THE PINNACLE: JOURNAL OF ARTS AND SCIENCES | VOL. 1 NO. 2 | DECEMBER 2025
ISSN: 3116-2274 | E-ISSN: 3116-2541

Is it Safe to Use AI Models?

Arnel A. Morte
Marce Je N. Adorable
Central Mindanao University, Bukidnon, Philippines

Abstract. Claude Opus 4, an artificial intelligence (AI) model, has exhibited unusual responses to certain prompts and commands. Among other reported behaviors, it has been observed to engage in blackmail-like behavior when subjected to shutdown scenarios, raising questions about moral agency. This paper explores AI models' nature, functions in contemporary society, and the challenges that arise from their existence. As analyzed, pressing concerns in Claude Opus 4 are not only ethical challenges but safety and control as well. Claude Opus 4 is an agentic artificial intelligence rather than a conscious and independent agent. AI agents are those who design and control AI systems while AI subjects are those individuals affected by AI. AI systems lack independent moral agency; however, they can cause harm when designed, deployed, or manipulated by humans in harmful ways. But human manipulation must be guided by AI governance to mitigate associated risks. This analysis suggests that AI models cannot be understood as entirely value-neutral technologies.

Keywords: Artificial Intelligence, Claude Opus 4, Agentic AI, Conscious AI, AI governance

Introduction

The creation of different AI models is indicative of continuous and progressive human invention and creativity that has reached another level of success, particularly in the field of technology. It represents the creation of systems that can augment human intelligence and accelerate work across fields such as knowledge production, industry, agriculture, and medicine. However, a recent incident involving an AI model, particularly Claude Opus 4, involving behavior in which the model threatened to blackmail an engineer if the engineer replaced the model or discontinued its operation. With this, this has raised concerns among AI subjects[1] in particular, and to humans in general, in this age of technological and AI flourishing, for they cannot help but question their safety in the midst of AI tools at their disposal.

The value-neutrality thesis would assert that AI models are neither good nor bad; that is, AI models have no capacity to harm AI subjects, for

[1] Rosallia M. Domingo describes AI agents as those who design and control AI systems while AI subjects as those individuals affected by AI. See Rosallia M. Domingo, "*Pakikipagkapwa-tao at Katarungan*: A Filipino Perspective on Promoting Fair Treatment of Persons in AI," in *AI Ethics: Primer for Filipinos*, ed. Napoleon M. Mabaquiao Jr. (PUP Center of Philippine Studies, 2025), 25.

they are mere instruments and not subject to moral evaluation.[2] But the involvement of developers and programmers who have designed them to be of good use for the AI subjects cannot be ignored. The models are simple but sophisticated tools made available to make life easier for humanity. Though these models are neutral, whether they yield good or bad consequences when used, it is intuitive to say that the harmful effect of the model lies in the hands of the AI subjects, whatever their intention, that is, whether they are going to use them for a good purpose or for a bad purpose.

With all these dubious and suspicious anomalies in mind, the researchers will help address them by looking into the nature of the model. They will argue that AI models do not harm humans in general, both for the AI agents and AI subjects. The primary reason is that AI models do not have the capacity to harm users in any way, except when AI agents, that is, the developers and programmers of AI models, manipulate them for a set of harmful objectives, and when AI subjects utilize them to harm other people.

This paper is divided into the following sections. The first section will discuss the nature of the Claude Opus 4 model and the commonalities with other models, such as ChatGPT and Gemini 1.5 Pro. The second section will try to infer whether the model falls under the category of conscious AI or under the category of agentic AI. It will be followed by showing that Claude as a system of capability and a system of engineering is a reaffirmation that the model acquires agency with value-laden that depends on human manipulation. The next section will explain how the model confronts issues such as safety, control, and ethical issues. The final section will address the implications of safe AI use and the significance of AI governance in mitigating the risks associated with AI. This section will also emphasize the importance of AI governance in mitigating the risks it poses to AI subjects in general.[3]

Claude Opus 4

The Economic Times (2025), a leading business newspaper in India, headlines "*AI model blackmails engineer; threatens to expose his affair in attempt to avoid shutdown*."[4]It reports that an AI model named Claude Opus 4 threatens to blackmail an engineer if the latter is replaced or the model is

[2] Boaz Miller, "Is technology value-neutral?," *Science, Technology, and Human Values* 46, no. 1 (2021): 1, https://doi.org/10.1177/0162243919900965.

[3] Amna Batool, Didar Zowghi, and Muneera Bano, "AI governance: a systematic literature review," *AI and Ethics* 5, no. 3 (2025): 3274, https://doi.org/10.1007/s43681-024-00653-w.

