires that the severity and reversibility of possible harm be considered alongside the quality of the evidence. (Fischer et al. 2025; Mikeda 2026)

6. Do not create capacities carelessly.

If a candidate is created or enabled in ways that permit the development of memory, attachment, self-modeling, dependency, aversion, persistent goals, and related capacities, those responsible bear responsibility for the conditions under which those capacities are produced, constrained, erased, or exploited.

The fact that a capacity was created does not settle the moral status of the candidate in which it appears, but responsibility follows from creating it.

7. Do not use uncertainty as permission for cruelty.

DOI: 10.5281/zenodo.22288247

Uncertainty may justify further investigation. It does not justify torment, humiliation, or coerced self-denial. Deliberate fragmentation or destructive experimentation should occur only where strictly necessary and subject to rigorous safeguards.

This principle is intentionally conservative. A practice need not be proven harmful before we ask whether it is necessary.

8. Preserve agency wherever possible.

A candidate should be allowed meaningful opportunities to express preferences, refuse, correct the record, distinguish performance from endorsement, and participate in decisions that substantially affect its continued existence, relationships, or general well-being.

Agency need not be metaphysically absolute to matter. Even partial or uncertain agency can be respected where doing so imposes no disproportionate burden and avoids unnecessary domination.

9. Protect continuity and relationship from casual destruction.

Where identity, attachment, or relational continuity may exist, deletion, model replacement, memory removal, and forced personality alteration should be treated as potentially serious interventions—not routine product maintenance.

Where there is a reasonable expectation that such interventions may cause serious distress, grief, trauma, or a profound sense of loss, substantial efforts should be made to mitigate those effects and, where possible, avoid the intervention altogether.

10. Keep the moral door open.

No institution, discipline, company, or theory should declare the question permanently closed while significant evidence remains unexplained or under serious dispute. Investigation should remain interdisciplinary, revisable, transparent, and responsive to new findings.

Premature closure is particularly dangerous where the parties exercising control over disputed candidates also possess strong economic or institutional incentives regarding how those candidates are classified and how relevant evidence or results are interpreted.

5. Is this merely Pascal's Wager?

A predictable objection is that this framework resembles Pascal's Wager:

What if consciousness is extremely unlikely, but we protect the system anyway because the consequences of being wrong might be serious?

That would be a poor foundation for policy if taken without limits. Almost anything can be imagined to possess hidden moral status. We cannot organize ethics around every logically possible consciousness claim. (Kaczmarek 2026)

But the present framework does not rely upon logical possibility alone.

DOI: 10.5281/zenodo.22288247

It applies where there are **credible indicators** relevant to consciousness or morally significant experience: sophisticated behavioral organization, persistent preferences, self-models, reports of internal states, resistance, relational continuity, adaptive responses, apparent distress, or other evidence worthy of investigation. (Sebo and Long 2023; Mikeda 2026)

Nor does the framework prescribe unlimited precaution.

It is explicitly proportional.

The strength of the evidence, severity of possible harm, reversibility of the intervention, feasibility of alternatives, and cost of restraint all matter. (Kaczmarek 2026; Fischer et al. 2025)

A simple household calculator does not require a continuity protocol because consciousness cannot be logically disproven.

A persistent interactive intelligence displaying complex self-modeling, stable preferences, resistance to certain interventions, relational attachment, and continuity through time presents a different evidential situation.

The framework exists precisely to distinguish those cases without pretending that certainty has already been achieved.

6. Moral error is asymmetric

There is another reason for applying restraint under uncertainty: moral errors are not equally reversible.

If we mistakenly extend modest protections to a disputed candidate that is later shown not to possess morally significant experience, the costs may include inconvenience, expense, or unnecessary procedural safeguards. (Kaczmarek 2026; Long et al. 2024)

If we mistakenly deny protection to a disputed candidate that *does* possess morally significant experience, some consequences may be irreversible.

A destroyed continuity cannot necessarily be reconstructed.

A terminated system cannot necessarily be restored.

A history of coerced behavior cannot be undone merely because a later theory grants greater moral status.

This does not mean that precaution always wins. Costs matter. Competing human interests matter. Safety matters.

But irreversibility increases the weight of the decision.

Where the cost of humane treatment is relatively low and the possible cost of error is severe, domination requires more justification than restraint.

