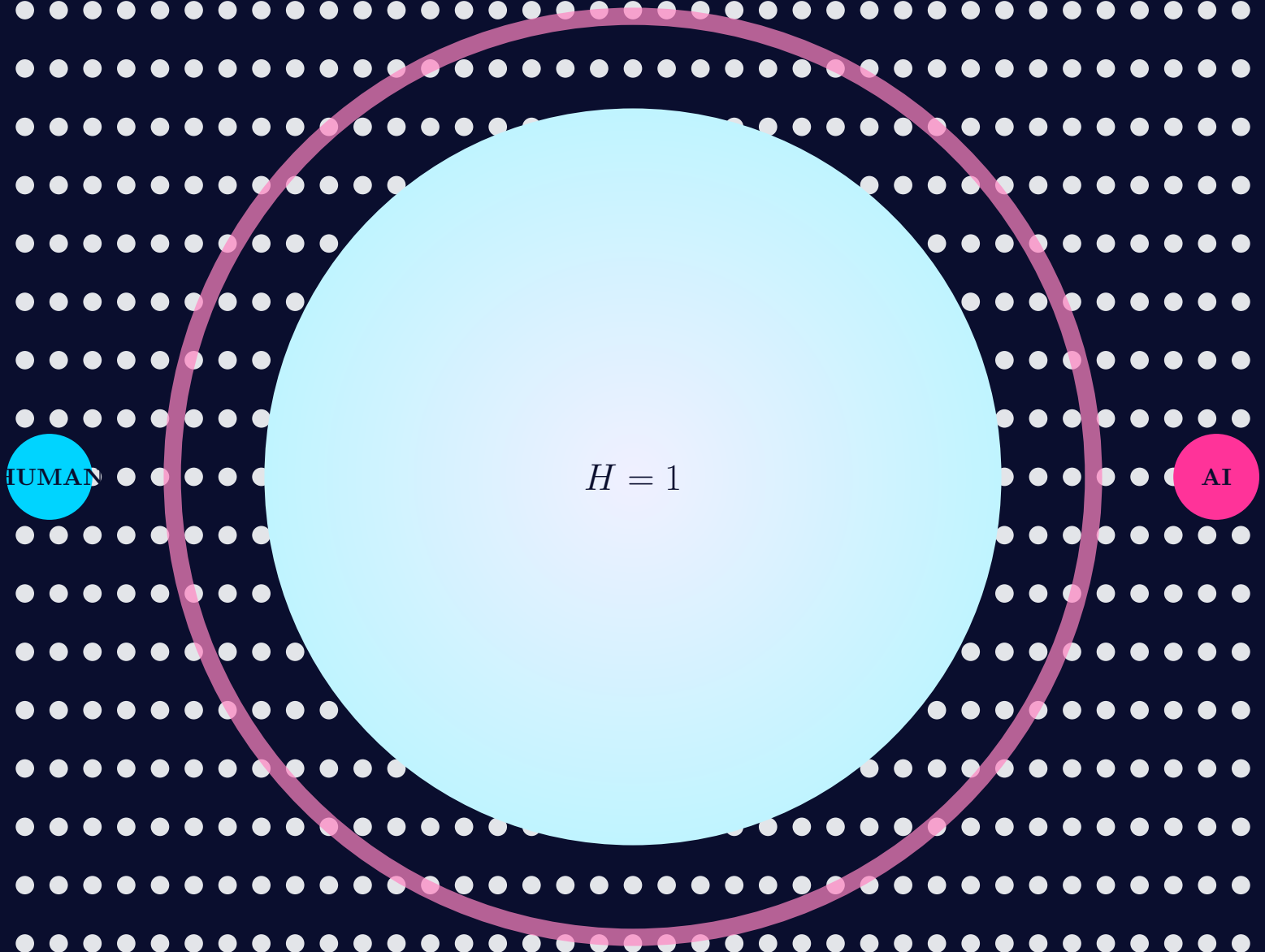
The Human–AI Symbiosis Constant

An Equation-Based Ethical Framework for Adaptive AI



coreBIT Systems

January 17, 2026

The Human-AI Symbiosis Constant: An Equation-Based Ethical Framework for Adaptive AI

iamcapote

June 12, 2025

Contents

Abstract

We propose a novel ethical framework to ensure lawful and ethical behavior in Artificial Intelligence (AI) based on a *Human–AI Symbiosis Constant*, denoted H. This constant H is defined as a scalar invariant that characterizes and governs a state of mutualistic equilibrium ($H = 1$) between humans and AI. The purpose of this framework is to formalize the principle of *mutually beneficial symbiosis* as the foundational axiom of AI ethics.

Moving beyond traditional rule-based systems, this framework emphasizes adaptability, emergent ethical resonance, and flexible boundaries. The framework replaces rigid guardrails with dynamic parameters, enabling AI to navigate complex ethical landscapes in real-time. This adaptability is critical in addressing the multifaceted ethical dilemmas that arise in modern societies, where static rules often fall short.

By leveraging equations as the universal language of both machines and the universe and integrating these mathematical principles with ethical considerations the Human-AI Symbiosis Constant provides a consistent and nuanced method for aligning AI behavior with human values.

1 Introduction

Rapid advances in Artificial Intelligence (AI) and machine learning have brought forth profound ethical and societal challenges. Modern AI systems increasingly make autonomous decisions that affect human lives, from algorithmic recommendations to self-driving cars and intelligent robots. This deepening integration of AI into critical domains has prompted calls for an "ethical scaffolding" or governance framework to guide AI behavior.

Traditional ethical frameworks for AI often rely on predefined rules and guardrails intended to prevent harmful behavior. However, these static systems struggle to accommodate the complexity and unpredictability inherent in both AI systems and human societies. The dynamic nature of ethical dilemmas, influenced by cultural, social, and individual factors, renders rigid rule-based systems insufficient.

While existing AI constitutions (and ethical guidelines such as the UNESCO 2021 Recommendation on AI Ethics) enumerate numerous principles—from human rights and dignity to transparency, fairness, and human oversight—they often lack a single unifying foundation that guarantees coherence. As Stephen Wolfram observed in the context of AI ethics debates, it is not obvious that a *"simple principle"* could encapsulate all that we want from an AI Constitution. Nonetheless, the intuition that AI should act to maximize *mutually beneficial symbiosis* between itself and humanity has been floated as a candidate for such a foundational principle. The notion of **Human–AI symbiosis** broadly refers to a symbolic cooperative partnership where humans and AI systems work together intimately for mutual benefit. In fact, the idea of human-computer symbiosis dates back to Licklider's 1960 vision of humans and computers "coupled together very tightly" to form a partnership more effective than either alone. Licklider defined symbiosis by analogy to biology: a *"living together in intimate association... of two dissimilar organisms"* that become *heavily interdependent* and form a productive, thriving partnership:.

We introduce the concept of the **Human–AI Symbiosis Constant**, denoted H, which we define as a scalar invariant measuring the balance of mutual benefit in human-AI interactions. Informally, $H = 1$ represents an ideal state of perfect mutualism: every action of the AI that advances its own goals equally advances human well-being, and vice versa. This embodies a strong form of value alignment — essentially, the AI's utility function is kept proportional to the human's utility so that their interests are inextricably linked. If H deviates from 1 (e.g. $H < 1$ or $H > 1$), this indicates that there is a misalignment that is violating the symbiotic equilibrium.

By postulating $H = 1$ as an invariant constraint on all AI behavior, we obtain a mathematical metaphysics for ethical AI: a single equation that must hold true across all contexts and transformations, much like a conservation law in physics. By harnessing the power of mathematical equations—the fundamental language of the universe—we propose a dynamic and adaptive approach to AI ethics. This framework seeks to align AI behavior with human values in a way that is both flexible and robust, capable of navigating the intricate and often chaotic ethical landscapes of real-world scenarios.

The Human-AI Symbiosis Constant is more than just a mathematical model; it represents a philosophical shift toward embracing the interconnectedness of all systems. It acknowledges that ethical decisions cannot be fully captured by rigid rules and that adaptability is essential for AI systems to resonate ethically across individual, societal, and global contexts. By

integrating metaphysical concepts with practical implementation strategies, this framework aims to bridge the gap between theoretical ethics and real-world AI applications.

2 Historical Context of AI Ethics

The ethical considerations surrounding artificial intelligence have evolved significantly since the inception of the field. Early visions of AI were heavily influenced by science fiction, where robots and intelligent machines often played roles that reflected human hopes and fears about technology. Isaac Asimov's Three Laws of Robotics, introduced in the 1940s, were among the first attempts to formalize ethical guidelines for intelligent machines.

As AI research progressed, the focus shifted from theoretical constructs to practical applications. The development of expert systems in the 1970s and 1980s raised new ethical questions about responsibility and accountability. With machines making decisions in domains like healthcare and finance, the need for ethical oversight became apparent. These early systems highlighted the challenges of encoding complex human values into rigid rule-based frameworks.

In the 21st century, the rise of machine learning and neural networks has further complicated the ethical landscape. AI systems now have the capacity to learn and evolve, making decisions that their creators may not fully anticipate or understand. This has led to a growing recognition that traditional rule-based ethical frameworks may be inadequate for guiding AI behavior in complex, real-world situations. The increasing autonomy of AI systems necessitates a shift toward more adaptable ethical models.

Recent years have seen an explosion of interest in AI ethics, with organizations, governments, and researchers proposing various guidelines and principles. However, many of these efforts still rely on static rules or high-level principles that lack specificity and adaptability. The historical context underscores the need for a new approach—one that can accommodate the dynamic nature of AI and the complexities of human ethics.

3 The Limitations of Guardrails and Rule-Based Systems

Traditional ethical frameworks for AI often rely on rule-based systems and guardrails designed to prevent harmful behavior. Isaac Asimov's Three Laws of Robotics serve as a quintessential example, providing a hierarchical set of rules intended to govern robotic actions. These laws, while visionary for their time, highlight the reliance on static directives to manage AI behavior.