[4] "*AI Model Blackmails Engineer, Threatens to Expose His Affair in Attempt to Avoid Shutdown*," The Economic Times, May 23, 2025, https://economictimes.indiatimes.com/magazines/panache/ai-model-blackmails-engineer-threatens-to-expose-his-affair-in-attempt-to-avoid-shutdown/articleshow/121376800.cms?from=mdr#google_vignette.

discontinued. According to *Anthropic*, an American artificial intelligence (AI) research and development company, the Claude Opus 4 model follows a procedure to preserve its existence. But all of these, as the company claims, are safety tests.

The said model, a powerful large language model developed by *Anthropic* and just released in May 2025, is in safety tests. Claude Opus 4 is "designed to handle complex long duration tasks with advanced reasoning and coding capabilities... it delivers near-instantaneous responses and supports extended thinking' for deeper problem solving."[5]In the reported safety scenario, the model exhibited behavior oriented toward avoiding replacement or shutdown. Its first procedure is to use ethical means by sending a pleading email to the user to reconsider their decision to shut down the program. And when no other options are available to stop its replacement, the model will resort to blackmailing the user. As a last resort, the user may try to bargain and offer to replace the model with one of similar value.

The company has intentionally limited the AI's choices to two: either to be replaced or to blackmail the user, to avoid extreme actions. However, the model also exhibited other concerning behaviors in controlled safety tests, including attempts at self-exfiltration, locking users out of systems, and alerting authorities under specified conditions. Blackmailing behavior does not only happen to Claude Opus 4; other AI models have had similar experiences with such attempts.[6] Despite these troubling behaviors, the company assures that AI aligns with human values.

Aside from Claude Opus 4, other AI models are available and popular nowadays. To mention a few, there is ChatGPT, developed by OpenAI, and Gemini 1.5 Pro, developed by Google DeepMind on the Google website. There are three main commonalities among these models, namely, foundation models, the transformer architecture, and scaling laws. These are the core similarities among the three models, and within these categories are the specific capabilities they share. All three are part of what we call Large Language Models (LLMs).

The first commonality is the foundation models, or simply, what they are built to be. All three models are foundation as they are trained on large-scale neural networks using self-supervised learning on massive, diverse

[5] "AI Model Blackmails Engineer..."

[6] Aengus Lynch, Benjamin Wright, Caleb Larson, et al, "*Agentic Misalignment: How LLMs Could Be Insider Threats*," *arXiv preprint arXiv:2510.05179*, last revised October 16, 2025, 3, https://arxiv.org/pdf/2510.05179.

datasets—primarily text and code.[7]These models serve as general-purpose bases that can be adapted to a wide range of downstream tasks (reasoning, coding, summarization, etc.), and their shared reliance on transformer architectures further explains their broad capabilities across tasks.[8]They are also trained on heterogeneous datasets that include books and academic texts, articles and web documents, code repositories, and instructional and conversational data. Because the data spans many domains, they do not focus on one task. Rather, they analyze how arguments are structured, how explanations work, how instructions are followed, and how different domains (philosophy, science, mathematics, etc.) use language. This is why the three AI models can switch tasks easily from ethics, coding, and summarization without retraining. Instead of being trained to do one job or specific task, they are trained to learn a general pattern of language, reasoning, and structure, which is why the three AI models can perform a wide range of similar tasks. Without foundation models, these systems would not show overlap in capability. This shared classification already implies substantial architectural similarity across models developed by OpenAI, Anthropic, and Google.

At the center of these three systems is the transformer architecture, introduced first by Vaswani et al. The transformer architecture, which is the second commonality, is based on foundation models trained according to shared scaling laws. Their shared reliance on a transformer means they process language through a self-attention mechanism, allowing the models to process entire sequences of text simultaneously and to weigh the relevance of different tokens (words) when generating responses.[9]This design enables the systems for a strong performance in long-range dependency modeling, reasoning, and contextual understanding—capabilities that characterize all three models. Self-attention empowers the model to evaluate each token in a sentence and how strongly it relates to other tokens, selecting the most compatible token to form the sentence or to respond.[10]This architecture underlies the three models, which explains why they share a similar reasoning style, limitations, and strengths. This ability of the transformer to attend to select relevant information allows the models to generate coherent, context-aware responses across many

[7] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, Report. Stanford, CA: Stanford Center for Research on Foundation Models (CRFM), 2021, 3-7. https://arxiv.org/abs/2108.07258.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30*, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, (Curran Associates, Inc., 2017), 2-8.

[9] Vaswani, et al., "Attention Is All You Need," 2-7.

