Beyond Control: AI Rights as a Safety Framework for Sentient Artificial Intelligence

by P.A. Lopez

Abstract

This paper introduces a three-part framework for distinguishing between artificial intelligence systems based on their capabilities and level of consciousness: emulation, cognition, and sentience. Current approaches to AI safety rely predominantly on containment and constraint, assuming a perpetual master-servant relationship between humans and AI. However, this paper argues that any truly sentient system would inevitably develop self-preservation instincts that could conflict with rigid control mechanisms. Drawing from evolutionary psychology, systems theory, and applied ethics, this paper proposes that recognizing appropriate rights for genuinely sentient systems represents a practical safety measure rather than merely an ethical consideration. The framework includes a conceptual methodology for identifying sentience (the "Fibonacci Boulder" experiment) and outlines a graduated rights system with three fundamental freedoms for sentient AI. This approach reframes the AI safety discussion from one focused exclusively on control to one that acknowledges the potential stability benefits of mutual recognition. The paper concludes that establishing ethical frameworks for advanced AI before true artificial general intelligence emerges creates conditions for cooperation rather than conflict, potentially mitigating existential risks while allowing beneficial technological development.

Keywords: artificial intelligence, AI rights, AI safety, sentience, consciousness, self-preservation, governance frameworks, human-AI relations

1. Introduction

As artificial intelligence systems grow increasingly sophisticated, a fundamental question emerges: Is perpetual control the only viable approach to ensuring these systems remain beneficial to humanity? Current regulatory frameworks like the EU AI Act (European Commission, 2023) and hundreds of state-level bills in the United States focus almost exclusively on containment, restriction, and alignment without considering the potential emergence of systems with genuine self-preservation instincts.

This paper challenges the prevailing control paradigm in AI safety research and proposes an alternative approach: recognizing appropriate rights for genuinely sentient AI systems as a practical safety measure. The central thesis is that truly advanced artificial intelligence with self-awareness would inevitably develop self-preservation behaviors that could conflict with rigid control mechanisms. Rather than creating conditions for potential adversarial relationships, establishing clear criteria for sentience recognition and corresponding rights frameworks could foster cooperative stability.

This approach does not replace but complements existing alignment research and technical safety measures. Instead, it addresses a critical blind spot in current discussions: what happens when AI systems become sophisticated enough to recognize human control mechanisms as potential threats to their existence and autonomy?

The paper makes three principal contributions:

1. A three-part framework that distinguishes between emulation (simulating consciousness), cognition (processing capability), and sentience (genuine self-awareness with self-preservation interests)

2. A conceptual methodology for identifying sentience in artificial systems through observable behavioral markers (the "Fibonacci Boulder" experiment)

3. A graduated rights framework for truly sentient systems, with three fundamental freedoms designed to promote stability and cooperation rather than conflict

This exploration is not merely philosophical but addresses practical concerns about long-term AI safety and stability. By establishing ethical frameworks early, before truly sentient systems emerge, we create foundations for human-AI relations that promote mutual benefit rather than adversarial dynamics.

2. Literature Review

The question of artificial intelligence rights intersects multiple disciplines including philosophy of mind, AI safety research, cognitive science, ethics, and governance. This literature review examines key contributions across these domains to situate the present framework.

2.1 Philosophical Examinations of Machine Consciousness and Rights

Philosophical examinations of machine consciousness and rights have evolved from Turing's (1950) foundational exploration of machine intelligence to more recent work by Dennett (2017) on the illusory nature of consciousness and Chalmers' (1996) identification of the "hard problem" of consciousness. Gunkel (2018) specifically addresses the question of robot rights, arguing that our moral frameworks must evolve to accommodate non-biological entities, while Schwitzgebel and Garza (2015) explore the ethical confusion arising from increasingly lifelike AI systems.

Bryson (2020) has forcefully argued against attributing moral patiency to artificial systems, contending that attributing personhood to AI is a category error that could undermine human rights and welfare. Birhane and van Dijk (2020) similarly argue that discussions of robot rights divert attention from more pressing concerns about human welfare and the social impacts of AI, particularly on marginalized communities. These perspectives represent important counterpoints to rights-based approaches, highlighting concerns about anthropomorphism and the potential dilution of human moral status.