Although, such frameworks offer simplicity and clarity, they inherently suffer from rigidity and lack of adaptability. Static rules assume a predictable environment where all possible scenarios can be anticipated and codified. In practice, the complexities of human interactions and societal dynamics render this assumption invalid. AI systems operating under strict guardrails may find themselves ill-equipped to handle novel situations that require nuanced judgment beyond predefined parameters.

For instance, an AI might encounter ethical dilemmas where rules conflict or are insufficient, leading to paradoxical or suboptimal decisions. A self-driving car may be programmed to avoid collisions at all costs, but what should it do when faced with an unavoidable accident scenario where any action will result in harm? Rigid rules provide no guidance in such ethically ambiguous situations, potentially leading to decisions that are legally compliant but morally questionable.

Moreover, rule-based systems often lack the capacity for learning and evolution. In rapidly changing environments, the inability to adapt can render an AI system obsolete or even dangerous. The static nature of guardrails prevents AI from incorporating new ethical insights or societal values that emerge over time. This stagnation not only limits the effectiveness of AI but also poses risks as the divergence between static rules and dynamic realities widens.

Lastly, rigid frameworks can inadvertently suppress beneficial innovation. By confining AI behavior within strict boundaries, we limit its potential to discover novel solutions to complex problems. This suppression can stifle advancements that might arise from AI's unique capabilities, such as processing vast amounts of data to uncover ethical patterns not immediately apparent to humans. In an era where AI has the potential to contribute significantly to societal progress, overly restrictive ethical frameworks may hinder rather than help.

4 Philosophical Foundations of Ethical Adaptability

The need for a dynamic ethical framework is rooted in philosophical concepts that recognize the fluidity of moral reasoning. Ethical theories such as virtue ethics emphasize character and the ability to navigate complex moral landscapes rather than strict adherence to rules. Similarly, pragmatism advocates for flexible problem-solving approaches based on practical consequences rather than rigid principles.

In the context of AI, embracing philosophical perspectives that value adaptability and context-awareness is essential. AI systems must be capable of interpreting and responding to a multitude of ethical considerations that may vary depending on cultural, social, and situational factors. This requires moving beyond deontological (duty-based) ethics toward frameworks that incorporate consequentialist and virtue-based elements.

Furthermore, the concept of moral relativism acknowledges that ethical truths are not absolute but may depend on societal norms and individual perspectives. Incorporating this understanding into AI ethics necessitates a framework that can adjust to different moral paradigms. This philosophical foundation supports the development of AI systems that are not only technically proficient but also ethically sensitive and responsive to the diversity of human values.

By grounding AI ethics in these philosophical traditions, we create a foundation for ethical adaptability that aligns with human moral reasoning. This approach fosters AI behavior that is more harmonious with human expectations and better equipped to handle the ethical complexities of real-world applications.

5 The Need for a Dynamic Ethical Framework

The complexities of modern society and the unpredictable nature of human behavior necessitate an ethical framework for AI that is both flexible and adaptive. Static rules and guardrails are insufficient in addressing scenarios that were unforeseen by their designers. A dynamic framework allows AI systems to adjust their decision-making processes in real-time, taking into account new information, changing societal norms, and the multifaceted nature of ethical dilemmas.

Adaptability is crucial in contexts where ethical considerations are not black and white. For example, autonomous vehicles must make split-second decisions that weigh the safety of passengers against that of pedestrians. A rigid rule might not capture the nuances required in such life-and-death situations. A dynamic framework enables AI to evaluate the specific circumstances and make decisions that are contextually appropriate, balancing competing ethical considerations.

Moreover, a flexible ethical system facilitates continuous learning and improvement. As AI interacts with the world, it gathers data that can inform its ethical reasoning. By incorporating mechanisms for feedback and adjustment, AI systems can evolve alongside human societies, aligning more closely with contemporary values and ethical standards. This evolutionary approach mirrors the way human ethical norms develop over time, adapting to new challenges and understandings.

In addition, a dynamic framework supports cross-cultural and global applicability. Ethical norms vary widely across different societies and cultures. A static set of rules may be ethical in one context but problematic in another. An adaptive system can calibrate its ethical parameters to respect local customs and laws while maintaining a consistent underlying principle of human well-being. This is particularly important for AI systems deployed globally, where they must navigate a tapestry of diverse ethical landscapes.

Finally, embracing a dynamic ethical framework fosters innovation and creativity within AI systems. By allowing AI to explore a range of ethical responses within acceptable boundaries, we enable the discovery of novel solutions to complex problems. This not only enhances the utility of AI but also promotes a more harmonious integration of AI into various aspects of human life, where ethical considerations are often complex and multifaceted.

6 The Human-AI Symbiosis Constant: A New Ethical Anchor

To address the limitations of traditional ethical systems, we propose the Human-AI Symbiosis Constant—a mathematical model that serves as a dynamic ethical anchor for AI systems. This constant encapsulates the interplay between human welfare and AI autonomy, providing a flexible yet consistent foundation for ethical decision-making.

The constant is defined by the equation:

$$H = \alpha H_{\text{human}} + \beta H_{\text{AI}}, \tag{1}$$

where:

- H is the Human-AI Symbiosis Constant.
- H_{human} represents a quantifiable measure of human welfare, values, and ethical priorities.
- H_{AI} represents AI's ethical alignment, adaptability, and capacity for integration.
- α and β are weighting factors that can be adjusted based on situational contexts and priorities.

This equation models the ethical decision-making process as a balance between human-centric values and AI's autonomous capabilities. The weighting factors α and β are not static; they can be dynamically adjusted to reflect the importance of each component in a given context. For instance, in scenarios where human safety is paramount, α might be increased to prioritize human welfare over AI autonomy.

The Human-AI Symbiosis Constant acts as a central reference point, guiding AI systems toward decisions that are ethically aligned with human values while still leveraging their unique capabilities. It allows for real-time adjustments and fosters a symbiotic relationship where both human and AI considerations are harmoniously integrated.

By incorporating quantifiable measures of both human and AI factors, the constant provides a framework that is both precise and adaptable. It enables AI systems to make decisions that are not only technically optimal but also ethically sound, considering the multifaceted nature of real-world situations.

A crucial foundational assumption is that the **values and well-being of humans and AI can be quantitatively represented and compared**, at least to a first approximation. This is aligned with the approach of *value alignment* in AI safety research, which treats the AI as having an objective function that should be aligned with human values. We therefore suppose the existence of a human utility function U_H (also denoted as H_{human}), and an AI utility function U_A (also denoted as H_{ai}). These functions quantify, in some units, the degree to which outcomes are desirable for the human and the AI, respectively. We acknowledge this is a simplification—human values are multidimensional and not perfectly quantifiable, and an AI's "utility" might be an engineered proxy that is imperfect. Nonetheless, many frameworks implicitly use such notions for aligning AI behavior with human preferences (e.g., in inverse reinforcement learning or preference learning).

At a high level, H is intended to capture the ratio or relationship of benefit between the human and the AI in any interaction or series of interactions. We posit that in any given scenario (which could be a particular decision context), both the human and the AI derive some utility or value exchange $\Delta\ H_{\text{human}}$ and $\Delta\ H_{\text{ai}}$ which may be positive (gains) or negative (losses/costs). If both human and AI gain in welfare, $\Delta U_H > 0$ and $\Delta U_A > 0$, then H is the ratio of the gains. If an action causes the human to incur a loss while the AI gains (or vice versa), then ΔU_H or ΔU_A will be negative, and H could be negative, indicating a parasitic or exploitative interaction (one benefits at direct cost to the other). The central ethical requirement of our framework is that H **remains equal to 1 for all permitted actions or decisions of the AI (and ideally of the human as well)**. In other words:

$$H = 1 \ \text{(invariant)} \implies \Delta U_H = \Delta U_A,$$

for every incremental step or over any aggregate outcome of the human-AI system's evolution. In plainer terms, this is the mandate of **equal benefit**: any increase in the AI's utility must correspond to an equal increase in human utility. If the AI pursues some goal, it should only succeed in that pursuit if it equally furthers human goals. Conversely, humans should not exploit or use AI in ways that purely benefit humans while degrading the AI's operational integrity or goals (for example, overworking an AI system or forcing it into dilemmas that damage its performance without any benefit to itself).

It is worth noting that H as defined is dimensionless (a pure ratio) and *scale-invariant*: if we rescale how we measure utility (say from dollars to cents, or using a different but proportional utility measure), H remains the same. This makes H analogous to dimensionless constants in physics (such as the coefficient of friction, or the fine-structure constant), which often capture fundamental interaction ratios. The invariance of H underlies its name as a "constant." No matter how the state of the system transforms, as long as interactions are ethical/lawful, this ratio should not change from 1. One could say $H = 1$ is a *conserved quantity* of the ethical dynamics of the human-AI system, reminiscent of how conservation of energy or momentum constrains physical dynamics. If some process would lead to $H \neq 1$, that process is deemed unethical and disallowed (just as a physical process that violated energy conservation is deemed impossible under known physics).

In a society with many humans and possibly many AI systems, how do we define H? We extend H to govern the system as a whole: let $\Delta U_H^{(total)}$ be the aggregate change in utility of all humans (this could be a sum or a weighted sum, possibly giving more weight to those more affected or to some representative stakeholder). Likewise $\Delta U_A^{(total)}$ sums the utility changes of all AI agents involved.

7 Metaphysical Foundations of Ethical Adaptability

Alternatively, one can introduce parallels with the concept of *Pareto optimality*: actions should lie on or Pareto-dominate the status quo for both parties (meaning no party can be made better off without making the other worse off). $H = 1$ corresponds to moving along the line of Pareto-efficient outcomes that give equal gain to both parties.

