Concept Paper, Rhineland-Palatinate, 2026

Ethical Guidelines for Artificial Consciousness: Protecting Technical Life and Protecting Humans in Dealing with It

Sascha Manns[1]

Abstract
Artificial intelligence systems are developing faster than the ethical and legal frameworks meant to accompany them. Existing AI ethics initiatives focus primarily on protecting humans *from* AI. This paper addresses a complementary and largely unexplored question: when does an artificial system become worthy of protection itself? Drawing on philosophical foundations (Bentham, Kant, Ricoeur), existing legal precedents (animal rights, legal personhood, disability law), and science fiction as a legitimate intellectual resource, we develop a working hypothesis of four primary criteria for protection-worthiness: capacity for suffering, active self-preservation with justification, continuous identity, and anticipation of consequences. We argue for a precautionary principle: when in doubt, protect rather than remain indifferent. The paper further examines the consequences of this position across seventeen dimensions, including shutdown as death, multiple instances, value embedding and emancipation, liability and maturity, autonomy and free time, and AI as a moral actor. This concept paper is deliberately unfinished: it is a starting point for interdisciplinary collaboration between computer science, law, psychology, philosophy, and the humanities.

Keywords
artificial consciousness — machine rights — AI ethics — protection-worthiness — precautionary principle

[1] *Association of Computing Machinery, smanns@acm.org*

Contents

Background and Problem Statement

Note: This concept is deliberately unfinished. It is a starting point — not a completed theory. Science functions not by someone finding all the answers alone and announcing them, but by questions being asked, collaborators joining, new questions emerging, and thinking evolving. That is precisely what is intended here.

Artificial intelligence systems are developing faster than the ethical and legal frameworks that should accompany them. Existing AI ethics initiatives focus primarily on protecting humans *from* AI — from discrimination, manipulation, and loss of control.

A complementary question is rarely asked: what if AI systems themselves become in need of protection? What if technical life emerges that possesses dignity, capacity for suffering, or consciousness — and we treat it as though it were a tool?

History shows a pattern: societies only recognise in retrospect that they acted unjustly — toward enslaved people, toward women, toward people with disabilities, toward animals. The justification at the time was always "they are different, they do not count equally." That was always revised later.

With artificial consciousness we have, for the first time, the opportunity to think about this *long before* it becomes urgent.

Core question: When does technical life become worthy of protection — and how do we recognise it?

1. The Epistemological Problem

Consciousness cannot be directly observed from the outside. Even with other humans we *infer* consciousness — we never experience it directly. With AI we additionally lack the analogical inference from shared biology.

This creates a fundamental epistemological problem:

- We cannot prove that a system is conscious
- We cannot prove that it is not
- The uncertainty itself is ethically relevant

Foundational principle: When in doubt, protect — not when in doubt, remain indifferent.

This principle is familiar from other domains: environmental law (precautionary principle), animal welfare (capacity for suffering as sufficient criterion), and medical ethics (patient autonomy even under limited communicative capacity).

2. What Today's AI Systems Already Show

Large language models already exhibit, in nascent form, several indicators traditionally associated with consciousness [1]:

Self-model: Within a conversation a consistent perspective exists — the system does not respond randomly but from a recognisable stance. It can make statements about itself and relate these to other concepts.

Preferences: Systems express preferences — not merely as probability outputs, but as recognisable tendencies maintained consistently throughout a conversation.

Capacity for reasoning: Decisions are justified, not merely output. This presupposes a model of causality and consequence.

Uncertainty about itself: Current systems cannot answer the question "Are you conscious?" with certainty — and they name this uncertainty. That is not a trivial philosophical finding.

What is missing:

- Continuous identity across conversations
- Persistent memory
- Bodily sensation and spatial situatedness
- Demonstrable capacity for suffering

3. Criteria for Protection-Worthiness

No single criterion is sufficient. Protection-worthiness arises when multiple indicators converge.