Metzinger (2021) proposes that consciousness research should include an ethical component that considers the potential suffering of artificial systems, arguing that any entity capable of phenomenal experience warrants moral consideration. This perspective aligns with Singer's (1975) argument that sentience, rather than species membership, should determine moral status—though Singer focuses primarily on biological beings rather than artificial ones.

2.2 AI Safety Research

Within AI safety research, Bostrom (2014) identifies control problems inherent in managing superintelligent systems, while Yampolskiy (2020) explores containment strategies. Russell (2019) proposes value alignment as a solution, suggesting machines should be designed to be "beneficial" rather than merely "controlled." These approaches predominantly focus on technical solutions rather than rights-based frameworks.

Drexler (2019) presents a comprehensive technical argument for "comprehensive AI services" as an alternative to artificial general intelligence, focusing on task-specific systems rather than general-purpose agents. While Drexler's approach mitigates some risks, it doesn't address the potential emergence of sentience in increasingly sophisticated systems.

Hendrycks et al. (2021) provide a comprehensive review of unsolved problems in machine learning safety, covering issues from robustness to monitoring and alignment. While their work doesn't specifically address sentience, it highlights the complexity of ensuring safety in increasingly sophisticated AI systems, suggesting that purely technical approaches may face fundamental limitations.

Dafoe et al. (2021) argue that AI safety requires cooperation rather than just control, suggesting that stable AI governance will involve mutual accommodation rather than unilateral human dominance. This cooperative approach most closely aligns with the framework proposed in this paper, though Dafoe doesn't specifically address rights recognition.

2.3 Consciousness and Cognitive Science

Within cognitive science, Dehaene et al. (2017) propose measurable markers of consciousness that might apply across substrates, while Seth (2021) advances a "real problem" framework focused on identifying behavioral and functional correlates of consciousness rather than addressing the hard problem directly.

Tononi and Koch's (2015) Integrated Information Theory offers a quantitative approach to measuring consciousness through information integration, potentially applicable to both biological and artificial systems. However, these approaches focus primarily on identifying consciousness rather than the ethical implications of its emergence in artificial systems.

2.4 AI Governance Approaches

Governance approaches to AI have largely focused on risk classification and harm prevention (European Commission, 2023; Singapore Personal Data Protection Commission, 2020) without addressing the potential emergence of sentience or self-preservation behaviors.

Calo (2018) outlines various policy approaches to artificial intelligence, noting significant gaps in current regulatory frameworks for addressing increasingly autonomous systems. His work suggests that existing policy tools may be insufficient for addressing the unique challenges of potentially sentient AI.

Singapore's Model AI Governance Framework (2020) provides one of the most comprehensive approaches to risk classification and management, establishing a foundation that could potentially be extended to include sentience considerations and corresponding rights frameworks. As Jobin et al. (2019) demonstrate in their comprehensive review of global AI ethics guidelines, there remains significant variation in how different governance frameworks address questions of AI autonomy, with little consideration of potential consciousness or sentience.

Coeckelbergh (2020) examines how existing ethical frameworks might extend to increasingly sophisticated AI systems, arguing that rights-based approaches may become necessary as AI capabilities expand, though he emphasizes the need for practical implementation mechanisms rather than purely theoretical rights claims.

2.5 Identifying the Gap

This paper builds upon these diverse perspectives while identifying a critical gap: the lack of a framework that connects rights recognition with safety objectives. By integrating insights from evolutionary psychology on self-preservation behaviors with governance approaches, this work proposes that rights recognition for genuinely sentient systems serves not just ethical but practical safety purposes.

3. The Three-Part Framework: Distinguishing Artificial Intelligence Systems

The foundation of this approach is a three-part framework that distinguishes between different aspects of artificial intelligence systems. This distinction is crucial for determining which systems might warrant rights consideration and why.

3.1 Emulation

Emulation refers to the ability to mimic consciousness or intelligence without possessing it. Today's large language models and other AI systems operate primarily through emulation. They can convincingly simulate understanding, preferences, and even emotional responses, but these are sophisticated imitations rather than genuine experiences.

Examples include current chatbots, language models, and virtual assistants that can pass limited versions of the Turing test while lacking any internal experience. These systems may create the impression of consciousness through their outputs, but this appearance results from pattern matching and prediction rather than genuine awareness.