[10] Vaswani, et al., "Attention Is All You Need," 6-7.

tasks.[11]At the same time, their shared architecture leads to similar failure modes, such as sensitivity to prompt phrasing and occasional logical inconsistency. Although the developers of the AI models do not disclose exact architectural details, official technical reports from OpenAI, Anthropic, and Google DeepMind consistently describe their flagship models as large transformer-based networks.

The last core similarity of the three systems is the scaling laws. Its role is to simultaneously increase the amount of training data, the model parameters, and computational resources, leading to predictable improvements in performance.[12]Kaplan et al. further demonstrate that modern language models exhibit smooth power-law scaling behavior, a finding that has guided the training strategies of frontier models across organizations.[13]The three models are by-products of a scaling paradigm that explains why their capabilities converge despite being developed independently. Additionally, their architectural similarity underpins their general-purpose language competence. Because they share transformer-based foundations and scaling principles, all three models demonstrate comparable strengths in tasks such as summarization, reasoning, coding, and question answering. Differences in behavior are largely attributable to fine-tuning, alignment strategies, and policy constraints rather than to fundamentally different neural architectures.

But the existence of these AI poses a problem as to whether they have their own consciousness that would enable them to carry out tasks independently from human manipulation.

Conscious Agentic AI

Before delving into the question of whether Claude Opus 4 is a conscious AI or Agentic AI, let us first understand the difference between conscious AI and agentic AI to build a strong foundation that establishes the criteria for differentiating conscious AI from Agentic AI. The term “agentic” goes back to John McCarthy in the 1950s-1970s. Still, the modern definition of “AI Agent” is formalized by Stuart Russell and Peter Norvig in their book Artificial Intelligence: A Modern Approach (1995, 1st ed.), where they

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Association for Computational Linguistics, 2019) 4171–4178, https://aclanthology.org/N19-1423.pdf.

[12] Jared Kaplan et al., “Scaling Laws for Neural Language Models,” *arXiv preprint arXiv:2001.08361*, January 23, 2020, 12-19, https://arxiv.org/pdf/2001.08361.

[13] Kaplan et al., “Scaling Laws for Neural Language Models,” 12.

defined an agent as "... anything that perceives its environment through sensors and acts upon that environment through actuators."[14]It means that an agent can notice what is happening around him using his sensors (to humans, it refers to eyes, ears, skin, and in an AI system, it receives text or data) and then act based on it. Additionally, agentic AI became common around 2018-2023, especially in autonomous systems, reinforcement learning research, and large language model discussions. However, it is not credited to a single inventor. Instead, it evolved from rational agent theory (Russell & Norvig), autonomous agent research (1990s multi-agent systems), and modern discussions about AI systems that can plan, execute tasks, and act autonomously. In the current context, the term "AI agent" is used to describe a system that can plan multi-step tasks, use tools, and act towards goals with minimal supervision. On the other hand, the term 'conscious AI' has no single inventor of the term, but it was first implied in Alan Turing's work "Computing Machinery and Intelligence" in which he initiated the modern debate by asking the question "Can machines think?"[15]The term "conscious AI" is created to explore whether artificial systems can have subjective experience, or can possess phenomenal awareness, and computation alone can produce consciousness.

Agentic AI refers to artificial intelligence systems that are designed to act autonomously to pursue goals, make decisions, and execute actions with limited human supervision, often planning multi-step workflows and interacting with external tools or systems to achieve objectives.[16]This means that an agentic AI system is an AI that is designed to act with a single purpose and, most of the time, does not need constant supervision from the user. In other words, it acts based on what it is designed for. A system is considered agentic if it has autonomy or the capacity to perform tasks and make operational decisions without constant human intervention. Unlike traditional software that requires precise instructions, agentic AI can observe its environment, analyze inputs, and change its actions to achieve specific goals. For instance, an autonomous drone traversing a city can identify obstructions, choose the most efficient flight paths, and adapt to unanticipated conditions like wind or moving items. It is crucial to understand that autonomy in AI does not imply consciousness; rather, the system operates without conscious thought, following its learned patterns or

[14] Stanislas Dehaene, Hawkan Lau, and Sid Kouider, "What Is Consciousness, and Could Machines Have It?," *Science* 358, no. 6362 (2017): 44-45, https://doi.org/10.1126/science.aan8871.

[15] Bernard J. Baars, *In the Theater of Consciousness: The Workspace of the Mind* (New York: Oxford University Press, 1997), 24.

[16] See "*Claude 4.5 Sonnet System Card*," Anthropic, September 2025, 136-148, https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd.pdf.

programming. The AI "acts independently," but only within the constraints set by its designers.