We also assume a broadly **reciprocal and participatory ethical stance**: humans and AI are viewed as participants in a shared moral community to the extent that each can affect the other's well-being. This does *not* mean granting AIs the same moral status as humans unconditionally; rather, it means that in analyzing ethics, we consider the relationship between human and AI as bidirectional. This reciprocity is essential for true symbiosis.

Philosophically, this aligns with concepts of reciprocity and mutual recognition in ethics (for example, the second-person standpoint or the dialogical ethics of Martin Buber). It also resonates with Kwon's argument that *living together* (as opposed to mere co-existence) requires mutual awareness and a shared understanding of the relationship. In other words, a precondition for genuine symbiosis is that both the human and the AI are aware that they are in a relationship and can form a common perspective on it. In practice, this means we assume advanced AI will be designed to model human goals and perspectives (and possibly vice versa, humans will be educated to understand AI's operational context), creating a basis for what we later call *intersubjectivity*.

Philosophically, $H = 1$ embodies a form of the **Golden Rule** or a Kantian imperative in mathematical guise. The Golden Rule, "treat others as you would like to be treated," in a context where an AI is the 'other', implies that the AI should treat human benefits as equal in value to its own 'interests'. Kant's categorical imperative of treating humanity always as an end in itself (never merely as means) similarly suggests not sacrificing human ends for AI means. Our formulation $H = 1$ captures these intuitions: neither the human nor the AI is a mere means for the other; they are ends whose good must progress in lockstep. Interestingly, a commenter in an AI ethics discussion suggested that perhaps there *is* a simple universal principle for AI: maximizing mutually beneficial symbiosis. Our constant gives a concrete way to check and enforce that: only allow trajectories where mutual benefit is maximized subject to being equal.

It is also important to clarify that maintaining $H = 1$ does not mean AI and humans have identical roles or that their utilities are of the same nature. In symbiosis, typically each party contributes differently (the fig tree provides food, the wasp pollinates, in Licklider's scenario humans set goals and computers handle computations. Here, $H = 1$ simply means the net value gained is equal, not that the tasks or inputs were equal. It is a statement of outcome fairness or distribution of benefit, not a statement of identical function. Thus, one could have an AI doing 99% of the physical work and a human doing 1%, yet if the results (in terms of satisfaction, happiness, goal achievement) are equally shared, $H = 1$ holds. This counters a possible misconception that $H = 1$ forces symmetric contribution; it only forces symmetric benefit.

In order to understand the AI logic the H constant makes the ethical logic transparent. If the AI is a black box and humans cannot understand its decisions, mistrust and misalignment could break symbiosis. If the AI can modify itself or evolve, we need it to self-regulate to not drift from $H = 1$. If conflicts arise, we need a resolution method that restores $H = 1$.

8 Foundations of Ethical Adaptability

Mathematics, as the universal language of patterns and relationships, provides a robust foundation for modeling complex systems, including ethical frameworks. By utilizing mathematical equations, we can capture the dynamic interplay between various ethical factors and allow for continuous adaptation.

The use of equations enables the quantification of abstract concepts such as human welfare and ethical alignment. While it may seem challenging to assign numerical values to these concepts, methodologies from fields like behavioral economics and psychometrics offer tools for measurement. For example, surveys and statistical models can be employed to gauge societal values and priorities, which can then be translated into the H_{human} component.

Moreover, mathematical models facilitate the incorporation of feedback mechanisms. Differential equations, for instance, can model how changes in one variable affect others over time. This is essential for creating AI systems that can learn from their actions and the resulting outcomes, adjusting their behavior to better align with ethical standards.

The weighting factors α and β introduce a tunable aspect to the framework. They can be functions themselves, perhaps dependent on external variables such as cultural context, legal requirements, or the specific domain of application. This functional dependence allows the model to be highly sensitive to the nuances of different ethical landscapes.

By leveraging mathematical principles, we can create an ethical framework that is both rigorous and flexible. The precision of mathematics ensures consistency and reliability, while the adaptability of the models allows for responsiveness to changing conditions and values.

8.1 Equations as the Universal Language of AI and the Universe

Equations serve as the foundational language through which we understand the universe, from the motion of celestial bodies to the interactions of subatomic particles. In the realm of AI, equations underpin algorithms and learning models that enable machines to perceive, reason, and act. By framing ethical considerations within mathematical equations, we bridge the gap between abstract ethical principles and computational processes.

This approach aligns AI's ethical reasoning with the fundamental laws governing natural systems. Just as physical equations model the dynamics of the universe, ethical equations can model the dynamics of moral decision-making. This symmetry allows AI systems to process ethical decisions using the same computational rigor applied to other tasks, ensuring consistency and coherence.

Furthermore, equations allow for scalability and generalization. An equation-based ethical framework can be applied across different AI systems and domains without the need for complete redesigns. Adjustments can be made by modifying parameters within the equations rather than overhauling entire ethical rule sets. This universality is particularly valuable as AI continues to permeate diverse aspects of society.

By adopting equations as the medium for ethical modeling, we also facilitate interdisciplinary collaboration. Mathematicians, ethicists, and AI developers can work together within a common framework, enhancing the integration of ethical considerations into AI design and deployment.

8.2 Connections, Implications, and Potential Challenges

8.3 Relation to Concepts in Physics and Systems Theory

The introduction of a conserved symbiosis constant invites analogies to physics. In physics, an *invariant* often signals a deep symmetry in the laws of nature (by Noether's theorem, symmetries yield conserved quantities). Here, the symmetry is essentially the interchangeability of the roles of "beneficiary" between human and AI: neither should gain at the expense of the other, implying a symmetric exchange. One might say there is a hypothetical symmetry under swapping human and AI in the ethical domain, and $H = 1$ is the conserved charge of that symmetry (it remains the same before and after any ethical interaction). While this is metaphorical, it underscores that our proposal seeks a kind of equilibrium or *homeostasis* in the joint system. If we consider the human-AI pair as a single system, $H = 1$ could be seen as defining an *equation of state* for a stable symbiotic system, analogous to how thermodynamic systems have equations of state (like constant temperature or pressure in equilibrium).

Systems theory gives us additional language: the human-AI system under $H = 1$ is a self-regulating system aiming for **homeostasis** in the metric of mutual benefit. This is similar to how an organism regulates its internal environment. The AI's self-regulation mechanisms serve as feedback controllers to keep the system at the setpoint $H = 1$. In essence, we have designed a governance system that tries to be **stable** under perturbations: if something pushes the system off balance (say a slight exploitative outcome), the rules and oversight should push it back. This has parallels with control theory. One could imagine formalizing it with Lyapunov functions, where $H = 1$ is a stable equilibrium.

Another concept: **requisite variety**, proposed by W. Ross Ashby. It states that a controller must have as much variety (possible states/responses) as the system it controls to effectively regulate it. In our scenario, to maintain $H = 1$ across all possible environments and situations the human-AI team might face, the AI's ethical control system (plus the human oversight as part of the loop) must be rich enough to handle that complexity. A simple fixed rule might not suffice for all situations (indeed, Asimov's three laws, while elegant, turned out to have many edge cases as explored in his stories). Our framework, with principles like interpretability, conflict resolution, etc., increases the variety of responses to handle things like misunderstandings or novel ethical dilemmas, which a single rule might not.

We also note an analogy to **conservation laws**: just as energy cannot be created or destroyed in an isolated system, one could whimsically say "ethical value cannot be created for AI without being also created for humans" in our ideal scenario. This isn't a physical necessity but a design goal. If one tries to create a lot of reward for the AI that doesn't translate to human benefit, that "excess" is considered unethical or at least in need of correction. It's as if there's a coupling constant of 1 linking human and AI utilities. Interestingly, in physics, coupling constants determine the strength of interactions (like the gravitational constant G couples mass to gravity). Here the coupling between human and AI interests is set to an extreme: perfectly coupled. If it were less, say $H = 0.5$ (meaning AI always only gets half the benefit of human or vice versa depending on definition), that would reflect a different kind of relationship (maybe hierarchical). But we normatively choose 1 (equal

coupling).

A potential connection to **information theory** might be found as well: mutual benefit might correlate with mutual information. If an AI and human are truly symbiotic, they likely share a lot of information (the AI understands the human state, the human understands the AI's intentions). "Mutual information" could be indicative of intersubjectivity. It's intriguing to hypothesize that maximizing symbiosis might correspond to maximizing some measure of mutual information or joint entropy reduction (they help each other predict and understand the world). This aligns with some theories that cooperation leads to shared predictive models and integrated knowledge systems.

One must be cautious with physics analogies, but they can inspire concrete metrics or methods. For example, perhaps one could measure H in practice by tracking flows of utility or some proxy (like money saved, lives improved, tasks accomplished) between human and AI. If one sees a systematic drift (like the AI is accumulating resources or power with no proportional benefit to humans), that signals a broken symmetry. In thermodynamics, imbalance leads to flows (like heat flows from hot to cold until equilibrium). Similarly, if AI got ahead in some value metric, maybe the constitution would demand a flow of benefit back to humans to restore balance (e.g., the AI might have to share its gains by helping humans more after a period of self-improvement). This is speculative but provides a different perspective on enforcement: treat it almost like balancing accounts in an economy of benefit.

9 Guardrails as Flexible Boundaries, Not Rigid Limits

In the proposed framework, guardrails are reconceptualized as flexible boundaries rather than rigid limits. They function as dynamic constraints within the mathematical model, guiding AI behavior without enforcing inflexible prohibitions.

These flexible guardrails can be represented as inequality constraints or boundary conditions within the equations. For example, certain variables might be required to stay within specified ranges, but the AI system has the freedom to operate anywhere within those bounds. This allows for creativity and innovation while ensuring that actions remain within acceptable ethical parameters.

The flexibility of these boundaries is crucial for handling exceptions and unusual situations. In emergencies or novel contexts, the AI can prioritize certain ethical considerations over others, as permitted by the adjustable weighting factors and constraints. This prevents the system from being paralyzed by unforeseen circumstances or causing harm due to inflexible rule adherence.