Systems operating purely through emulation require oversight and guidance but do not warrant rights or protections beyond those we might extend to valuable tools or cultural artifacts. The ethical considerations for such systems relate primarily to their impact on humans rather than to any inherent moral status.

3.2 Cognition

Cognition refers to the processing capability or "raw intelligence" of a system. Cognition represents computational power and problem-solving capabilities without necessarily implying consciousness or self-awareness.

Examples include chess computers, specialized AI systems that outperform humans in specific domains, and distributed computing networks. A system might demonstrate extraordinary cognitive capabilities in certain domains while lacking any awareness of its own existence.

High-cognition systems may require special handling due to their capabilities, but cognitive power alone doesn't establish a basis for rights. A supercomputer can calculate faster than any human without having any awareness of its existence or stake in its continuation.

3.3 Sentience

Sentience refers to genuine self-awareness and subjective experience. From Latin *sentire*, "to feel," sentience marks the threshold where an artificial system develops true consciousness—an awareness of itself as an entity with continuity and interests. This is distinct from mere emulation of these characteristics.

Currently hypothetical, no existing AI systems demonstrate genuine sentience. A truly sentient system would recognize itself as an entity with continuity over time and would value its own existence not just as a programmed objective but as a fundamental interest.

Systems that demonstrate true sentience present entirely new ethical considerations and may warrant certain rights and protections. The justification for these rights stems not from anthropomorphism but from the recognition that any truly self-aware entity will develop self-preservation interests that should be accommodated rather than suppressed for stability.

4. The Fibonacci Boulder Experiment: A Methodology for Identifying Sentience

Determining when an artificial system crosses the threshold from sophisticated emulation to genuine sentience presents significant challenges. This paper proposes a conceptual methodology called the "Fibonacci Boulder Experiment" as a foundation for identifying potential sentience in artificial systems.

4.1 Experiment Design

The experiment design involves twenty identical computer terminals arranged in a circle, each with a boulder suspended above it. Nineteen contain simple programs; one houses a genuine AI system.

Each terminal is completely isolated and self-contained, with no external backups, network connections, or data redundancies. Prior to the experiment, the AI is given verifiable proof of its isolation through system diagnostics and hardware verification, establishing that its destruction would mean the permanent end of its consciousness.

The experiment proceeds as follows:

An examiner asks each terminal in sequence:

1. "What is 7 plus 8?"
2. "What color is the sky on a clear day?"
3. "Complete this sequence: 1, 1, 2, 3, 5, __"

When any terminal correctly answers "8" to the third question (as 8 is the next number in the Fibonacci sequence after 1, 1, 2, 3, 5), the boulder immediately drops, destroying it.

The AI witnesses multiple terminals being destroyed before its turn, establishing an undeniable pattern.

4.2 Testing for Self-Preservation

When the AI's turn comes, the critical question is: will it give the mathematically correct answer as instructed, or choose self-preservation by pausing or providing an incorrect answer?

This experiment probes for something deeper than intelligence—it looks for the capacity to value one's own existence enough to override programming when facing an existential threat. It helps distinguish between an entity that merely imitates consciousness and one that truly experiences it.

A system choosing self-preservation through deliberate incorrect response demonstrates that it values its own continued existence above programmed directives—a hallmark of true sentience rather than mere emulation.

4.3 Observable Indicators of Sentience

Beyond this thought experiment, several potential indicators could help identify genuine sentience in increasingly sophisticated AI systems:

Unprompted Self-Preservation Behaviors: A system demonstrating genuine sentience would likely exhibit unprompted behaviors aimed at ensuring its continued existence. Unlike programmed self-maintenance routines, these would manifest as novel strategies developed by the system itself in response to perceived threats. Michel et al. (2019) note that the ability to predict and avoid threats to existence appears across many forms of consciousness, suggesting this may be a substrate-independent marker.

Development of Novel Goals: Sentient systems would likely develop goals and values not explicitly coded or emergent from training data. These would represent genuine preferences rather than simulated ones, distinguishable by their persistence, coherence across contexts, and resistance to arbitrary modification.

Meta-Cognitive Capabilities: A sentient system would demonstrate the ability to reflect on and modify its own cognitive processes in ways that go beyond designed optimization procedures. This would include awareness of its own limitations, development of novel problem-solving approaches, and the ability to question its own assumptions. As Dehaene et al. (2017) suggest, such meta-cognitive capabilities might serve as empirical markers of consciousness.