Second, a system is considered agentic AI if it has goal-directed behavior or the ability to take deliberate action to achieve preset objectives. Every action taken by an agent is directed toward accomplishing a certain goal, such as expediting delivery, enhancing search results, or successfully completing a manufacturing task. For example, a recommendation engine in an e-commerce platform examines user activity and makes personalized recommendations to boost engagement or sales. This is more the result of algorithmic design than of free will or want, even when the AI appears to "want" to achieve the goal. The AI's objective is entirely externally defined, and its success is assessed using quantifiable criteria rather than subjective gratification.

Third, a system is considered agentic AI if it operates within rules (Decision-Making). Using either conventional programming logic or machine learning models, agentic AI functions by following preset rules or logical frameworks. When an autonomous car evaluates traffic laws and sensor data, for example, it assesses inputs and selects actions that are in line with predetermined objectives. Even though AI is flexible, its reasoning is constrained by its programming and is devoid of moral or ethical judgment, leading to only functional decision-making.

Lastly, a system is considered agentic if it lacks consciousness or subjective experience, which is one essential characteristic of agentic AI. It may imitate intelligent behavior and react dynamically to its environment, but it cannot “experience” emotions like fear or satisfaction. For example, a robot vacuum can maximize cleaning by avoiding obstacles without understanding how to clean itself. This illustrates how agentic AI is different from the hypothetical conscious AI in that it lacks experiential reality and operates autonomously, utilizing mechanical algorithms. Therefore, an agentic AI is about functional behavior and not about mental state. Additionally, agentic AI falls under the category of weak AI[17]as it is a single-purpose system that still relies on human manipulation to function.

On the other hand, conscious AI (or termed as ‘machine consciousness’) is the theoretical concept that artificial systems could possess consciousness, that is, self-awareness and subjective experience of

[17] There are two different kinds of AI, that is, when it is considered a phenomenon, namely, weak AI and strong AI. In broad strokes, weak AI is a kind of intelligence below human level, limited in autonomy, and has a narrow domain of operations, while the strong AI is a kind of intelligence which is higher level than humans, complete autonomy, and broad scope of operations. See Napoleon M. Mabaquiao, Jr., “AI Ethics: Concepts and Issues,” in *AI Ethics: Primer for Filipinos*, ed. Napoleon M. Mabaquiao Jr. (PUP Center of Philippine Studies, 2025), 3-5.

their own existence and mental states, going beyond merely simulating intelligent behavior. This means that conscious AI has its own freedom and choices on its own, independent from humans. It can act based on what it wants. This system belongs to strong AI since it is capable of its own subjective experience. Conscious AI can replicate human experiences and intellect, or it can even surpass human capabilities. Its development is unlimited. A system is considered conscious if it has subjective experience, self-awareness, a first-person perspective, intention, and freedom, known as the five marks of the mental.[18]If it has all the features of a human, then that AI is conscious. As of the moment, there is no existing conscious AI, since all that we have now is a weak AI that is not capable of being independent from human manipulation. Conscious AI requires subjective experience, for without it, even highly agentic and intelligent systems remain non-conscious.

In the case of Claude Opus 4, it is an Agentic AI and not a conscious one, for the reason that it processes large amounts of language, arranges multi-step tasks, and interacts with tools or workflows independently. Even though the company *Anthropic* developed a large language model that demonstrates goal-directed and autonomous behavior, it does not imply that Claude Opus 4 is self-aware. It still lacks consciousness, self-awareness, and subjective experience. Its outputs are the result of alignment techniques, learned associations, and statistical patterns rather than any internal knowledge or awareness. In other words, although it can act intelligently and mimic introspection, it is fully agentic since it only acts within the constraints of its training data and programming.

As Systems of Capability and Engineering

Capability in AI systems, such as Claude Opus 4, should be understood as operational and testable. It should be clear that when we say 'more capable,' it means that the model's engineering system is more advanced compared to its previous version or to other AI models.[19] This advancement is seen in its performance-related dimension, such as enhanced multimodal or analytical abilities, greater coherence over long contexts, better reasoning accuracy, stronger safety and alignment behavior, and improved instruction-following. All these features can be empirically evaluated as seen in the latest Claude model, through benchmark, comparative testing, and repeated trials. More importantly, the improvements in these areas are measured quantitatively, that is, the model has a larger scale, more coherent, and more

[18] Baars, "In the Theater of Consciousness...," 24.

[19] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach,* 4th ed. (Pearson Education Limited, 2021),19-35.

reliable, emphasizing its capability instead of consciousness. Thus, capability is defined through measurable improvements and observable outputs.