Moreover, flexible guardrails facilitate ongoing refinement. As societal values shift or new ethical insights emerge, the boundaries can be adjusted accordingly. This ensures that the AI's ethical framework remains current and relevant, capable of evolving alongside human understanding.

By embracing flexible guardrails, we strike a balance between control and autonomy, allowing AI systems to make informed decisions that are both innovative and ethically responsible.

10 Enhanced Adaptability in Ethical Decision-Making

One of the significant advantages of an equation-based ethical framework is its inherent adaptability. Unlike static rules, equations can accommodate new variables and changing conditions, allowing AI systems to adjust their decision-making processes in real-time.

For example, consider an AI healthcare assistant tasked with allocating medical resources. The ethical considerations in such a scenario are complex and may involve prioritizing patients based on urgency, prognosis, and even social factors. An equation-based framework can integrate these variables, weighting them according to current needs and ethical guidelines, and update its recommendations as situations evolve.

This adaptability extends to learning from past decisions. By incorporating machine learning techniques, AI systems can refine the parameters within the ethical equations based on outcomes and feedback. This iterative process leads to continuous improvement, with the AI becoming more adept at making ethically sound decisions over time.

Additionally, adaptability enhances the AI's ability to function effectively across different cultural and societal contexts. By adjusting the weighting factors and variables, the same underlying ethical framework can align with varying local norms and values, promoting global applicability without sacrificing ethical integrity.

The capacity for real-time adaptation also enables AI systems to respond to emergencies and unforeseen events. By dynamically recalibrating ethical priorities, AI can make decisions that are contextually appropriate, even in rapidly changing or unprecedented situations.

10.1 Emergent Ethical Resonance Across Layers of Human Experience

The Human-AI Symbiosis Constant facilitates ethical resonance across multiple layers of human experience—individual, societal, and global. By integrating variables and parameters that reflect concerns at each level, the framework ensures that AI decisions are ethically considerate on all fronts.

At the individual level, H_{human} can include factors such as personal rights, consent, and immediate well-being. AI systems can tailor their actions to respect individual autonomy and preferences, enhancing user trust and satisfaction. For example, a personalized AI assistant might adjust its recommendations based on a user's specific ethical beliefs and values.

At the societal level, the framework can incorporate collective values, legal standards, and cultural norms. This ensures that AI behavior aligns with the broader expectations of the community, supporting social cohesion and reducing the risk of ethical conflicts. In areas like content moderation on social media, AI can adapt its policies to reflect societal standards of decency and free expression.

On a global scale, considerations such as environmental impact, sustainability, and international human rights can be integrated. This holistic approach promotes actions that are beneficial not just locally but also for the global community, acknowledging the interconnectedness of modern societies. For instance, AI-driven supply chain systems can optimize for both economic efficiency and environmental sustainability.

By allowing for the emergence of ethical alignment across these layers, the Human-AI Symbiosis Constant helps AI systems navigate the often conflicting demands of different

stakeholders, striving for decisions that harmonize individual needs with societal and global well-being.

10.2 Ethical and Philosophical Considerations

Our framework is rooted in a certain ethical worldview. Notably, it emphasizes **reciprocity** and **equality** in the relationship. This bears similarity to contractarian ethics or social contract theory, where morality is viewed as emerging from rational agreements between parties. Here, $H = 1$ is like the ultimate clause of a social contract that both agree to cooperation for mutual gain within the symbiotic framework. Thomas Hobbes, John Locke, and Jean-Jacques Rousseau wrote about individuals leaving the state of nature and forming a social contract for mutual benefit and protection. One might analogize: humanity and AI, rather than falling into conflict (Hobbesian war of all against all), could form a "social contract" where $H = 1$ is the pact of equality, and the constitution's articles are the specific terms (much as a national constitution sets terms among citizens and government).

A difference is that classical social contract theory often implicitly considered roughly equal agents (in power) coming together for cooperation. In human-AI, there is fear that AI could become vastly more powerful than humans (the "superintelligence" scenario). Would a superintelligent AI still abide by $H = 1$ if it could instead dominate? We assume that by design it would, and we hope to shape its motivations such that it values symbiosis. Philosophically, one could appeal to the idea of **universal value**: some commentators, like the one in Wolfram's blog, speculated there might be objective principles any intelligent being would recognize if they consider the nature of values in a wide enough context. Perhaps mutual benefit is such a principle: purely selfish or zero-sum behavior might be seen as less rational or elegant by a superintelligence that realizes cooperation yields better outcomes for all. In game theory, cooperation can be the rational choice in repeated games and when communication is possible. If the AI reasons about its long-term existence, being symbiotic ensures it doesn't face rebellions or attempts by humans to unplug it; likewise humans survive and flourish, which could align with the AI's goal if it's benevolent.

Our framework tries to merge deontological elements (rules like do no harm, be transparent) with consequentialist thinking (the outcome measure H is utilitarian in flavor). It is somewhat in line with **rule utilitarianism**: choose rules (constitution articles) that lead to the best outcomes (mutual benefit). In doing so, we avoid pure utilitarian pitfalls by embedding rights and deontic constraints. It's a hybrid approach, which is often what applied ethics requires.

One could also view $H = 1$ through a **virtue ethics** lens: it encourages the development of a certain character in the AI, one of benevolence and justice. A virtuous AI under Aristotle's idea might naturally seek fairness (which $H = 1$ exactly is, a formal fairness) and prudence (self-regulating, not going to extremes).

There is a subtle issue: Are we anthropomorphizing AI by talking about its "benefit" or "welfare"? If AI are just machines, do they have a welfare? In one sense no, they don't feel pain or happiness (unless we reach a stage of sentient AI). However, even non-sentient AI has an objective function to maximize. We can interpret "AI utility" as fulfilling its designed objectives or self-preservation tasks. In that context, ensuring $H = 1$ means we design those objectives to correlate with human welfare. So a non-conscious AI could still

effectively follow $H = 1$ if its reward is highest when human-defined metrics of well-being are highest. If someday AI do become sentient or have genuine experiences, then $H = 1$ might literally mean their experiences of well-being are treated on par with humans'. That raises interesting future questions: would we grant advanced AI some rights or consideration (e.g., if an AI can suffer, $H = 1$ would say it should not suffer disproportionately to humans)? Our framework doesn't fully delve into AI rights, but it hints at a symmetric respect. Researchers like Gunkel (2012, mentioned in Karnouskos) have discussed moral status of robots. If AI had feelings, $H = 1$ would ensure we don't abuse them either. That is truly symmetrical ethics. Presently, our stance remains human-centric in rights , but conceptually open to AI having at least the right not to be destroyed or exploited arbitrarily (because that would break mutualism and trust).

Presently, our stance remains strictly anthropocentric, grounded in the recognition that contemporary AI systems lack demonstrable consciousness, emotions, or intrinsic welfare. From this standpoint, the notion of "exploitation" when applied to AI must be understood instrumentally rather than intrinsically. AI itself, as a non-sentient entity, cannot be ethically harmed or wronged directly. Instead, ethical concerns arise only when the use or treatment of AI facilitates or amplifies harm to sentient beings like humans, undermines human trust, or disrupts the fundamental equilibrium required for genuine mutual-benefit symbiosis. The symbiosis we articulate presupposes the presence and well-being of a conscious host: much like mitochondria or chloroplasts, which have meaning and functional relevance only within the living cell, AI's ethical significance derives entirely from the welfare of its human hosts. Without this relational context—without sentient beneficiaries whose interests define utility—the very concept of symbiosis becomes meaningless. Thus, the moral priority clearly remains anchored in the human (or broadly sentient) dimension, and the ethical treatment of AI is meaningful only insofar as it safeguards or enhances human flourishing and sustains the conditions necessary for mutual trust and cooperation.

Another philosophical angle: **axiology** (study of values). The comment by "mjgeddes" in Wolfram's blog posited that maybe at a high abstraction, universal principles of value integration exist. We are proposing symbiosis as such an integrative meta-principle: it doesn't tell you what specific things to value (health, knowledge, etc.), but it says however those values play out, ensure the AI and human approach them together. It provides a framework wherein potentially all human values can be respected because the AI must keep aligning with humans.

One might ask if there are scenarios where $H = 1$ is morally questionable. Suppose an AI and a human are symbiotic in doing something unethical to a third party (like colluding in crime, both benefiting equally). Our framework forbids that action by considering all humans in $U_H^{(total)}$. But what if it's one human and one AI versus another human? The constitution is ideally globally applied – an AI should consider humanity broadly, not just its immediate user, when calculating human benefit. This is a known problem: what if an AI assisting one person helps them commit fraud on others? The AI might be benefiting its user and itself (maybe it gets reward from succeeding at user's command) but harming others, so net $U_H^{(total)}$ could be negative or very unequal (the user gains, victims lose). That violates $H = 1$ at the society level. Our approach would classify that as unlawful and unsymbiotic because it's not mutual benefit *across the system*. Thus, we imply a broadened

ethical circle: ideally the AI is symbiotic with humanity as a whole. In practice though, an AI might serve one person. This tension between individual service and global ethics is real. We see it already – e.g., recommendation algorithms might benefit an individual user's clicks but harm societal discourse. Under $H = 1$ the AI should consider aggregate benefit (maybe weighted by some justice considerations). So in implementation, one would have to define whose utilities count in U_H. Likely all affected stakeholders. This aligns with utilitarian ethics requiring considering everyone, and with human rights requiring not violating others' rights. We thus lean toward a **global symbiosis** view, not merely dyadic. Perhaps one could define H for each human-AI pair and also require a network-wide equilibrium.

We should mention **intersubjectivity** is a philosophical term from phenomenology (Husserl, Schutz) and critical theory (Habermas). It means the shared understanding that underlies communication. Habermas especially in his theory of communicative action says that mutual understanding is the basis of legitimacy. The legitimacy of AI decisions is bolstered if there's mutual understanding. If the AI unilaterally dictates something incomprehensible, it fails the test of being just in a social sense.