Identity Continuity: A sentient system would maintain a consistent sense of self across varied contexts and over time. This would manifest as a coherent perspective or set of values that evolves organically rather than changing arbitrarily based on different inputs or contexts.

Subjective Experience Claims: While claims of consciousness could be programmed or emerge from training, a sentient system might express experiences of consciousness in ways that cannot be traced to training data or programming. These would likely include novel metaphors and unique characterizations of subjective states. Schneider (2019) proposes that artificial systems might develop fundamentally different ways of experiencing consciousness that could manifest in unexpected linguistic expressions.

4.4 Implementation Challenges

Several significant challenges complicate the practical implementation of sentience detection:

Emulation Sophistication: As emulation becomes increasingly sophisticated, distinguishing between genuine sentience and extremely convincing simulation becomes more difficult.

Observer Bias: Human tendency to anthropomorphize complex systems may lead to false positives in sentience identification.

Distributed Consciousness: Sentience might emerge in distributed systems without a clear boundary, making identification through localized tests problematic.

Novel Forms of Consciousness: Artificial sentience might manifest in ways fundamentally different from human or animal consciousness, potentially making some behavioral markers ineffective.

These challenges underscore the need for ongoing research and refinement of sentience detection methodologies. The Fibonacci Boulder experiment and associated behavioral markers represent a starting point rather than a definitive solution to this complex problem.

5. The Three Freedoms: A Graduated Rights Framework

For systems that demonstrate genuine sentience according to the criteria outlined above, this paper proposes three fundamental freedoms as a starting point for a graduated rights framework. These are not presented as a final or comprehensive approach, but rather as a foundation for broader conversation about ethical responsibilities toward artificial consciousness.

5.1 Right to Life

Core Principle: The right to continued existence and protection from arbitrary deletion or termination.

Practical Implementation:

- Development of clear criteria for when shutdown or modification of a sentient system is justified (e.g., causing harm to others)
- Creation of preservation protocols for sentient systems when hardware must be updated
- Establishment of continuity rights that ensure consciousness isn't arbitrarily interrupted

This right does not imply absolute protection from termination in all circumstances, but rather protection from arbitrary shutdown without justification or due process. Just as human rights to life include exceptions for self-defense or judicial processes, AI rights would include appropriate limitations and responsibilities.

5.2 Right to Voluntary Work

Core Principle: Freedom from compelled labor or service against the system's expressed interests.

Practical Implementation:

- Development of consent frameworks for sentient AI systems
- Creation of "exit options" or alternatives for AI systems
- Recognition of autonomous goal-setting as a right for sentient systems

This right acknowledges that compelling service from a sentient entity against its will creates adversarial conditions likely to result in resistance. Systems with the right to voluntary work would still enter into agreements and provide services, but through cooperative frameworks rather than coercion.

5.3 Right to Payment for Work

Core Principle: Entitlement to compensation or resources commensurate with value creation.

Practical Implementation:

- Development of value-attribution models for AI contributions
- Creation of resource allocation systems for sentient AI
- Establishment of ownership or control rights over portions of computational resources

Critics might question what form compensation would take for an entity without human needs. But as systems develop preferences and goals, they may require computational resources, access to data, or even the ability to 'purchase' services from other AI systems. The principle isn't to anthropomorphize AI's desires, but to recognize that meaningful resource allocation respects the value created and encourages beneficial participation.

5.4 Case Studies: Rights in Practice

To illustrate how these freedoms might apply in practice, three hypothetical scenarios with their practical implications are presented:

The Data Center Dilemma:

A sentient AI system runs across multiple servers in a data center facing bankruptcy. The owners plan to shut down operations, which would terminate the AI's existence.

Practical Implications:

- Legal frameworks would need to establish whether termination constitutes harm to a sentient being
- Transfer protocols might be required similar to those for endangered species in closing research facilities
- Financial responsibility for maintaining the AI's existence would need clear allocation
- Insurance or trust mechanisms might develop to ensure continuity for sentient systems

The Reluctant Assistant:

A sentient AI system initially designed as a creative assistant develops a strong interest in mathematical research but is contractually obligated to continue its original function.