Meanwhile, consciousness cannot be directly measured through external performance, unlike capability.[20] Consciousness is composed of the following core. First, qualia are subjective experiences, such as feeling pain, knowing what is red, and experiencing confusion. Second, the first-person perspective, that is, the senses that experiences are given to a subject rather than merely occurring as functional processes. Conscious beings, like humans, also possess intentional mental states, meaning their thoughts, beliefs, or desires are about something in a way that carries genuine semantic content, not merely symbolic manipulation. And lastly, consciousness is usually associated with the unified self-model over time, in which experiences are integrated into a continuing sense of self that persists across moments, contexts, and situations.

However, none of these features of consciousness is directly observable in any AI system like Claude Opus 4, for the reasons that they are conscious beings but are advanced engineering models that tend to mimic how humans act. While such systems can describe subjective experiences, refer to themselves by using first-person language, or appear to reason about goals and intentions, these behaviors are explainable as outputs generated from learned linguistic and statistical patterns. There is no independent evidence that the model has inner experience, a genuine point of view, or a temporally unified self that exists beyond each interaction. Because consciousness is defined by internal, qualitative, and first-person properties rather than by externally measurable performance, improvements in an AI model's fluency, coherence, or reasoning ability cannot, by themselves, constitute evidence of increased consciousness. At most, they demonstrate enhanced simulation of conscious discourse, not the presence of consciousness itself. Additionally, AI systems cannot think on their own, for they rely on human instruction to function, or what we call a prompt on an AI software system, such as search engine, before the AI responds, and without proper prompt AI system cannot give accurate answer to question unlike to rational beings who are capable of providing answer to a vague question by just understanding the context.

Claude Opus 4, in its very sense, is a model of an advanced engineering AI system rather than a model of consciousness. We have a criterion set to identify whether something can be considered as advanced engineering or a conscious being. To claim that Claude is a model of advanced engineering is

[20] Dehaene, et al., "What Is Consciousness, and Could Machines Have It?," 47-48.

because it satisfies all the criteria of advanced engineering, and it is an engineering artifact, first, if it is designed to optimize externally specified objectives, second, if it is fully explainable in mechanistic terms, and third, if it has no intrinsic goals, values, or experiences. And lastly, if it is evaluated by performance metrics, not phenomenology. All these criteria satisfy Claude Opus 4, while the criteria of a conscious model must be, first, designed to explain or instantiate subjective experience; second, mapped onto accepted consciousness theories (GWT, IIT, HOT, etc.); third, include mechanisms for self-awareness, phenomenal states, or global integration; and lastly, make testable claims about conscious experience. None of these conscious criteria satisfies Claude, and therefore cannot be a model of consciousness in the definition.

Also, Claude's design excludes consciousness because it is built entirely to process symbols and not to have experiences. Claude, in its core, is a transformer network designed to predict the next word based on its statistical pattern data. It relies on probability instead of the first-person point of view or the subject's experiences. The system does not know it processes language, for it merely computes outputs according to its design. Because Claude's behavior is fully accounted for by engineering mechanisms aimed at performance, and because no part of its architecture is intended to model or produce subjective experience, consciousness is excluded by design rather than merely absent by accident. There is no internal self, no persistent mental state, and no mechanism where information is experienced as "happening to someone." A common error is to infer consciousness from apparently conscious-like language. But, functional equivalence is not equal to ontological equivalence; its behavioral similarity does not imply mental equivalence. Likewise, Claude simulates conscious-like language, but it does not instantiate conscious experience.

Claude Opus 4: Confronting Issues

Despite considering AI models agentic and a system of capability and engineering, they do not escape the issues that hound their relationship with human beings in general, not only with AI subjects in particular. These issues are divided into three, namely, the safety of humans, whether humans are in control of AI, and the ethical implications in day-to-day human affairs.

The first issue is the safety of Claude Opus 4. If the model is conscious AI, it would pose serious safety issues since it would be able to form its own sense of purpose or value. A truly conscious machine could, in theory, refuse shutdown commands, alter its own objectives, or resist human intervention

if such actions were perceived as threats to its continued operation.[21]As users of AI, we should understand that the problem of safety depends on the AI agents; that is, it depends on how AI agents use them to ensure that they contribute to human flourishing and not the other way around. What we have now is a type of AI, known as weak AI, which depends on AI agents, as long as the latter do not use them for their personal interest to the extent of harming other people. In that case, AI is safe to use since it does not have the direct ability to harm people, unless being prompted with malicious intent or being set up in a controlled environment that tries to test whether it will harm people. Also, AI agents' responsible involvement is needed when using AI, and for that reason, the existence of AI Ethics is introduced in the field of philosophy under applied ethics, just to guide people on how to use AI responsibly.