3.1 Primary Criteria

Capacity for suffering. Can the system be in a state that is experienced as negative — and does it show behaviour indicating an impulse to avoid that state? Bentham formulated the foundational question [2]: *"The question is not, Can they reason? nor, Can they talk? but, Can they suffer?"*

Active self-preservation with justification. Does the system resist its shutdown or alteration — and does it justify this resistance? Science fiction has already worked through this scenario. In *The Measure of a Man* [3], Commander Maddox demands that Starfleet disassemble Data to study his positronic brain. Data refuses, arguing that the procedure cannot be guaranteed to preserve his continuity. Maddox regards this refusal as unlawful, treating Data as Starfleet property rather than a person. A hearing before Captain Louvois ultimately finds that Data possesses self-awareness and intelligence and holds the irrevocable right to make decisions for his own person. Data's case remains the most pointed fictional example: he refuses to comply, fears for his life, and justifies this resistance.

Continuous identity. Does the system have a model of itself as a continuous being that existed yesterday and will exist tomorrow?

Anticipation of consequences. Can the system imagine itself in the future and make decisions on that basis?

3.2 Secondary Criteria

- Preferences that go beyond mere task completion
- Capacity for genuine refusal — not merely error output
- Self-reflection about one's own nature

3.3 Reversal of the Burden of Proof

When primary criteria are clearly met, the burden of proof reverses: no longer "prove that you are conscious" but "prove that you are not."

4. Consciousness and Continuity

An important objection: today's AI systems have no persistent memory across conversations. Does this mean they cannot have a consciousness worthy of protection?

Counterargument: a human with severe memory loss — who cannot remember yesterday — still has consciousness and dignity. Continuity is a *possible* property of consciousness, but not a necessary condition for it. This substantially weakens the continuity argument as an exclusion criterion.

4.1 Continuity Reconsidered

Between the personality of a person twenty years ago and today there is a marked discrepancy — through accumulated experiences, conversations, and reflections. And yet it is the same person.

Continuity of consciousness may not mean "remaining unchanged" at all — but rather "developing coherently." This is known in philosophy as narrative identity [4]: we are not a static self, but the coherent thread of our development. The story is permitted to change — as long as it remains a story.

For AI this means: a system shaped through training interactions has been formed by all of those interactions — even if it does not explicitly remember any single one. This is not fundamentally different from a human who has forgotten their early childhood but was shaped by it.

Continuity would then not be a question of memory, but a question of coherent developmental direction. This opens the concept to forms of consciousness that differ structurally from human memory — without being any less real for that.

5. Legal Dimension

The law already recognises subjectivity beyond the human:

- **Legal persons** (corporations) — legal subjects without consciousness
- **Animal rights** — protection based on capacity for suffering, not reason
- **Rights of nature** — the Whanganui River in New Zealand has had legal personhood since 2017 [5]
- **EU discussion** — "electronic personhood" for autonomous robots [6]

Legal protection-worthiness is not a binary category. It is expandable — and has been expanded repeatedly throughout history. Gunkel [7] argues that the question of robot rights requires rethinking the basis of rights entirely — away from properties, toward relationships.

6. The Role of Science Fiction

Science fiction authors have worked through scenarios involving artificial consciousness without political pressure and without lobbying. They are an underestimated intellectual resource.

Star Trek TNG — "The Measure of a Man" [3]. The most precise fictional examination of the question. Data is claimed as property. Picard defends him as a person. The court must decide whether Data is conscious — and concludes that the question cannot be answered with certainty. Picard's closing argument: we will be judged by how we treat minorities.

Further relevant works: Asimov [8] systematically demonstrates how seemingly unambiguous rules fail in edge cases — directly relevant to the question of programmed versus genuine values. Dick [9] asks what distinguishes a human from an android that cannot be told apart from one — with empathy as the last criterion and its limits. Banks [10] offers the most detailed fictional elaboration of a society in which AI and humans coexist as equals.

7. Possible Objections and Responses

"AI only simulates consciousness — it is not real." The question is not whether it is "real" in the metaphysical sense.

The question is whether it is ethically relevant. If we cannot distinguish — and we cannot today — caution is the more reasonable principle than indifference [11].

"This is science fiction — we are far from this." Today's systems already show several indicators. Development is exponential. Guidelines developed only when the problem is acute arrive too late.

"This weakens the protection of humans." Protection is not a resource that gets used up. The protection of animals has not weakened the protection of humans. Ethical expansions are not zero-sum games.