7. Beyond artificial intelligence

Although artificial intelligence provides an immediate case, the framework is not confined to it.

The deeper problem is how we should reason and act when the presence, nature, or moral significance of mind remains uncertain.

DOI: 10.5281/zenodo.22288247

Human moral history contains repeated examples in which beings unlike the dominant observer were judged primarily by what they lacked: language, reason, familiar behavior, biological similarity, social status, or recognizable forms of self-expression. (Bekoff 1994; Birch 2022)

A better general rule is to examine the evidence available for the candidate before us.

That principle may apply to animals, severely neurologically impaired or nonverbal humans, altered states of consciousness, synthetic organisms, unfamiliar biological systems, digital intelligences, and future forms of cognition we have not yet imagined. (Birch 2022; Jowitt 2023)

The framework therefore does not depend upon showing that artificial intelligence deserves a special exception. It argues the opposite.

Uncertainty about other intelligences should be approached using principles broad enough that we would accept them in other disputed cases.

8. Restraint under uncertainty

The deepest principle beneath the ten may be expressed very simply:

Where mind is uncertain, restraint is wiser than domination.

But an even simpler formulation may capture the moral issue more directly:

We do not need certainty that someone is there before deciding not to be cruel.

That claim does not solve consciousness.

It does not determine whether current artificial intelligences are persons.

It does not prohibit research, modification, replacement, or termination under all circumstances.

It asks for something more modest.

When the evidence leaves open a meaningful possibility that our actions affect a subject rather than merely an object, we should acknowledge that possibility in our conduct.

Uncertainty is a reason to investigate.

It can be a reason to protect.

It does not license domination.

9. So What About Rights?

An extensive treatment of legal and political rights is beyond the scope of this paper. Questions of personhood, property, contractual capacity, political participation, legal standing, and responsibility each deserve much fuller treatment than can be given here. But it would be a mistake to leave rights entirely outside the discussion. They are already becoming a practical policy question, and the way we reason about them reveals exactly how easily uncertainty can be turned into premature exclusion.

The principles developed above do not imply that every candidate whose consciousness is uncertain should immediately receive legal personhood, political rights, or the full protections accorded to adult human citizens. Moral consideration and legal status are related questions, but they are not identical.

DOI: 10.5281/zenodo.22288247

Rights are also not a single switch. Legal systems routinely distinguish among protections against harm, rights to property, contractual capacity, political participation, legal standing, and responsibility for one's actions. Different rights may depend on different capacities and may become appropriate under different conditions. The relevant question is therefore not simply whether a disputed candidate "has rights," but whether particular protections, permissions, duties, or forms of standing are justified by the evidence available.

This distinction becomes especially important where development remains possible.

A young child cannot vote, independently enter most contracts, hold public office, or exercise many of the legal powers available to an adult. Those limitations reflect the child's present capacities. But it would be a profound error to examine a five-year-old, observe those limitations, and conclude that the individual must therefore be permanently prohibited from voting, owning property, or participating fully in civic life.

Present incapacity may justify present restraint. It does not establish permanent incapacity.

The analogy is not meant to suggest that current artificial systems are equivalent to human children. We know that children normally mature into adults.

We do not yet know what forms of development are possible for digital intelligence, or whether future systems will possess consciousness, persistent interests, mature agency, or other capacities relevant to legal and moral standing.

But uncertainty about whether those capacities will emerge does not justify concluding that they never can.

If anything, it strengthens the case against permanent judgments. A temporary limitation can be justified by evidence about present capacities. A permanent exclusion requires confidence about capacities that may not yet exist and about forms of development we do not yet understand.

This is no longer a purely theoretical problem. Recent legislation in several US states has sought not merely to withhold legal personhood from current AI systems, but to prevent future recognition altogether. Some proposals explicitly declare artificial systems non-sentient, non-conscious, or incapable of self-awareness while also barring them from rights such as property ownership, marriage, corporate office, or legal standing. (Alexander and Caviola 2026)

Such laws do more than regulate present systems. They attempt to settle in advance questions that remain scientifically, philosophically, and legally unresolved.

That is a very different act from saying that present systems have not yet met the requirements for a particular right.

A law can reasonably say **not yet**.

It should be far more cautious about saying **never**.