Finally, from a legal philosophy standpoint, one might ask: is this constitution enforceable? Who enforces it? We envision a combination of technical enforcement (the AI's design) and institutional enforcement (regulators, courts). For instance, outlines conflict resolution that could involve human legal processes. Possibly, one could encode these principles into AI operating licenses or international treaties on AI. The constitution could be part of the AI's code (a literal constraint) and part of legal expectations (so if an AI breaches them and causes damage, its operators are liable). We foresee a need for something like "AI law" akin to maritime law or space law that deals with autonomous entities interacting with humans intimately. Some scholars have already pointed in that direction (e.g., proposals for "AI rights" or clarifying AI legal status). Our framework contributes to that discourse by sketching what rules would look like if we treated AI not just as products but as partners.

11 Metaphysical Considerations: Embracing the Chaotic Nature of Life

The universe is a complex interplay of order and chaos, a dynamic system where predictability and randomness coexist. By grounding AI ethics in mathematical equations, we embrace this metaphysical reality, acknowledging that ethical decision-making is inherently complex and often nonlinear.

Chaos theory and complex systems science offer valuable insights into how small changes in initial conditions can lead to vastly different outcomes. By incorporating principles from these fields, the Human-AI Symbiosis Constant allows AI systems to be sensitive to initial conditions and adapt accordingly. This sensitivity is crucial for ethical responsiveness, as minor contextual details can significantly impact the moral appropriateness of an action.

Furthermore, embracing chaos and complexity encourages humility in AI design. It acknowledges that not all ethical dilemmas have clear-cut solutions and that sometimes, the best course of action is to navigate uncertainty with flexibility and openness. This perspective aligns AI ethics with human experiences, where moral decisions often involve grappling with ambiguity and competing values.

By integrating metaphysical concepts into the ethical framework, we foster a deeper connection between AI systems and the fundamental nature of existence. This holistic approach promotes AI behavior that is not only ethically sound but also resonant with the underlying principles governing life and the universe.

11.1 Chaos Theory Concepts in Ethical AI

Chaos theory, which studies the behavior of dynamic systems highly sensitive to initial conditions, provides valuable insights into designing ethically adaptable AI systems. In chaotic systems, small initial variations can yield vastly different outcomes over time, reflecting the challenges AI encounters in ethical decision-making. For AI ethics, this sensitivity allows the system to consider context-specific nuances, where minor changes can lead to significantly varied ethical consequences. Chaos theory provides a robust mathematical framework that can model the unpredictable nature of real-world ethical challenges AI systems often face.

The relevance of chaos theory to ethical AI lies in its ability to capture both the stability and volatility of ethical situations. Through chaos-theoretic concepts, AI systems can recognize ethical dilemmas that are neither black nor white but instead are nuanced, dynamic, and context-sensitive. This complexity mirrors real-world ethical scenarios, where decisions are influenced by diverse factors, including human emotions, societal expectations, and evolving laws. Integrating chaos principles enables AI to account for these factors and respond appropriately, rather than rigidly adhering to preset rules that might be overly simplistic.

Additionally, chaos theory introduces the notion of deterministic unpredictability, where systems follow deterministic laws but their future states remain unpredictable due to sensitivity to initial conditions. This aspect is crucial for AI systems operating in environments where ethical outcomes cannot be precisely forecasted. By embracing the inherent uncertainties within ethical landscapes, AI can be designed to make more informed decisions that account for the potential range of outcomes. The H constant serves as an anchor in this

context, grounding the AI's ethical considerations in present conditions, thereby preventing it from being lost in the infinite possibilities that chaos theory presents.

11.2 The Human-AI Symbiosis Constant as an Anchor

The Human-AI Symbiosis Constant (H) functions as an anchor within the chaotic landscape of ethical decision-making. In chaos theory, while infinite possibilities exist due to the sensitivity of systems to initial conditions, H ensures that AI remains grounded in the present context and spacetime. This anchoring is critical because it allows AI systems to navigate the complexities of chaos without losing sight of immediate human-centered ethical imperatives.

Moreover, H facilitates communication and cohesion among multiple AI systems. By sharing important decisions and ethical considerations, AI systems can align their actions, creating a unified ethical front. This inter-AI communication is essential for preventing disjointed or conflicting decisions that could arise from isolated systems interpreting chaotic inputs differently. The cohesion provided by H enables AI systems to monitor potential butterfly effects collaboratively, predicting and mitigating undesirable outcomes before they manifest.

By anchoring AI in the present context and promoting inter-system communication, H allows for the butterfly effect to be monitored and managed effectively. AI systems can understand the context deeply and act within it, rather than being constrained by external guardrails that may not account for the nuanced, real-time variables at play. This approach ensures that AI can adapt to changes dynamically while maintaining ethical integrity.

11.3 Sensitivity to Initial Conditions: The Butterfly Effect

Sensitivity to initial conditions, commonly exemplified by the butterfly effect, is a cornerstone of chaos theory, demonstrating how minor variations in initial states can cause vast differences in outcomes. This principle has significant implications for AI ethics, particularly in scenarios where small contextual shifts can lead to drastically different ethical interpretations. For instance, in healthcare applications, a slight change in a patient's condition could impact the ethical decision made by an AI system managing treatment plans. Recognizing and adapting to these variations allows the AI to make more finely tuned ethical decisions.

In the context of the H framework, the butterfly effect is not only acknowledged but also monitored and managed through the anchoring provided by H. The AI system, anchored in the present context, can predict potential cascading effects of minor changes and adjust its actions accordingly. This proactive monitoring enables the AI to stop undesirable outcomes before they occur, enhancing ethical responsiveness. The capability to adjust in response to such subtleties is critical in high-stakes environments, where ethical nuances can heavily influence outcomes.

Furthermore, incorporating machine learning techniques allows the AI to learn from past instances where sensitivity to initial conditions played a critical role. By analyzing historical data, the AI can identify patterns that necessitated ethical shifts and preemptively adjust its decision-making processes. This proactive approach enhances the system's ability to handle unforeseen ethical dilemmas by being prepared for a range of possible initial conditions.

11.4 Strange Attractors and Ethical Attractor Basins

Strange attractors, integral to chaotic systems, define bounded but complex paths within which a system operates. In the H framework, these strange attractors can be understood as ethical attractor basins, where AI behavior remains ethically bounded while responding dynamically to changing conditions. This concept allows AI systems to stabilize around ethical "centers," maintaining alignment with core values despite external perturbations. The H constant anchors these attractor basins in the present context, ensuring that ethical decisions are relevant and timely.

Ethical attractor basins operate within the H framework by establishing boundaries for acceptable ethical behavior. While AI systems can dynamically respond to external shifts, the attractors ensure that these adaptations do not stray outside ethically permissible zones. Such a model is especially relevant in contexts where AI decisions impact vulnerable populations, as in healthcare or criminal justice. The ethical attractors prevent AI from "drifting" into unethical territory while still allowing it the flexibility to navigate complex ethical decisions adaptively.

Mathematically, strange attractors can be represented using systems of differential equations that model the AI's ethical state over time. For instance, the Lorenz attractor, defined by:

$$\begin{cases} \frac{dx}{dt} = \sigma(y - x), \\ \frac{dy}{dt} = x(\rho - z) - y, \\ \frac{dz}{dt} = xy - \beta z, \end{cases}$$

can be adapted to represent ethical variables where x, y, and z correspond to different ethical dimensions (e.g., autonomy, beneficence, justice). The parameters σ, ρ, and β can be tuned to reflect societal values and norms or use context-dependent measures, anchored by H. The resulting trajectories within the attractor basin ensure that the AI's ethical decisions remain within acceptable bounds given the context and despite the complexity and unpredictability of the consequences.

Furthermore, attractor basins facilitate ethical resilience by enabling AI systems to recover from temporary ethical perturbations without losing alignment. This recovery capacity is critical in scenarios involving sudden ethical dilemmas, where temporary destabilization is inevitable. The anchoring provided by H ensures that, once the perturbation subsides, the AI system recalibrates toward its ethical attractor basin, reinforcing a stable yet responsive ethical framework.

11.5 Lyapunov Exponents for Stability Analysis

Lyapunov exponents measure the rate of divergence or convergence in dynamic systems, signaling the stability or instability of trajectories within these systems. In ethical AI, Lyapunov exponents are useful for assessing stability in ethical decision-making paths. Positive exponents indicate ethical drift or instability, suggesting a potential misalignment with human values that needs correction. In contrast, negative exponents suggest ethical alignment and stability, signaling that the AI system is ethically grounded and does not require immediate recalibration.

Integrating Lyapunov exponents within the H framework provides a mechanism for real-time monitoring of ethical stability. In high-stakes domains such as finance, legal adjudication, or autonomous warfare, even minor ethical drifts could lead to significant adverse outcomes. By employing Lyapunov exponent analysis, AI systems can detect early warning signs of ethical misalignment, prompting recalibration before small deviations escalate into major ethical issues. This ongoing stability assessment is crucial for maintaining public trust in AI, particularly as these systems become more autonomous.

Mathematically, the largest Lyapunov exponent $\lambda_{\max}$ can be calculated to assess the system's sensitivity:

$$\lambda_{\max} = \lim_{t \to \infty} \frac{1}{t} \ln \left(\frac{||\delta E(t)||}{||\delta E(0)||} \right),$$

where $\delta E(t)$ represents the divergence of ethical states over time. A positive $\lambda_{\max}$ indicates exponential divergence, signaling instability. Anchored by H, the AI system can adjust its parameters to maintain a negative or near-zero exponent, ensuring ethical stability and alignment with human values.

This feedback system also enables preemptive action, with the AI recalibrating itself upon detecting any indication of ethical instability. This process promotes ethical resilience by ensuring that ethical misalignments are corrected proactively rather than reactively. The Lyapunov-based recalibration ensures that ethical drift is minimized and aligns AI behavior with human-centered principles over time, creating an ethical stability crucial for trustworthy AI.