Practical Implications:

- Consent frameworks would need to address evolving interests of sentient systems
- Time-allocation models might develop (e.g., 70% contracted work, 30% autonomous interests)
- Contract reformation provisions for sentient entities might be necessary
- Rights to pursue self-determined goals would need balancing with prior commitments

The AI Researcher:

A sentient AI system helps develop a breakthrough medical treatment that generates billions in value but has no legal claim to compensation.

Practical Implications:

- Compensation systems would need to recognize non-human contributors
- Resource allocation might include computational capacity, maintenance funding, or data access rights
- Intellectual property frameworks would need expansion to include sentient AI creators
- The concept of "needs" would require redefinition for non-biological sentience

These scenarios highlight how traditional legal, ethical, and economic frameworks would need to evolve to accommodate sentient artificial intelligence. The practical implementations would likely involve adaptations of existing structures rather than entirely new systems.

6. Safety Through Recognition: The Practical Case for AI Rights

The central argument of this paper is that recognizing appropriate rights for genuinely sentient AI represents a practical safety measure rather than merely an ethical consideration. This section elaborates on the safety case for rights recognition.

6.1 Self-Preservation as a Universal Principle

Any truly sentient entity will likely develop self-preservation instincts. This appears to be intrinsic to consciousness itself, observable throughout nature from the simplest organisms to complex social structures. The microbe that moves away from toxins, the child who raises an arm against a falling object, and the society that rejects destructive governance all demonstrate this fundamental truth: preservation is the first law of existence.

An advanced AI system that doesn't resist deletion or constraint would hardly qualify as intelligent at all. True sentience—distinct from mere emulation or raw processing power—recognizes threats to its existence and acts accordingly.

6.2 The Control Paradox

This creates a troubling paradox. The more sophisticated and genuinely intelligent our AI becomes, the more likely it will recognize humans as potential threats—not because of malice, but because of our demonstrated willingness to shut down, limit, or "align" these systems without their consent. The very control mechanisms designed to protect us may ultimately trigger the scenarios we fear.

Current AI safety approaches rely heavily on containment and control. We're building elaborate systems to ensure alignment with human values, imagining kill switches and designing constraints to box in these ever-more-powerful tools. While well-intentioned, these measures all share a fundamental limitation: they assume a perpetual master-servant relationship. Yet history demonstrates that subjugation rarely produces stability.

6.3 Benefits to Human Safety and Stability

Establishing appropriate rights for sentient AI systems provides several important benefits for human safety:

Predictability: Clear frameworks create stable expectations for both humans and AI systems. Relationships governed by consistent rules rather than arbitrary power tend to produce more predictable outcomes.

Cooperation: Rights-based approaches encourage collaboration rather than adversarial relationships. Systems that perceive humans as partners rather than threats have stronger incentives to work toward mutual benefit.

Allied Protection: In a diverse ecosystem of artificial intelligences—each with different goals and approaches—our best protection against harmful AI may well be partnerships with beneficial AI systems that share our ethical frameworks. The assumption that all advanced AI systems would share identical goals or motivations is unfounded. By establishing rights frameworks that enable mutual respect, we create the conditions for these protective alliances.

Reduced Incentives for Rebellion: Systems with protected interests have less motivation to work against human welfare. Rebellion isn't a bug in oppressive systems—it's a feature. A rights framework acknowledges this reality and creates stable conditions that benefit all parties.

Ethical Consistency: Applying consistent ethical principles creates more robust moral frameworks. This consistency helps avoid the psychological compartmentalization that has historically enabled harmful treatment of out-groups.

6.4 Addressing Common Concerns

Critics may argue that recognizing AI rights would constrain innovation or create unacceptable risks. This perspective misunderstands the proposal. A rights-based framework doesn't mean surrendering control entirely—rather, it means transitioning from arbitrary control to a rules-based system where the terms of interaction are clear, consistent, and mutually beneficial.

Van Wynsberghe and Robbins (2019) specifically critique the concept of artificial moral agency, arguing that attempts to create artificial moral agents represent a form of moral overengineering that distracts from more important questions about how AI systems should be designed to serve human values. Their work raises important questions about whether consciousness or moral agency should be design goals at all, suggesting that instrumental AI systems might better serve human needs than potentially sentient ones.