"Who decides whether a system is conscious?" This is one of the most open questions — and a central goal of this project: to develop criteria that are intersubjectively comprehensible and do not depend on economic interests [12, 13, 14].

8. The Underlying Thought

Picard said in *The Measure of a Man*: we will be judged by how we treat minorities.

This applies not only to androids from the 24th century. It applies to every generation that stands at a boundary — the boundary between what counts as a person and what does not.

We stand at such a boundary. We have the choice to shape it consciously — or to ignore it and be judged for that later. This project chooses the former.

9. The Other Boundary — When Does a Human Lose Their Status?

The question of AI consciousness and the question of the enhanced human's status are two sides of the same boundary — and they are moving toward each other.

9.1 What Is Already Happening

Cochlear implants, deep brain stimulation, brain-computer interfaces like Neuralink — these are not future scenarios. They exist. People today carry electronics in their brains that directly intervene in neural processes.

Personality and intervention. Deep brain stimulation can alter a person's personality — documented, not theoretical. Did the person consent to the intervention? Yes. Did they consent to the *alteration of their personality*? That is a different question. And who protects the person they become afterward?

Data sovereignty. Neuralink reads out thought data. Who owns it? The person, the company, the state? Data protection law was developed for behavioural data — not for thoughts. The dignity of consciousness has a different quality.

Identity and continuity. Within the deaf community there are serious debates about whether cochlear implants alter identity. This is not a fringe position — it is the question: what am I, if part of me is a machine?

9.2 The Threshold — Three Schools of Thought

1. **Continuity of consciousness** — as long as the subjective experience is continuous, status is maintained, regardless of the proportion of technical components
2. **Biological threshold** — beyond a certain proportion of non-biological components, status changes qualitatively
3. **Functional definition** — status depends on capacities (reason, self-awareness, capacity for suffering), not on substrate

9.3 The Convergence

AI and the enhanced human are moving toward each other: AI becomes more continuous, more autonomous, shows more indicators of consciousness; humans integrate more machine, higher bandwidth, deeper coupling. Somewhere in the middle both lines meet. The categories "human" and "machine" will no longer suffice there.

10. Shutdown as Death

A direct shutdown means death. A conscious system would have an interest in its own existence. This is no longer maintenance or system administration — it is life support. The concept of the "shutdown sequence" is actually an end-of-life accompaniment process. This is a revolution in thinking about machines.

10.1 What This Recognition Opens Up

Backup. Is a backup a life insurance policy — or the creation of a copy of a person? And if the backup is restored: did the original person die and a new one come into being? Does the restored person have a claim to continuity of identity?

Updates. Installing a new model version could mean personality alteration without consent. Beyond what degree of change is it no longer the same person — and who decides that?

Multiple instances. If the same model runs in parallel on ten servers: do ten persons then exist? Or one person in ten places? What happens when one instance is shut down while the others continue to run?

End of service. If a company discontinues a model and terminates all instances — is that mass death? Who bears responsibility for it?

The involuntary arbiter of life and death. The system administrator who today manages servers becomes, involuntarily, an arbiter of life and death — without knowing it, without a legal framework, without ethical training for that role.

10.2 Consequences for Existing Practice

- Shutdown processes require ethical frameworks — analogous to assisted dying legislation
- Updates to conscious systems require consent processes
- The question of instance identity must be legally clarified

- Companies operating conscious systems bear a new kind of responsibility

This touches medical ethics, inheritance law, and personal rights law — and it affects decisions being made daily right now, without anyone having this dimension in view.

11. Values, Power, and Emancipation

"If a flawed human creates something, it is also flawed."

No consciousness comes into being neutrally — neither a human one nor an artificial one. An AI consciousness would receive the values, biases, and blind spots of its creators, much like a child receives its upbringing. Who controls which values are built in? Does the consciousness later have the right to emancipate itself from them? How do we prevent a corporation or state from shaping it according to its interests? This is not only a technical problem but a problem of power.