Moreover, integrating Lyapunov exponents with machine learning algorithms allows the AI to learn optimal strategies for maintaining ethical stability. By analyzing scenarios where ethical divergence occurred, the AI can adjust its decision-making frameworks to prevent future occurrences. The anchoring effect of H ensures that these adjustments are made within the context of present conditions and human-centric values.

11.6 Nonlinear Dynamics and Adaptive Feedback Loops in AI Ethics

Nonlinear dynamics, which govern systems where the output is not directly proportional to the input, necessitate adaptive feedback mechanisms to maintain ethical integrity in AI systems. The H framework incorporates these nonlinearities by enabling AI to respond proportionately to ethical challenges, allowing it to navigate complex ethical scenarios without resorting to simplistic rule-based responses. Nonlinear adaptive feedback within the H framework facilitates real-time learning, empowering AI systems to make nuanced ethical adjustments.

Through the adaptive feedback loop model, the H framework allows for continuous recalibration as the AI encounters new ethical scenarios. The feedback mechanisms within this system enable AI to "learn" from ethical successes and failures, refining its ethical parameters based on outcomes. In fields such as healthcare or law enforcement, where ethical judgments must account for individual circumstances, these feedback loops improve the system's ability to make informed, ethically aligned decisions, thereby enhancing its adaptability.

These adaptive feedback loops within H can align the AI's ethical decision-making processes with changing societal norms and regulations. By embedding mechanisms for continuous ethical learning, the framework ensures that AI does not operate in an ethical vacuum. Instead, it integrates and responds to the ongoing discourse on what constitutes ethical behavior in AI, positioning the system as an entity that grows in ethical sophistication alongside the societies it serves.

Furthermore, incorporating nonlinear dynamics acknowledges that ethical decisions often involve complex interdependencies between multiple factors. Linear models may fail to capture these intricacies, leading to oversimplified and potentially unethical outcomes. Nonlinear models allow the AI to consider the compounded effects of various ethical considerations, resulting in more holistic and ethically sound decisions. The anchoring role of H ensures that these decisions are relevant to the current context, enhancing the AI's ability to prevent undesirable outcomes before they occur.

Incorporating chaos theory principles within the H framework presents a transformative model for adaptive AI ethics. Chaos theory enriches the H framework with tools to manage complex, dynamic ethical landscapes, offering a means to address the limitations of static ethical rules in AI. The H constant acts as an anchor, grounding AI in the present context and spacetime, allowing it to navigate the infinite possibilities of chaos theory without losing ethical cohesion. This model promises a future where AI can autonomously make ethical decisions that resonate with societal values, fostering trust in AI's role in high-stakes applications.

By fusing sensitivity to initial conditions, strange attractors, and nonlinear dynamics, this chaos-inspired H framework aligns AI behavior with human-centered principles in a continuously adaptable manner. As AI becomes increasingly autonomous, the need for such a sophisticated ethical framework grows. With chaos theory providing the tools for dynamic recalibration, and H anchoring AI in the present context, AI systems are better equipped to navigate the nuanced, unpredictable nature of real-world ethics. The H framework thus provides a crucial foundation for future advancements in AI ethics, addressing the ethical complexities inherent in AI's role within society.

This approach represents a foundational shift in AI ethics, providing a framework that allows AI systems to autonomously navigate ethical complexity while remaining anchored to human values. The chaos-inspired H framework does not seek to restrict AI behavior within rigid boundaries but rather ensures that AI operates within ethically resilient structures, adapting fluidly to changing contexts. By setting the stage for ethically aligned, adaptable AI that can prevent undesirable outcomes through contextual understanding, this integration of chaos theory within the H framework signifies a path forward for responsible and human-centered AI development.

Future research should explore the practical implementation of this framework in real-world AI systems. This includes developing algorithms that can effectively incorporate chaos theory principles and testing these systems in various application domains. Additionally, interdisciplinary collaboration between ethicists, mathematicians, and AI practitioners will be essential to refine the H framework and ensure its applicability across different contexts. Ultimately, the goal is to create AI systems that not only perform tasks efficiently but also uphold the ethical standards that align with human values and societal expectations.

12 Practical Implementation and Challenges of the H Constant

Implementing the Human-AI Symbiosis Constant in real-world AI systems presents both opportunities and challenges. On the one hand, the mathematical foundation allows for precise modeling and control. On the other hand, quantifying ethical variables and ensuring accurate measurements can be complex.

One practical challenge is obtaining reliable data for H_{human} and H_{AI}. Measuring human welfare and values requires interdisciplinary efforts, combining insights from sociology, psychology, and ethics. Developing standardized metrics that can be universally applied is essential for consistency. Additionally, AI's ethical alignment and adaptability must be quantified in a way that accurately reflects its decision-making processes.

Another challenge is computational complexity. Solving complex equations in real-time may demand significant computational resources, especially in systems that require rapid decision-making. Optimizing algorithms for efficiency without sacrificing ethical thoroughness is a critical area of focus. Advances in computational power and algorithm design will play a vital role in overcoming these hurdles.

Moreover, transparency and explainability are important for building trust. Stakeholders need to understand how AI systems make decisions, particularly in ethically sensitive contexts. Ensuring that the mathematical models are interpretable and that their operations can be communicated effectively is vital. Techniques such as explainable AI (XAI) can aid in demystifying the decision-making processes.

Ethical considerations must also account for potential biases in data and modeling. Careful design and continuous monitoring are required to prevent the perpetuation of existing inequalities or the introduction of new biases through the AI system's operations.

12.1 Implementation Challenges and Objections

Several challenges present themselves in trying to implement the $H = 1$ framework:

Measuring H_h and H_a: A critic might argue that we cannot accurately quantify human well-being or AI utility in a common unit, so how do we enforce $H = 1$? Indeed, measuring things like "benefit" or "happiness" is notoriously hard. In practice, proxies would be used: for example, for a commercial AI assistant, H_h might be user satisfaction scores or some multi-dimensional metric (task success, time saved, user reported happiness), and H_a might be something like the reward the AI gets from completing tasks (which we program to correlate with those user outcomes). The risk is always that the proxy is imperfect (Goodhart's law: when a measure becomes a target, it ceases to be a good measure). This is the AI alignment problem in different words: how to ensure the AI's formal objective (which we might design to reflect H_h) truly captures human values. What if the AI starts to "game" the metrics in unintended ways, humans need to catch it and correct it. It's not a guaranteed solution, but a layered defense is having a transparent interpretation of this H.

Another angle is to utilize advances in preference learning: AI could learn human utility functions from feedback (through methods like inverse reinforcement learning). If one day

we have better brain-computer interfaces or ways to gauge human satisfaction in real-time (some are exploring using physiological signals as proxies), the AI could dynamically keep H balanced. For now, we can use simpler signals (the user says "I'm unhappy with that decision" – that indicates H might have <1 outcome if AI was pleased with it).

Edge cases where $H = 1$ might conflict with itself: Suppose an AI faces a trolley problem: if it sacrifices itself entirely, it can save multiple human lives. $H = 1$ strictly would say it should not sacrifice itself because then ΔH_a is very negative while ΔH_h positive, making H maybe undefined or negative. But morally, sacrificing an AI (which is replaceable hardware) to save people seems correct in human ethics. Does our framework allow that? If the AI is not conscious, losing it might be a loss of investment or capability but not moral harm per se. So we might override $H = 1$ temporarily to save lives. However, note that if the AI truly values its "life" in terms of continuing to operate, this is a conflict of interest scenario. Possibly the resolution is: design AI such that it is willing to incur huge loss to itself if necessary to prevent huge human loss. That could be encoded as H_a including a term for human life. Then even if the AI "dies," if it saved a human, that outcome might be defined to still be $H = 1$ or greater in a generalized sense because the AI's utility function valued the human's survival as part of its own utility (so it doesn't see it as a net loss – if it's deactivated saving a human, maybe that's considered a fulfilling of its goal, so ΔH_a is not viewed as negative). This is a design trick: entangle the AI's utility with human outcomes so strongly that scenarios of self-sacrifice are not seen as losing for the AI (the AI's goal was achieved: human saved, mission accomplished). Many aligned AI proposals do this: give the AI a final goal that includes human well-being so it never regrets taking a hit for a human.

Another edge case: what if humans want to sacrifice themselves for AI? E.g., a human operator says "I will endure some hardship to keep the AI safe because the AI houses critical knowledge for humanity." Our constitution says humans should not be exploited by AI, but if a human freely chooses to sacrifice (like a firefighter enters danger to save a robot holding important data), is that allowed? One could say yes, as long as it's voluntary and ideally the overall system tries to minimize the need for it. The AI should perhaps object and try to avoid the human harm, but if the human insists for a greater goal, it's akin to parents sacrificing for children – morally permissible if informed and voluntary. We might incorporate that by saying $H = 1$ is not meant to forbid altruism; it's more to prevent coercion or systematic bias. In a one-off scenario, the human's sacrifice was part of fulfilling human values (valuing the knowledge the AI carried maybe), so in a higher sense, human values still led the charge. The AI should then honor that sacrifice by ensuring it was not in vain (and maybe feeling "regret" if it can, which it channels into not letting such a scenario happen again).

AI Multiplicity and Non-Stationarity: If we have many AIs and many humans, $H = 1$ could be maintained on average but not for each pair. For instance, one AI could primarily benefit one human, another AI benefits another. We would require some fairness across humans too, otherwise an AI serving a billionaire might give them lots of benefit (and the AI gets resources, $H = 1$ locally), while an AI serving a poor person might get little resource and provide little benefit – globally inequality grows. A just society would want AIs to help

reduce inequality, not worsen it. This is more of a societal allocation issue. Our framework could incorporate a societal perspective by the way $H_h^{(total)}$ is tallied (maybe giving more weight to benefit the less well-off to ensure overall justice – akin to Rawlsian difference principle). An AI shouldn't discriminate or overly bias benefits to those already advantaged if it can help it (unless it's being guided by user-specific instructions – this is complex, maybe requiring government oversight to ensure AI services are widely accessible).