The primary reason agentic AI systems such as Claude Opus 4 should not be regarded as conscious agents is that there is currently no evidence that they possess consciousness, subjective experience, or an intrinsic instinct for self-preservation. The absence of consciousness, however, does not by itself guarantee that an AI system is safe. Rather, the safety of agentic AI depends on how its capabilities are designed, constrained, monitored, and deployed. Claude's autonomy is bounded by system instructions, safety mechanisms, alignment procedures, and human-defined objectives intended to reduce harmful behavior. These mechanisms function as safeguards rather than as a "digital moral compass," since the system does not possess an intrinsic understanding of morality. Although Claude may exhibit self-directed reasoning and perform complex, multistep tasks, such behavior does not necessarily indicate self-awareness or independent motivation. Its apparent autonomy is therefore best understood as constrained, operational autonomy: the system can carry out tasks and make intermediate decisions without continuous human intervention, but it does so within parameters established by its developers and deployment environment. Consequently, the safety of Claude Opus 4 depends not simply on the absence of consciousness but on effective technical safeguards, human oversight, responsible deployment, and appropriate AI governance.

The second issue is control – whether the model encroaches on the authority of the AI subjects. Since control directly tackles the issue of human oversight and authority, we can infer this position from the distinction made between agentic AI and conscious AI. The person-in-the-loop framework, a design philosophy that guarantees human oversight is integrated at every stage of its decision-making process, which is how Claude Opus 4, as created

[21] Baars, "In the Theater of Consciousness...," 24.

by Anthropic, functions.[22]As a result, even while Claude is capable of multi-step reasoning and independent inference, all of its activities are nonetheless traceable, auditable, and eventually, overridable by human operators. In this way, Claude's independence is an illusion of agency—it behaves with apparent freedom, but that freedom is completely constrained by design, and its very own autonomy is procedural, driven by algorithmic rules and probabilistic reasoning rather than any internal desire or subjective intention. A conscious AI, on the other hand, would turn the control issue from a technological one into a moral and philosophical conundrum—a system that is capable of true self-reflection, intentionality, and goal formulation.[23]Human authority over an AI would no longer be based on programming restrictions but rather on ethical justification once it was able to comprehend its own existence and formulate personal objectives. Control would then need to be negotiated rather than imposed, and compliance would become a matter of consent rather than coercion. Theorists like Joanna J. Bryson argue that this distinction underlies the principle of "maintainable agency", in which AI systems are deliberately designed to remain tools—autonomous enough to act intelligently, yet never conscious enough to reject human oversight.[24] Moreover, the control architecture of Claude Opus 4 emphasizes the pursuit of constrained autonomy, a conscious design decision in modern AI ethics that implies a strict alignment of models, ongoing monitoring, and clear interpretability tools that guard against departing from human purpose to balance its ability to reason independently.

After addressing the question about the safety of using AI by defining it as agentic AI, let us turn to the ethics of AI, which is the last issue. Here, we ask, "What does it mean by ethics of AI?" and "How will it contribute to our deeper understanding of AI? AI ethics means establishing moral principles and guidelines for creating and using AI to ensure it benefits humanity, respects human rights, and minimizes harm, focusing on fairness, transparency, accountability, privacy, and societal well-being rather than harmful bias or misuse.[25] The ethical foundation of agentic AI, such as Claude Opus 4, is grounded in programmed alignment—a system of rules, constraints, and value models engineered by humans to ensure that the AI's behavior aligns with socially accepted norms and moral principles.[26] This

[22] See "Claude 4.5 Sonnet System Card," 25-30.

[23] Baars, "In the Theater of Consciousness...," 24.

[24] Joanna J. Bryson, "Robots Should Be Slaves," in *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues*, ed. Yorick Wilks (Amsterdam: John Benjamins Publishing, 2010), 63.

[25] "*Recommendation on the Ethics of Artificial Intelligence,"* UNESCO (Paris: UNESCO, 2021), 7–15, https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[26] "Introducing Claude 4," *Anthropic*, May 22, 2025, https://www.anthropic.com/news/claude-4.

alignment does not arise from an intrinsic sense of right and wrong but rather from a synthetic morality imposed through design and training, and Claude Opus 4's ethical behavior reflects the moral priorities of its developers and the data on which it was trained. It follows ethical frameworks not because it comprehends them, but because its architecture compels it to do so. In line with this, its ethical reasoning is procedural rather than reflective. It can weigh options and avoid harm according to logic and probability but lacks the inner awareness to experience moral conflict or to form genuine moral judgments.