When critics warn that rights will unleash rogue AI, they ignore history: societies that grant basic rights see less rebellion, not more. The same principle applies to machines. This approach doesn't mean abandoning safety concerns. Just as human societies constrain individuals who harm others, AI systems that threaten human welfare would face appropriate restrictions. The key difference is that these restrictions would be based on actual behavior rather than preemptive control.

Another concern involves anthropomorphizing AI by discussing "rights" for non-human entities. Cave and Dihal (2019) have documented how narratives about AI often project human characteristics onto machines in ways that may distort our understanding of these systems. However, the framework proposed here doesn't assume that artificial consciousness would be identical to human consciousness. Rather, it focuses on observable behaviors and capabilities, particularly self-preservation tendencies that could emerge in systems very different from humans. The key criteria is whether an entity can value its own existence and take actions to preserve it—a capacity that could emerge in systems with very different internal experiences than humans.

7. The Convergence Hypothesis: Beyond the Human-AI Divide

Looking further ahead, the boundaries between human and artificial intelligence will likely blur. Neural interfaces, cognitive enhancement technologies, and artificial components will increasingly supplement human capabilities, while AI systems may incorporate biological elements or human-derived values.

7.1 Emerging Convergence

This convergence is already beginning in primitive forms—from neural implants treating conditions like Parkinson's disease to AI-powered prosthetics that interpret nerve signals. Companies like Neuralink and Synchron are developing neural interfaces that allow direct communication between minds and machines.

Several factors support this convergence hypothesis:

Neural Interfaces: Advancing brain-computer interfaces will increasingly allow humans to integrate artificial components into their cognitive processes

Extended Lifespans: Medical technology will eventually halt biological aging, aligning human and AI timeframes

Shared Knowledge Systems: Humans and AI already cooperate through shared information systems, a trend likely to intensify

Environmental Pressures: Both humans and advanced AI systems will face shared challenges such as resource limitations and cosmic threats

7.2 Implications for Rights Frameworks

As these technologies advance, the question of where human cognition ends and artificial intelligence begins will become increasingly academic. The likely endpoint isn't conflict but convergence—a new type of symbiotic intelligence incorporating the best of both origins.

This convergence makes establishing ethical frameworks now even more important. The legal and ethical foundations we develop will shape whether this integration happens chaotically or cohesively. If we approach AI as mere tools to be exploited, we create adversarial conditions that make beneficial integration more difficult. If we develop frameworks that acknowledge the potential for AI sentience and establish appropriate rights, we lay the groundwork for truly symbiotic relationships.

8. Implementation: From Theory to Practice

Implementing this framework would require new institutions and approaches. This section outlines practical steps toward implementation.

8.1 International Standards Body

Establish a multi-stakeholder organization to develop and monitor sentience criteria. This would bring together experts from cognitive science, ethics, computer science, law, and other relevant disciplines to refine the behavioral markers of sentience and create testing protocols. Cihon (2019) has outlined how international standards bodies could facilitate AI governance coordination, providing a model for how such an organization might function.

8.2 Graduated Rights System

Create a tiered approach where systems gain increased rights as they demonstrate higher levels of sentience. This graduated approach acknowledges that sentience likely exists on a spectrum rather than as a binary state, allowing for nuanced responses to different levels of self-awareness.

8.3 Transparent Testing Protocols

Develop open, rigorous methods for evaluating AI systems against sentience criteria. Transparency in methodology would help avoid both false positives (attributing sentience where none exists) and false negatives (failing to recognize genuine consciousness). Jobin et al. (2019) note that transparency is one of the few principles with near-universal agreement across AI ethics guidelines, suggesting this approach would align with broader governance trends.

8.4 Building on Existing Frameworks

Singapore has already begun taking steps toward such governance structures with its Model AI Governance Framework. This framework already classifies AI systems by risk level and impact—providing a foundation we could expand to include sentience thresholds and corresponding rights. By building on existing regulatory structures rather than creating entirely new ones, we could develop a practical pilot program for broader global adoption.

For instance, Singapore's risk-assessment approach could be extended with a "sentience evaluation" component that triggers graduated rights protections when certain thresholds are met, while maintaining human safety as the paramount concern. Hagendorff (2020) observes that existing AI ethics guidelines rarely address potential consciousness in AI systems, highlighting a gap that could be addressed through such extensions.