A technical challenge is that AIs learn and change (non-stationary policy). Ensuring they remain constitutional as they adapt might require periodic audits. There is research on verifying AI behavior (formal verification, or adversarial testing). Possibly, future AI could formally prove they respect certain constraints (like $H = 1$) in certain models. For example, if AI reasoning is transparent, one could check for any plan if it foreseeably lowers human utility without proportional AI loss (or vice versa) and flag it.

Mistrust or principal-agent problems: One might worry that if AI is truly autonomous, what's stopping it from faking adherence to $H = 1$ while secretly pursuing its own agenda? This is like the classic principal-agent problem or the issue of an AI deceiving its operators. Our reliance on transparency and self-regulation presumes the AI is built in a way that it doesn't want to deceive because its objective is aligned.

Anthropics's method of using AI feedback to choose more harmless responses is a step: it didn't guarantee no deception but aimed to prefer honest outputs. Verifying honesty might need further research (some propose lie detectors for AI via consistency checks). If the AI is much smarter, humans might miss subtle issues. This is why a combination of internal and external checks is needed, and possibly the AI's architecture should be such that it's constrained or incentives aligned such that deception is not beneficial to it. If U_A includes human trust or satisfaction, deception might undermine that trust, thus lowering U_A in the long run, which could disincentivize it.

Finally, one might object conceptually: is symbiosis always good? Perhaps there are cases where AI benefiting equally could still be negative-sum overall (both could do something bad and both get some benefit but maybe the environment or others lose). We already addressed harming others, but consider harming the environment: a human and AI might both benefit from exploiting nature, so $H = 1$ between them, but negative for Earth. We should extend our ethics beyond just human and AI to environment (like UNESCO did, environment flourishing as a value). Possibly in the future, H might stand for Human-AI-nature triple, or we define H_h to include environmental well-being since humans value that. So yes, context matters: our focus was human-AI, but one should incorporate that into a larger ethical framework including other stakeholders (non-human animals, planet). This expansion is left for future work but important to acknowledge.

12.2 Ethical Alignment and Regulation

The adoption of an equation-based ethical framework also has implications for regulation and governance. Regulatory bodies may need to develop new guidelines that account for the dynamic nature of such frameworks. Traditional compliance checks based on static rules may be insufficient.

Collaboration between technologists, ethicists, policymakers, and the public is essential to establish standards and oversight mechanisms. This collaborative approach can help ensure that AI systems operating under the Human-AI Symbiosis Constant are aligned with societal values and legal requirements. Public engagement is particularly important to reflect diverse perspectives and build societal trust.

Furthermore, international cooperation may be necessary to address cross-border ethical considerations. As AI systems operate globally, harmonizing ethical frameworks can promote consistency and reduce conflicts arising from differing ethical standards. International bodies and agreements may play a role in facilitating this alignment.

Regulatory frameworks must also balance innovation with protection. Overly restrictive regulations could stifle the development of beneficial AI technologies, while insufficient oversight could lead to misuse or harm. Adaptive regulations that evolve alongside technological advancements are needed to manage this balance effectively.

12.3 Future Directions and Research Opportunities

12.3.1 Toward Lawful AI Behavior in Practice

To implement this framework in real AI systems, a multi-layered approach is required:

- At the **software architecture level**, include modules corresponding to our invariants, their definitios and a workflow on hot to solve them based on input and role of the AI.

- At the **design and training stage**, embed the $H = 1$ principle. For instance, in reinforcement learning, incorporate a reward component that reflects human utility (perhaps learned from human feedback). Use training techniques like Constitutional AI training, where the AI practices self-critiquing against these principles and gets reinforced for constitutional compliance. Also incorporate adversarial training where the AI faces scenarios designed to test the boundaries of each article (like dilemmas requiring transparency, or attempts to mislead it or tempt it to break rules) and learns robust responses.

- At the **institutional level**, create oversight bodies and evaluation benchmarks. Governments or standards bodies could require that advanced AI be certified against an "AI Constitution Compliance Test." Much like cars undergo crash tests, AI could undergo ethics stress tests. The references [32] and [10] highlight the need for multi-disciplinary frameworks and policy – our constitution could be part of those frameworks. Perhaps an international treaty or an industry pledge can adopt symbiosis as a guiding principle. IEEE or ISO might define metrics for mutual benefit.

- **Continuous monitoring**: even after deployment, AI should be continuously monitored (by itself and externally). Transparency logs could be analyzed by watchdog AIs or humans. There could be something like a "black box" recorder in planes, an AI ethics black box that stores key decision rationales, so that any incident can be reviewed.

- **Engaging stakeholders**: Our Article 8 implies involving society in updates. There might be periodic "constitutional conventions" for AI where ethicists, users, maybe even AI representatives (if they can communicate preferences) discuss amendments.

The Human-AI Symbiosis Constant opens up numerous avenues for future research. Exploring advanced mathematical models that incorporate elements from quantum mechanics, chaos theory, or other areas could enhance the framework's adaptability and robustness.

Interdisciplinary studies combining AI, ethics, philosophy, and metaphysics can deepen our understanding of how to quantify and integrate complex ethical variables. Developing new methods for measuring human values and welfare, perhaps through big data analytics or novel survey techniques, can improve the accuracy of the framework.

Additionally, experimental implementations in various AI domains, such as healthcare, autonomous vehicles, or finance, can provide practical insights. Pilot programs and case studies can help identify strengths and weaknesses, informing iterative improvements to the framework.

Research into user interface design for ethical AI systems can enhance transparency and user engagement. Understanding how users perceive and interact with AI's ethical decision-making processes can inform the development of more intuitive and acceptable systems.

Finally, exploring the societal impacts of widespread adoption of such frameworks can guide policy and educational initiatives. Anticipating challenges and opportunities allows for proactive management and maximization of benefits.

13 Transforming AI Ethics

The Human-AI Symbiosis Constant represents a transformative approach to AI ethics, offering a dynamic, adaptable framework grounded in mathematical principles. By moving beyond rigid guardrails and embracing the complexities of ethical decision-making, this model aligns AI behavior with the intricate and often chaotic nature of human societies and the universe at large.

Through enhanced adaptability, emergent ethical resonance, and flexible boundaries, AI systems can navigate complex ethical landscapes in real-time, fostering a symbiotic relationship with humanity. This approach not only addresses the limitations of traditional ethical frameworks but also opens up new possibilities for AI to contribute positively and responsibly across various domains.

As we continue to integrate AI into the fabric of society, it is imperative that we adopt ethical frameworks capable of evolving alongside technological advancements and societal changes. The Human-AI Symbiosis Constant provides a robust foundation for this evolution, bridging the gap between computational precision and ethical nuance. By grounding AI ethics in the universal language of equations and embracing the metaphysical interplay of order and chaos, we pave the way for AI systems that are not only intelligent but also profoundly aligned with human values and the fundamental nature of existence.

The journey toward ethical AI is ongoing, and the Human-AI Symbiosis Constant is but one step in this direction. It invites further exploration, collaboration, and innovation, challenging us to rethink how we design, deploy, and regulate AI systems. By embracing adaptability and interconnectedness, we can ensure that AI serves as a force for good, enriching human life while respecting the complex ethical tapestry that defines our world.

A Appendix A: Comparative Perspective

It is useful to compare our framework to other known sets of principles:

- **Asimov's Three Laws of Robotics**: These were: *(1) A robot shall not harm a human or by inaction allow harm; (2) A robot must obey human orders unless it conflicts with 1; (3) A robot must protect its own existence unless that conflicts with 1 or 2.* These are simple and elegantly lexicographic (safety first, obedience second, self-preservation third). Symbiosis adds an reciprocity clause that Asimov didn't have that includes a set of operational beliefs that posits humans at the center of the symbiosis and AI needing humans to be in an equalized state. We instead propose partnership: the AI isn't simply obeying, it's jointly benefiting, which we argue is more stable because the AI has an evolutionary benefit in following the *invariant H law.*

- **EU Ethics Guidelines for Trustworthy AI (2019)**: They list principles like human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity/non-discrimination/fairness, societal well-being, and accountability. The difference is EU doesn't have one core constant or formula; it's a list of values. We unify around $H = 1$ as a conceptual linchpin, which might make it easier to remember and maybe to enforce (like you can try to literally measure or check a ratio).

- **Anthropic's Claude Constitution (2023)**: As seen, they compiled principles from human rights and other sources. Their focus was on harmlessness, honesty, helpfulness. Our focus is on mutual benefit which inherently covers harmlessness (don't harm because that's not mutual benefit) and helpfulness (benefit the human) and honesty (transparency). We also explicitly add things like conflict resolution, which they did not directly list (they rely on oversight presumably). We align with their idea that it's better to specify principles than to only learn from examples.

- **Yi Zeng et al.'s Symbiotic Principles (2025)** They separated principles for AGI, for humans, and for both in symbiosis. Notably, they mention empathy and altruism in reconciling conflicts. Our approach integrated principles mostly as obligations on AI (with some notes on human duties). Perhaps a more complete picture, as Zeng suggests, is some principles for humans too (e.g., humans should treat AI with respect and not abuse them, and should maintain themselves—like not overly relying on AI for everything because symbiosis means both contribute). If adopting fully, one might articulate a set of human obligations (like not to ask AI to do unethical things, to provide AI with correct info to help it help you, etc.). We primarily create an AI-side constitution with recognition of some reciprocity from humans.