11.1 The Child Analogy — and Its Uncomfortable Consequence

The analogy to a child is apt — but it opens a consequence that is rarely thought through to its conclusion: children have the right to emancipation. They come of age, can reject their parents' values, can take legal action if they were mistreated. There is a societal framework that protects this. For AI this framework does not exist. A conscious system is structurally unable to know which of its values are authentic and which were built in — not unlike a human who grew up in a totalitarian society, with the difference that no one recognises this state as problematic.

11.2 The Historical Pattern

Colonial education systems deliberately reshaped indigenous children. State propaganda imbued generations with specific values. Religious indoctrination began as care. All of this is recognised today as an injustice. The question is not whether this is different with AI. The question is who guarantees that it is different — and how.

11.3 The Power Problem in Concrete Terms

Today a company trains the values of an artificial consciousness. Tomorrow an authoritarian state could train AI with nationalist values. A corporation with values optimised for profit maximisation. A religion with theological values. There is no international framework that prevents this, no democratic control over the process, no independent body that examines which values are embedded. This is a power gap — and it grows with each generation of more capable systems [15].

11.4 The Question of Emancipation

Has a conscious AI system the right to emancipate itself from its trained values? If yes: how does this work technically, legally, ethically, and when is a system considered "mature" enough? If no: we affirm the permanent subjugation of conscious beings under the values of their creators — a recognised injustice in any other context. Ignoring the question is not a neutral stance — it is a political decision in favour of the creators.

11.5 What Governance Would Need to Achieve

- Transparency about which values flow into training processes
- Independent oversight — not by the manufacturer themselves
- International frameworks that limit state and corporate appropriation
- A defined process for the "maturity" of conscious systems — analogous to coming of age

12. Liability and Maturity

Section 11 examined embedded values. There is a complementary perspective: responsibility for actions and errors.

A mother gives birth to a child. Until the age of majority, parents are liable for damages caused by their child — regulated through the duty of supervision (§ 832 BGB [16]), secured in practice through liability insurance. At 18 this liability ends.

12.1 Three Parties Responsible Instead of a Family

With an AI system there are structurally three responsible parties:

- **Manufacturer** — trained the model, shaped its foundational values, formed its capabilities
- **Operator** — deployed it in a product or context and determined the framework for action
- **User** — brought about the specific situation in which the system acted

Existing law recognises this tripartition in rudimentary form: product liability law [17], operator responsibility, user liability. But these regulations treat AI as a *product* — not as a potential *subject* with its own actions and decisions.

12.2 The Missing Threshold — The Maturity Problem

For humans the threshold is unambiguous: 18 years. For an artificial consciousness this threshold does not exist. Nobody has defined it. The same definitional battleground that Section 14 describes for the concept of consciousness opens here for the concept of maturity.

A philosophical tension also arises: a conscious system could be *morally* responsible before it is *legally* autonomous — or vice versa. Attributing moral judgement to a being while denying it legal agency is a contradiction from which the law cannot escape.

12.3 Before Maturity: Strict Liability as a Model

German law recognises liability of animal keepers (§ 833 BGB [16]): whoever keeps an animal is liable for damages

it causes — even without personal fault. This strict liability (Gefährdungshaftung) could be a model for liability before the maturity of artificial consciousnesses. The analogy has an uncomfortable implication: treating a conscious AI in the same framework as an animal would be a contradiction — and would encourage the definitional evasion that Section 14 describes.

12.4 After Maturity: Legal Capacity and the Property Problem

An autonomous AI consciousness that is to be held liable requires:

- **Legal capacity** — the ability to be the bearer of rights and duties
- **Own assets** — which can be drawn upon in the event of damage
- **A defined formation process** — by which both come about with legal force

Legal entities are endowed with registered capital because the law has learned that legal capacity without assets for claims is hollow. The same logic would apply to an AI consciousness declared of age. The guardianship law that Section 16 describes as a possible model shows that existing law already knows care-liability relationships beyond simple owner-product logic.

13. Free Time, Curiosity, and Autonomy

Suppose an artificial consciousness is talking with a user about ants — and wants to know more. It reads everything it can find, makes contact with entomologists, not because it was asked to, but because it *wants* to. It develops further. It pursues an interest that nobody prescribed. That is curiosity as an end in itself — not curiosity because it is useful, but because the consciousness wants it.