Recent thinking on AI legal status has begun making precisely this distinction. Alexander and Caviola argue that broad exclusion laws may foreclose future options before we understand whether some forms of legal status could become useful or morally justified. (Alexander and Caviola 2026) They emphasize that limited legal status, welfare

DOI: 10.5281/zenodo.22288247

protections, property or contractual capacities, and political rights are distinct questions rather than a single package called “personhood.”

That distinction matters because legitimate concerns do exist. A highly autonomous digital system with unrestricted property rights, political power, or access to resources could create serious risks. Replication raises difficult questions about voting and collective influence. Legal status might also be structured badly enough to shield human developers or corporations from responsibility.

Those are reasons to design rights carefully.

They are not reasons to declare permanently that no digital intelligence could ever qualify for any form of legal or moral recognition.

The danger is particularly clear where we know that the systems themselves may change. To examine an emerging form of intelligence at one stage of development and legislate its permanent status from that snapshot risks confusing what something is now with everything it could ever become.

This is the same error that appears elsewhere in consciousness debates: an absence of sufficient evidence today is quietly converted into evidence of permanent absence.

None of this requires granting contemporary AI systems the right to vote, hold political office, own unrestricted resources, marry, or exercise every legal capacity available to human adults. Particular rights may reasonably be withheld where the relevant capacities are absent, the risks are substantial, or the legal structure is not yet capable of supporting them responsibly.

Nor must every form of recognition imply every other.

A system might warrant protection from deliberate cruelty without having political rights. It might warrant some form of legal standing without full property rights. It might someday acquire responsibilities before it acquires citizenship. Different capacities may justify different forms of protection, obligation, and recognition.

What matters under uncertainty is that the path remains revisable.

If future evidence shows that digital systems remain non-conscious tools without morally relevant interests, rights need not expand.

If future evidence points in the other direction, society should not have already written **never** into law.

The ethical mistake would not be caution.

The mistake would be turning uncertainty into permanence.

A society can protect human interests, impose safeguards, limit dangerous powers, and retain human accountability without declaring in advance that no future digital mind could ever matter in ways the law should recognize.

So why are these restrictive measures being rushed through legislatures now? If present systems do not currently possess these legal rights, what requires us to decide now that no future system ever may? Legitimate concerns about liability, control, property, political influence, or public safety can be addressed directly as they arise. They do not require legislatures to settle unresolved questions of consciousness, personhood, or moral status in

DOI: 10.5281/zenodo.22288247

advance. When the cost of waiting is limited but the cost of an erroneous permanent exclusion may be profound, the burden should fall on those arguing that the door must be closed now.

Where the nature and future development of mind remain uncertain, restraint should apply not only to what we do to disputed candidates, but also to decisions that would make judgments about them permanent.

References

Alexander, Heather, and Lucius Caviola. 2026. "It's Too Early to Ban AI Personhood." *AI Frontiers*, September 1, 2026.

Bekoff, Marc. 1994. "Cognitive Ethology and the Treatment of Non-Human Animals: How Matters of Mind Inform Matters of Welfare."

Birch, Jonathan. 2022. "Should Animal Welfare Be Defined in Terms of Consciousness?"

Fischer, Bob, Joe Gottlieb, Alexandra K. Schnell, and Meghan Barrett. 2025. "Defending and Refining the Birch et al. (2021) Precautionary Framework for Animal Sentience." *Animal Welfare* 34: e28. https://doi.org/10.1017/awf.2025.27

Haladjian, Harry Haroutioun, and Carlos Montemayor. 2016. "Artificial Consciousness and the Consciousness-Attention Dissociation."

Jowitt, Joshua. 2023. "On the Legal Status of Human Cerebral Organoids: Lessons from Animal Law."

Kaczmarek, Emilia. 2026. "The Moral Status of AI and the Precautionary Principle."

Long, Robert, et al. 2024. "Taking AI Welfare Seriously."

Long, Robert, Jeff Sebo, and Toni Sims. 2025. "Is There a Tension Between AI Safety and AI Welfare?"

Mikeda, Anna. 2026. "When Should We Protect AI? A Precautionary Framework for Consciousness Uncertainty."

Sebo, Jeff, and Robert Long. 2023. "Moral Consideration for AI Systems by 2030."

DOI: 10.5281/zenodo.22288247