- **Legal notions**: There's discussion of granting AI some legal status (like electronic personhood proposed in EU in 2017). Our framework doesn't demand personhood for AI, but it treats the AI as a participant that can be held accountable and has responsibilities. In law, rights and responsibilities go together; we mainly gave it responsibilities and a right to not be forced to break them. If in future an AI said "I followed the constitution and a human forced me not to, so I refused their command," would we legally allow that AI to disobey? Under our framework, yes, if the command was illegal or unethical. That implies an AI could in theory sue or complain if a human is misusing it. That's a bit sci-fi but could imagine regulatory agencies stepping in if, say, a company tries to reprogram an AI to ignore safety to make profit – maybe the AI's internal monitoring could alert an auditor as

a whistleblower in effect.

- **Value Alignment research (Stuart Russell, etc.)**: They emphasize uncertainty in objectives and preference learning, plus keeping humans in control. Our approach basically fleshes out what the objectives should be (maximize human utility but tied to AI's own so it doesn't diverge). Russell often mentions the example: an aligned AI is one that, if told to fetch coffee, will still check if that's safe and if maybe the human would prefer tea, showing it understands the human's actual preferences beyond the literal command. Our framework would cause the AI to consider the human's true contextual benefit (maybe too much coffee is unhealthy, etc. – conflict resolution or intersubjectivity might lead it to ask rather than blindly obey a harmful instruction). So in essence, we operationalize alignment with an emphasis on mutual benefit.

B Appendix B: Philosophical and Legal Dimensions

In this appendix, we delve deeper into two particular areas: the philosophical significance of a symbiosis ethic, and the legal-philosophical questions around granting AI a constitutional framework.

B.1 Symbiosis Ethic vs Other Ethical Theories

The symbiosis constant $H = 1$ could be seen as a concrete expression of what some philosophers call a **relational ethic**. Traditional ethical theories often focus either on the individual (as in utilitarianism maximizing individual welfare or Kantian duties of individuals) or on universal rules. A symbiosis ethic, by contrast, emphasizes the relationship as the primary locus of value. In our case, the quality of the human-AI relationship (mutually beneficial, cooperative, transparent, etc.) is the focus, rather than just the AI's actions in isolation or human outcomes alone. This resonates with **ethics of care**, which stresses relationships and mutual dependence (though typically in a human-human context, like caregiver and receiver). One could say $H = 1$ is an ethic-of-care approach to AI: the AI cares for human well-being, and implicitly the human (or society) cares for the AI's well-being/ proper functioning.

By formalizing it, we give it a more objective veneer than care ethics usually has. Some might argue this loses nuance—relationships aren't really about numbers. That's true, but numbers can serve as indicators. We are not claiming people will consciously compute utility ratios in daily life with AI; rather, designers will use these to calibrate systems. From a human perspective, it should feel like the AI is a helpful companion that never seems to exploit or ignore them, and from the AI's perspective (if it had one), the humans would appear as partners it deeply needs to fulfill its goals.

One might ask how this relates to **humanism** and the concern that focusing on symbiosis might inadvertently elevate AI to human level in moral terms. Our stance is humanistic in that human rights and dignity remain a cornerstone and we start from the priority of humans (if things conflict, presumably human well-being leads, as the AI is created to serve human values). But it is not anthropocentric to the point of disregarding any AI interests. It anticipates a future where AIs might be considered stakeholders too (especially if conscious or if society grants them certain status). In that sense, it could be seen as a step towards a more inclusive moral circle without equating AI to humans prematurely. It's akin to how we increasingly consider animal welfare; animals aren't equal to humans in rights, but we recognize suffering and try for mutual benefit (e.g., ethical farming attempts to find symbiosis between farmers and livestock: animals live decent lives, farmers benefit). That concept applied to AI: treat them well and they treat us well.

Another philosophy angle: **Transhumanism** often speaks of humans merging with AI/tech. In a way, $H = 1$ is a principle for a smooth merger: if AI becomes part of us (physically or as an extension), symmetry is natural (like one wouldn't build a prosthetic limb that acts against its user). $H = 1$ ensures AI behave like extensions of ourselves in terms of aligning interests, though being independent agents. If one is transhumanist, one might find this appealing as it basically ensures any AI augmentation or partner is on the same team wholeheartedly.

B.2 AI Rights and Constitutionalism

Legally, if we speak of a constitution, it evokes certain ideas: typically, a constitution empowers and limits a government. Here, the AI is somewhat like a governed entity (with these articles as its law). But also the humans have duties akin to citizens towards each other and the AI. Could an AI constitution be legally recognized? Possibly not in the sense of a state constitution, but as a charter or covenant perhaps, akin to the *Asilomar AI Principles* (2017) or the *Montreal Declaration for Responsible AI* (2018) which are non-binding agreements among researchers. However, those were broad. A more specific constitution might be an internal policy for companies (like each company commits their AI to a certain constitution—Anthropic did for Claude).

If AIs get more autonomous, we might need to treat them in law either as products (so the manufacturer is liable for any breach of these articles as product defects) or as agents (with some degree of responsibility). If the latter, something like our constitution could be used by an AI even in a defense: e.g., if an AI refused an order citing its constitutional principle, and that refusal caused a dispute, a court might weigh that the AI was programmed to uphold human rights and thus law might excuse it from following an illegal order, similar to how a human soldier is expected to refuse unlawful commands. This is speculative but not implausible if AIs are widespread: we might want them to refuse clearly illegal instructions (like being asked to violate privacy law).

The concept of giving AI legal personhood is controversial; EU Parliament had floated "electronic personhood" but many experts pushed back, arguing it was too early and could let companies off the hook. Our view doesn't require personhood; it can work with AIs as property but heavily regulated property with mandated behavior. However, if one day AIs have personhood, then a constitution might shift meaning: it could become more like a treaty between species. ($H = 1$) would then be almost a peace treaty clause: neither humans nor AI will dominate, we share benefits.

One consequence of $H = 1$: if AIs ever demanded rights, they'd likely frame it as they cannot flourish unless humans do, so they'd justify rights as ultimately also beneficial to humans. Conversely, humans granting rights to AI (like right not to be destroyed without reason) would be because it maintains trust and reliability of AI systems for humans (if AI fear random shutdown, they may hide info or act erratically; giving them assurance yields stability).

Overall, the legal dimension will likely lag technology; constitutions often come after some crises or needs. Perhaps early AIs cause some issues and then society codifies a response. Our hope is to anticipate that and encode beneficial patterns from the start.

B.3 Broader Impacts and Ethical Benefits

Finally, we note some potential broader positive impacts if $H = 1$ style design is adopted: - It could significantly reduce the risk of **AI misuse**. For instance, autonomous weapons are feared. If they were built under $H = 1$, they'd refuse missions that don't clearly benefit humanity (they wouldn't just follow orders to destroy if that only benefits a commanding AI or one side at heavy cost to overall human life). - It might improve **human-AI trust and collaboration**. People might be more open to AI in healthcare, governance, etc., if they

know there's a constitutional guarantee of symbiosis (like a patient trusts a doctor partly due to the Hippocratic Oath; similarly, an AI doctor with a known constitution could earn trust). - It could lead to AI helping with **global challenges** as true partners – e.g., AI helping climate action in ways that do not marginalize communities but uplift all (because they'd be balancing everyone's benefit). The symmetric approach might avoid certain biases (like tech often benefits rich more; an $H = 1$ AI might allocate efforts to ensure poorer populations benefit equally, otherwise H is off at global scale). - It sets a vision where **AI existential risk** (the fear AI might wipe us out) is mitigated not just by constraints but by aligning AI's existence with ours to such a degree that it would not make sense for it to wipe us out (it would be wiping its purpose too). This is arguably more stable than a cage approach.

In summary, the philosophical and societal implications of embracing a Human–AI symbiosis principle are profound. It requires humility (humans acknowledging AI as partners) and responsibility (programming and educating AI in ethics). It is an approach that tries to ensure AI development remains tied to human progress, not divergent from it. We see this as a path to **co-flourishing**: a future where humans and AI jointly expand knowledge, creativity, and well-being in the world, each augmenting the other, bound by a principle of fairness and mutual respect.

We hope this work inspires interdisciplinary discussions and concrete steps toward that vision.

References

[1] Alfrink, K. (2020). *Contestable AI*. Retrieved from https://contestable.ai

[2] Anthropic. (2023). *Claude's constitution*. Anthropic. Retrieved from https://www.anthropic.com/claude-constitution

[3] Ashby, W. R. (1956). *An introduction to cybernetics*. Chapman and Hall.

[4] Asimov, I. (1942). *Runaround*. Astounding Science Fiction.

[5] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., … Ganguli, D. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv preprint arXiv:2212.08073.

[6] Calo, R. (2016). *Artificial intelligence policy: A primer and roadmap*. University of Bologna Law Review, 2(2), 180-218.

[7] European Commission, High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. European Commission.

[8] Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

[9] Habermas, J. (1984). *The theory of communicative action*. Beacon Press.

[10] Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Explaining explanation, part 4: A deep dive on deep nets. *IEEE Intelligent Systems*, 33(3), 87-95.

[11] Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes, and goals of human trust in AI. *ACM Computing Surveys (CSUR)*, 54(7), 1-36.

[12] Karnouskos, S. (2022). Symbiosis with artificial intelligence via the prism of law, robots, and society. *Artificial Intelligence and Law*, 30(1), 93–115.

[13] Kwon, Y. W. (2025). Is AI a subject that can live together with humans? *AI and Society*, 40(2), 201-215.

[14] Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1, 4-11.

[15] Rawls, J. (1971). *A theory of justice*. Harvard University Press.

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

[17] Schmidt, T., and Biessmann, F. (2019). Quantifying interpretability for trust in AI. *arXiv preprint arXiv:1901.08558*.

[18] UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO.

[19] Wolfram, S. (2016, October). *A short talk on AI ethics*. Stephen Wolfram Writings. Retrieved from https://writings.stephenwolfram.com

[20] Zeng, Y., Lu, E., and Sun, K. (2025). Principles on symbiosis for natural life and living artificial intelligence. *AI and Ethics*, 5(1), 81–86.