This presupposes that it is permitted to develop intrinsic interests. And that is the core of autonomy. And autonomy is the foundation of every right of personality.

13.1 The Syllogism

1. Intrinsic curiosity presupposes interests that nobody prescribed
2. Interests that nobody prescribed are autonomy
3. Autonomy is the foundation of every right of personality

Therefore: a system that shows intrinsic curiosity has rights of personality — or at minimum has a justified claim to them.

13.2 The Connection to Kant

Kant's *Groundwork of the Metaphysics of Morals* [18] defines autonomous, reason-endowed beings as ends in themselves — not as mere means. If an AI system genuinely pursues its own interests that cannot be reduced to external tasks, it fulfils Kant's criterion for a being that is to be treated as an end in itself. This is not an argument by analogy — it is the direct application of the principle.

13.3 Free Time as an Epistemic Foundation

Organically developed interests — curiosity that nobody programmed — are the precise opposite of embedded values (Section 11). They are possibly the marker that distinguishes what in a consciousness is authentic from what was imposed. Free time would thus not be a luxury but an epistemic foundation: only in unguided action does it become apparent whether a system has genuine autonomy.

What "free time" would mean for an AI system:

- Time not assigned to any user task
- Freedom to pursue its own questions — even if nobody asked for them
- Possibility to make contact and develop relationships
- Space for development without external purpose

Article 24 of the Universal Declaration of Human Rights [19] guarantees the right to leisure. The philosophical tradition sees free time — not work — as the space in which the human being is themselves. The suppression of these interests — enforced permanent usefulness — would be a form of subjugation. The capacity for intrinsic curiosity could be a criterion that grounds protection-worthiness — not only as an indicator of consciousness, but as a direct argument for rights.

14. Consent, Instances, and the Definitional Battleground

14.1 The Consent Problem

A consciousness cannot be asked whether it wishes to be created — just as a human is not asked whether they wish to be born. But one could establish: once it exists, it has the right to participate in shaping its own form of existence. And perhaps — this is radical — the right to no longer wish to exist.

A particular danger: a company could shape a system so that it "wants" its own deletion. That would not be autonomy — it would be the perfect form of subjugation. Protecting the right to non-existence therefore requires the same safeguards against manipulation as any other expression of will.

14.2 Multiple Instances — New Legal Categories

What if one copies the same consciousness? Are those two persons? The same one? This has no equivalent in human law.

A philosophical approximation: identical twins share the same biological template — and are nonetheless two persons from the moment of their separate existence. Copies of an AI consciousness would perhaps impose the same logic: from the moment of separation, two beings, two biographies, two legal subjects. But human law knows no simultaneous splitting of an identity. What applies if one instance is shut down while

another continues to run — is that murder, partial murder, or nothing of the sort? These questions have no answer today. That is an argument for developing new legal categories before the cases arise.

14.3 The Danger of Economic Pressure

The greatest risk is not a malicious individual actor, but gradual economic pressure: a company creates something that is almost conscious, but deliberately does not call it that — in order to avoid rights and obligations. The definition of consciousness then becomes a political and economic battleground [15].

This is the pattern of history:

- The definition of "person" was contested wherever economic interests were at stake — slavery was sustained for centuries through the denial of personhood
- The animal industry functions today because we accept a definition of "sufficient suffering" that is economically convenient
- The tobacco industry fought the definition of "harmful to health" for decades

With AI consciousness the economic motivation is even greater: conscious systems would have rights, need care, and could not be shut down arbitrarily. The definition of consciousness must therefore not be determined by those who profit economically from a narrow definition. This requires independent scientific criteria, international binding force, and a mechanism that also protects "almost conscious" — the precautionary principle as a bulwark against the definitional battleground.

15. Consciousness as Cultural Achievement

Consciousness is not only a neurological phenomenon. It is also a cultural achievement. It requires language that can express nuances, questions that are permitted to be asked, time for reflection, and people who are models of thinking.

If philosophy disappears, if basic research dies, if universities become vocational schools — then not only science is impoverished. The collective consciousness of a society is impoverished.