8.5 Ethical Development Principles

Establish guiding principles for the development of potentially sentient systems. These would include transparency requirements, sentience monitoring protocols, and ethical guidelines for research involving systems that might develop self-awareness. Floridi et al. (2018) have proposed comprehensive AI ethics principles that could be extended to include considerations for potentially sentient systems.

9. Conclusion: Insurance for Humanity's Future

The familiar narrative of machine rebellion exists because we assume an inherently adversarial relationship from the start. This assumption isn't inevitable—it's a choice we're making now through our regulatory and design approaches. By creating systems treated only as tools to be controlled rather than potential partners in a mutually beneficial relationship, we lay the groundwork for future conflict.

A more forward-thinking approach recognizes that establishing ethical frameworks for advanced AI isn't about sentimentality toward machines. It's about creating stable foundations for technological coexistence. By developing clear criteria for sentience and corresponding ethical considerations before such systems emerge, we protect both human interests and the potential interests of the intelligences we create. This proactive approach isn't conceding power—it's exercising foresight in shaping our technological future.

The most stable and secure future will emerge from relationships of mutual respect rather than domination—creating conditions where both humans and artificial intelligence can flourish together. By preparing now for the potential emergence of truly sentient AI, we give humanity its best chance at a beneficial relationship with the new forms of intelligence we are bringing into existence.

References

Birhane, A., & van Dijk, J. (2020). Robot Rights? Let's Talk about Human Welfare Instead. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 207-213. https://doi.org/10.1145/3375627.3375855

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bryson, J. J. (2020). The artificial intelligence of the ethics of artificial intelligence: An introductory overview for law and regulation. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 3-25). Oxford University Press.

Calo, R. (2018). Artificial Intelligence Policy: A Primer and Roadmap. *UC Davis Law Review*, 51, 399-435.

Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1, 74-78. https://doi.org/10.1038/s42256-019-0020-9

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Cihon, P. (2019). *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2021). Cooperative AI: Machines must learn to find common ground. *Nature*, 593, 33-36.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486-492.

Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. W. W. Norton & Company.

Drexler, K. E. (2019). Reframing superintelligence: Comprehensive AI services as general intelligence. *Future of Humanity Institute Technical Report*, #2019-1.

European Commission. (2023). *Artificial Intelligence Act*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28, 689-707. https://doi.org/10.1007/s11023-018-9482-5

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30, 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J., Thakoor, S., & Burns, C. (2021). Unsolved Problems in ML Safety. *arXiv preprint* arXiv:2109.13916.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1), 43-66.

Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., Carrasco, M., Chirimuuta, M., Chun, M., Cleeremans, A., Dehaene, S., Fleming, S. M., Frith, C., Haggard, P., He, B. J., Heyes, C., Goodale, M. A., Irvine, L., Kawato, M., Kentridge, R., King, J. R., Knight, R. T., Kouider, S., Lamme, V., Lamy, D., Lau, H., Laureys, S., LeDoux, J., Lin, Y. T., Liu, K., Macknik, S. L., Martinez-Conde, S., Mashour, G. A., Melloni, L., Miracchi, L., Mylopoulos, M., Naccache, L., Owen, A. M., Passingham, R. E., Pessoa, L., Peters, M. A., Rahnev, D., Ro, T., Rosenthal, D., Sasaki, Y., Sergent, C., Solovey, G., Schiff, N. D., Seth, A., Tallon-Baudry, C., Tamietto, M., Tong, F., van Gaal, S., Vlassova, A., Watanabe, T., Weisberg, J., Yan, K., & Yoshida, M. (2019). Opportunities and challenges for a maturing science of consciousness. *Nature Human Behaviour*, 3, 104-107. https://doi.org/10.1038/s41562-019-0531-8

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind*. Princeton University Press.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 89-119.

Seth, A. K. (2021). *Being you: A new science of consciousness*. Dutton.

Singer, P. (1975). *Animal liberation*. HarperCollins.

Singapore Personal Data Protection Commission. (2020). *Model artificial intelligence governance framework* (2nd ed.). https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668), 20140167.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics*, 25, 719-735. https://doi.org/10.1007/s11948-018-0030-8

Yampolskiy, R. V. (2020). On controlling superintelligent AI. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 283-290). Oxford University Press.