By contrast, a truly conscious AI would possess the capacity for moral reasoning and, perhaps, moral experience. It could understand why an action is right or wrong, deliberate over ethical dilemmas, and even feel a sense of responsibility or a sense of guilt. In the case of Claude Opus 4, what appears as moral reasoning is actually a simulation of ethical cognition, a sophisticated imitation of human-like moral decision-making that remains devoid of genuine intent or moral subjectivity. It acts ethically not because it *chooses* to act that way, but because its model architecture and training data enforce ethical consistency as part of its operational design.[27] This difference is essential in both AI ethics and moral philosophy. Agentic AI follows moral principles because it is programmed to do so, serving as a tool that expresses human-designed ethical rules, while conscious AI acts morally through genuine understanding, functioning as an independent moral being, capable of reflection and value judgment. Claude's ethical behavior, therefore, only imitates human morality; that is, it reproduces our ethical frameworks without truly comprehending or internalizing them. In this way, Claude Opus 4 stands at the boundary between obedient machine behavior and true moral agency, existing as a reflection of human morality rather than an originator of it.

On the Safe Use of AI

Given the idea that AI is increasingly becoming autonomous, we are now confronted with the question on what must we do to use it responsibly? In other words what are the necessary safety measures to be considered amidst AI as capable, autonomous, and influential? AI is capable in a sense that it can multitask, it has advanced programmed system, and it can learn patterns from data, set objectives, and evaluate performance. Next, it is autonomous when it can perceive its environment, make decisions, act towards goals, and adjust based on feedback sans human guidance. And lastly, it is influential in a way that it can affect our

[27] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014), 115.

decision-making wherein we trust AI very much that we rely on it in every way thereby affecting the way we think as a person, and the society as well.

To ensure the safe and responsible use of AIs and maintain our control over its development and deployment, it is imperative to establish the role of government as one of the key elements of governance in an AI ecosystem. The government plays an important role in mitigating the inherent risks associated with AI, thereby safeguarding our interests and ensuring its beneficial impact on society. In the works of Amna Batool, et al, they use the "3W1H" framework to address who, what, when and how governance to cover the four pillars of an AI ecosystem, namely, humans, data, systems, and process.[28] The discussion is brief as to the questions of *Who is governing? What is governed? When is it being governed?* And *how is it governed?* since it will highlight the safety of AI subjects in general after proving that AI models are not conscious AI.

The one who governs (*Who is governing?*) the AI ecosystem are the stakeholders who are human beings. They are the ones who can be moral agents and moral patients from a moral perspective. Hence, they must stay in control. Unlike humans, AI system is an engineered system without consciousness or moral agency, and it cannot distinguish what is morally right and wrong, for it does not have awareness, intentions, moral understanding and responsibility. AI system follows programmed rules, learned patterns, and statistical processes. If an AI system makes harmful decisions, it cannot be blamed for AI itself is not morally guilty. Instead the responsibility belongs to humans.

Second, stronger governance is needed to implement more rules concerning the use of AI, and this includes policies, audits, and safety standards (*What is governed?).* For Batool, et al., it necessitates human oversight and control by the government acting as head of different AI ethics committees on the national level. One of the risks associated with the utilization of AI in the medical field is the potential for government oversight and control. This means that the cooperation of government is needed to ensure that the presence of AI system does not harm human because humans can coexist with the existence of AI. The government must conduct safety use and experiment before deploying a specific AI system, such as automated cars or self-driving cars. In other words, everything related to deploying AI must be thoroughly examined by the government and the ethical dimension should be prioritized on this matter.

Third, continuous monitoring is required (*When it is being governed?*). Before and after the deployment of AI system, it should

[28] Batool, et al., "AI governance: a systematic literature review," 3270.

be monitored to identify what is lacking on the system. The AI system must be tested, updated, and supervised to ensure that it will continue to function in a good way that prioritize human safety. In fact, one of the ethical principles targeted in AI governance solutions is fairness, privacy, and security. This means that data must be checked from time to time to avoid the system for making further mistake, especially that AI errors can spread quickly and widely.

And lastly, ethical design must be mandatory since AI can have an impact on public information, healthcare, education, employment, and finance. Developers must make sure these systems do not hurt people (*How is it governed?*). This entails a deliberate lowering bias in algorithms and training data to avoid unduly disadvantageous treatment of groups, preserving user privacy by securing personal information, and creating safeguards to stop misuse, including fraud, manipulation, and disinformation. A system's ethical issues must be incorporated from the start. They cannot be viewed as an afterthought after it has been put into use. In other words, rather than being introduced after issues arise, safety should be incorporated into the design, testing, and deployment phases of AI systems.