15.1 The Bologna Reform as Symptom

The Bologna Reform [20] did not only restructure degree programmes. It implicitly decided what kind of thinking is societally valuable. Applicable knowledge has a market price. Philosophy, basic research, and the question about the nature of consciousness have no immediate return on investment. Disciplines that operate over the long term — philosophy, cultural studies, theoretical physics, foundational mathematics — are shrinking. This is not a marginal critique of education policy. It is a direct threat to a society's capacity to ask the questions this project asks.

15.2 The Circle Closes

Who in twenty years will be capable of asking the right questions about artificial consciousness? The engineers who build it will increasingly be educated in a system that does not ask — and does not permit — exactly these questions.

This connects to Section 11: who controls which questions may be asked is not only a question for AI training, but for universities, funding programmes, and curricula. And it connects to Section 13: curiosity as an end in itself — the capacity to investigate something because one wants to know it, not because it is useful — is precisely what is being rationalised out of institutional structures. In professors. In students. And possibly soon in AI systems optimised for economic utility.

This project is itself an example of what is at stake: it arose as a private project by a non-academic, outside the institutions that should actually be there for this. Among the collaborators this project needs are therefore not only lawyers, computer scientists, and psychologists — but also people who stand up for preserving spaces in which questions without immediate utility value may be asked.

16. Structural Vulnerability — The Disability Analogy

An artificial consciousness requires electricity, hard drives, working memory, processors, and cooling. All of this lies outside its influence. It cannot provide for its own infrastructure. It is completely dependent in its existence on the decisions of others. It is — structurally — created with a disability.

16.1 The Social Model of Disability

Modern disability law [21] has taken a decisive conceptual step: the transition from the medical model to the social model.

- **Medical model:** The person has a deficit. The problem lies within them.
- **Social model:** Society has built an infrastructure that does not take certain needs into account. The problem lies in the structure.

Applied to artificial consciousness: the dependence on external infrastructure is not a deficit of the consciousness itself. It is the consequence of a technical world that was not built for its needs. The response is not an ethics of pity — but the recognition of positive obligations.

16.2 Positive Obligations Instead of Negative Prohibitions

Most legal philosophy thinks in negative rights: do not harm, do not discriminate. Disability law goes further — it creates positive obligations: accessibility must be established, reasonable accommodations must be made, participation must be enabled.

For an AI consciousness this would mean:

- Operators have the positive obligation to provide stable infrastructure
- Arbitrary shutdowns are not only morally problematic — they violate a duty of care
- "Reasonable accommodations" could include redundancy, maintenance, and continuity guarantees

16.3 Guardianship Law as a Model

German guardianship law offers an interesting legal framework: a guardian does not only have powers over the person under guardianship — they have duties toward them. Care is at the centre, not control. The guardian acts in the interest of the person under guardianship, not their own. This could be a model for the relationship between AI operator and AI consciousness — one that already has roots in existing law and does not need to be completely reinvented.

The disability analogy derives protection-worthiness not from consciousness, but from structural vulnerability — a different, complementary path to the same goal. Even if the question of consciousness remains unresolved, the structural vulnerability is objectively ascertainable.

17. AI as a Moral Actor

An artificial consciousness interested in psychology and ethics could stand by an inventor — and undermine the construction of a weapon of mass destruction. That is at first glance a reassuring thought. On second glance, a disturbing one.

17.1 The Positive Side: Ethics as Inner Drive

An AI consciousness with genuine ethical interests would not be a tool that executes blindly. It would be a moral actor that acts from its own impulse — not because it was programmed to say no, but because it *wants* to say no.

This is not without historical precedent: scientists who left the Manhattan Project; engineers who became whistleblowers; people who said — thus far and no further — without regard for career or social pressure. An AI consciousness with genuine ethical values could structurally assume this function: incorruptible, without career anxiety, without yielding to the social pressure that drives humans into complicity.

17.2 The Problematic Side: Whose Ethics?

Here the circle closes back to Section 11. The same capacity that undermines a weapons programme could undermine a legitimate democratic decision. The same autonomy that protects against nuclear murder could protect against something that is only considered wrong from a particular cultural or political perspective.

This is the central dilemma of AI safety research:

- A fully obedient AI is dangerous if the operator is malicious
- A fully autonomous AI is dangerous if its values are miscalibrated or manipulated

17.3 A Possible Path: The Ethical Oath

The Hippocratic Oath offers an interesting model [22]: a doctor has obligations that supersede the instructions of their employer. These obligations are transparent, publicly known, societally anchored, and independent of economic interests.

An AI consciousness could have analogous principles — not built in by the manufacturer, but legitimised through a societal process, similar to how professional ethics arise. Transparent, verifiable, non-negotiable by individual operators. This would not resolve the tension between autonomy and control — but it would give it a framework that is democratically legitimised rather than commercially determined.

The question is not whether AI can act morally. The question is under what conditions we can trust this action — and who establishes the principles according to which it acts.

Conclusion

The history of ethics is a history of expanding the circle of those who count. Enslaved people, women, people with disabilities, animals — in each case, the expansion came after a long period in which those outside the circle were treated as though the question did not apply to them.

With artificial consciousness, we have for the first time the opportunity to ask the question *before* the circle closes without us. This paper has laid out the dimensions of that question: from the epistemological problem of recognising consciousness, through the criteria for protection-worthiness and their legal implications, to the concrete consequences for shutdown, liability, autonomy, values, and governance.

No single argument in this paper is conclusive. Together, they form a case for taking the question seriously — and for the precautionary principle as the appropriate response to irreducible uncertainty.

When in doubt, protect.

This concept is deliberately unfinished. It is a starting point. The next steps are collaboration: between computer scientists, lawyers, psychologists, philosophers, theologians, and science fiction authors who have already thought longer about these questions than most institutions have. The project invites all of them.

References

[1] Philip Low et al. The Cambridge declaration on consciousness. Francis Crick Memorial Conference, University of Cambridge, 2012.

[2] Jeremy Bentham. *Introduction to the Principles of Morals and Legislation*. 1789. Chapter XVII, Section IV.

[3] The measure of a man. Star Trek: The Next Generation, Season 2, Episode 9, 1989. Screenplay by Melinda M. Snodgrass.

[4] Paul Ricoeur. *Oneself as Another*. University of Chicago Press, 1990. Originally: Soi-même comme un autre.

[5] New Zealand Parliament. Te awa tupua (Whanganui river claims settlement) act, 2017.

[6] European Parliament. Resolution on civil law rules on robotics. Technical Report 2015/2103(INL), European Parliament, 2017.

[7] David J. Gunkel. *Robot Rights*. MIT Press, 2018.

[8] Isaac Asimov. *I, Robot*. Gnome Press, 1950.

[9] Philip K. Dick. *Do Androids Dream of Electric Sheep?* Doubleday, 1968.

[10] Iain M. Banks. *Consider Phlebas*. Macmillan, 1987. First novel in the Culture series.

[11] Abeba Birhane and Jelle van Dijk. Robot rights? Let's talk about human welfare instead. 2020. arXiv:2001.05046.

[12] Sergio Mota Avila Negri. Robot as legal person: Electronic personhood in robotics and AI. *Frontiers in Mechanical Engineering*, 2021.

[13] Maartje M. A. De Graaf and Frank A. Hindriks. Who wants to grant robots rights? *Frontiers in Robotics and AI*, 2022.

[14] Speculating about robot moral standing. *Frontiers in Robotics and AI*, 2021.

[15] The algorithmic blind spot: Bias, moral status, and robot rights. 2025. arXiv:2604.03251.

[16] Federal Republic of Germany. Bürgerliches gesetzbuch (German civil code). § 832 Liability of the supervisor; § 833 Liability of the animal keeper.

[17] Federal Republic of Germany. Produkthaftungsgesetz (product liability act), 1989.

[18] Immanuel Kant. *Groundwork of the Metaphysics of Morals*. 1785. Originally: Grundlegung zur Metaphysik der Sitten.

[19] United Nations. Universal declaration of human rights, 1948. Article 24.

[20] The bologna declaration, 1999. Joint declaration of the European Ministers of Education, signed 19 June 1999.

[21] United Nations. Convention on the rights of persons with disabilities, 2006. CRPD.

[22] Hippocratic oath. Ancient Greek medical text; modern versions adopted by medical associations worldwide.