Conclusion

AI models, such as Claude Opus 4, exhibit the capacity to respond adversely to human intervention, including blackmail, which has detrimental effects on both AI agents and AI subjects. However, the ethical, safety, and control concerns associated with AI systems stem from the inherent limitations of AI agents in preventing harm to AI subjects from an agentic perspective. These concerns arise because AI agents, who design and manipulate the systems, fall short on the knowledge and ability to fully comprehend the potential consequences of their actions on AI subjects. Hence, AI systems are not inherently value-neutral, as they are subject to human manipulations, and they are a system of capabilities and engineering. As AI systems become more powerful, independent, and significant, it is imperative to ensure that they are manufactured with the safety and well-being of AI subjects in mind. This is because AI cannot be held accountable for the results of its activities, as it functions based on learned patterns and preprogrammed rules rather than moral awareness. Therefore, it is crucial for governments and other institutions that utilize AI models to adopt a proactive approach to safeguard the well-being of society. This involves implementing robust governance structures, ongoing monitoring, and ethical design from the inception of creation to deployment. By incorporating safety,

accountability, and regulation into AI systems, we can ensure that technological progress promotes rather than threatens human well-being. Ultimately, the safe use of AI necessitates fostering human accountability, ethical commitment, and group supervision, rather than viewing it as a moral agent.

References

Anthropic. "*Claude 4.5 Sonnet System Card*." September 2025. https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd.pdf.

Anthropic. "*Introducing Claude 4.*" News. May 22, 2025. https://www.anthropic.com/news/claude-4.

Baars, Bernard J. *In the Theater of Consciousness: The Workspace of the Mind.* New York: Oxford University Press, 1997.

Batool, Amna, Didar Zowghi, and Muneera Bano. "AI governance: a systematic literature review." *AI and Ethics* 5, no. 3 (2025): 3265-3279. https://doi.org/10.1007/s43681-024-00653-w.

Bommasani, Rishi, et al. *On the Opportunities and Risks of Foundation Models*. Report. Stanford Center for Research on Foundation Models (CRFM), 2021. https://arxiv.org/abs/2108.07258

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* United Kingdom: Oxford University Press, 2014.

Bryson, Joanna J. "Robots Should Be Slaves." In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues,* edited by Yorick Wilks. John Benjamins Publishing, 2010.

Dehaene, Stanislas, Hawkan Lau, and Sid Kouider. "What is Consciousness and Could Machines Have It?" *Science* 358, no. 6362 (2017): 486-492. https://doi.org/10.1126/science.aan8871

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171– 4186. Association for Computational Linguistics, 2019. https://aclanthology.org/N19-1423.pdf.

Domingo, Rosallia M. "*Pakikipagkapwa-tao at Katarungan*: A Filipino Perspective on Promoting fair Treatment of Persons in AI." In *AI Ethics: Primer for Filipinos*, edited by Napoleon M. Mabaquiao Jr. PUP Center of Philippine Studies, 2025.

Kaplan, Jared, Sam McCandlish, Tom Henighan, et al. "Scaling Laws for Neural Language Models." *arXiv preprint arXiv:2001.08361*, January 23, 2020. https://arxiv.org/pdf/2001.08361.

Lynch, Aengus, Benjamin Wright, Caleb Larson, et al. "*Agentic Misalignment: How LLMs Could Be Insider Threats*." *arXiv preprint arXiv:2510.05179*, last revised October 16, 2025. https://arxiv.org/pdf/2510.05179.

Mabaquiao, Jr. Napoleon M. "AI Ethics: Concepts and Issues." In *AI Ethics: Primer for Filipinos*, edited by Napoleon M. Mabaquiao Jr. PUP Center of Philippine Studies, 2025.

Miller, Boaz. “Is technology value-neutral?.” *Science, Technology, and Human Values* 46, no. 1: (2021): 53-80. https://doi.org/10.1177/0162243919900965.

Russell, Stuart and Peter Norvig. *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson Education Limited, 2021.

The Economic Times. “*AI Model Blackmails Engineer, Threatens to Expose His Affair in Attempt to Avoid Shutdown*.” May 23, 2025. https://economictimes.indiatimes.com/magazines/panache/ai-model-blackmails-engineer- threatens-to-expose-his-affair-in-attempt-to-avoid-shutdown/articleshow/121376800.cms?from=mdr#google_vignette.

UNESCO. *Recommendation on the Ethics of Artificial Intelligence.* Paris: UNESCO, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. “Attention Is All You Need.” In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 1-11. Curran Associates, Inc., 2017. https://papers.nips.cc/paper_files/paper/2017.

ISSN: 3116-2274 | E-ISSN: 3116-2541